

**WEB TABANLI OTOMATİK DİL TANIMA VE ÇEVİRME SİSTEMİ
GELİŞTİRİLMESİ**

Uraz YAVANOĞLU

**YÜKSEK LİSANS TEZİ
BİLGİSAYAR MÜHENDİSLİĞİ**

**GAZİ ÜNİVERSİTESİ
FEN BİLİMLERİ ENSTİTÜSÜ**

**MAYIS 2009
ANKARA**

Uraz YAVANOĞLU tarafından hazırlanan WEB TABANLI OTOMATİK DİL TANIMA VE ÇEVİRME SİSTEMİ GELİŞTİRİLMESİ adlı bu tezin Yüksek Lisans tezi olarak uygun olduğunu onaylarım.

Prof. Dr. Şeref SAĞIROĞLU

Tez Danışmanı, Bilgisayar Mühendisliği ABD

Bu çalışma, jürimiz tarafından oy birliği ile Bilgisayar Mühendisliği Anabilim Dalında Yüksek Lisans tezi olarak kabul edilmiştir.

Prof. Dr. Sezai DİNÇER

Elektrik Elektronik Mühendisliği ABD, Gazi Üniversitesi

Prof. Dr. Şeref SAĞIROĞLU

Bilgisayar Mühendisliği ABD, Gazi Üniversitesi

Doç. Dr. M. Ali AKCAYOL

Bilgisayar Mühendisliği ABD, Gazi Üniversitesi

Tarih: 15/05/2009

Bu tez ile G.Ü. Fen Bilimleri Enstitüsü Yönetim Kurulu Yüksek Lisans derecesini onamıştır.

Prof. Dr. Nail ÜNSAL

Fen Bilimleri Enstitüsü Müdürü

TEZ BİLDİRİMİ

Tez içindeki bütün bilgilerin etik davranış ve akademik kurallar çerçevesinde elde edilerek sunulduğunu, ayrıca tez yazım kurallarına uygun olarak hazırlanan bu çalışmada orijinal olmayan her türlü kaynağa eksiksiz atıf yapıldığını bildiririm.

Uraz YAVANOĞLU

**WEB TABANLI OTOMATİK DİL TANIMA VE ÇEVİRME SİSTEMİ
GELİŞTİRİLMESİ
(Yüksek Lisans Tezi)**

Uraz YAVANOĞLU

**GAZİ ÜNİVERSİTESİ
FEN BİLİMLERİ ENSTİTÜSÜ**

MAYIS 2009

ÖZET

İnternetin hızla yaygınlaşması bu ortamlara sunulan hazır bilgi ve belgeleri zenginleştirdiği kadar farklı dildeki içeriklerinde internette sunulmasını kolaylaştırmıştır. Farklı dildeki bilgilere ulaşmak ise günümüzde farklı bir sorun haline gelmiştir. Bu çalışmada, uluslararası kabul gören İngilizcenin yanında internet ortamında her dile ait sayısal belgelerin dilini tanımak, internette sunulan bilgilerin içeriğini farklı dillere çevirme; üzerinde farklı işlemler yapmayı kolaylaştırmasını sağlamak için kullanıcıların içerik dillerini hiç bilmedikleri internetteki sayısal dokümanları yapay sinir ağları ile birlikte geliştirilen yöntemler kullanılarak istenilen dile otomatik olarak dönüştürebilecekleri web tabanlı otomatik dil tanıma ve 41 dile otomatik çevirme platformu geliştirilmiştir. Bu çalışmanın, internet içeriklerinin kullanıcılar tarafından daha etkin olarak kullanılmasını sağlaması beklenmektedir.

Bilim Kodu : 902.1.014
Anahtar Kelimeler : Dil tanıma, dil dönüştürme, web platformu, yapay sinir ağı, web tabanlı yazılım
Sayfa Adedi : 147
Tez Yöneticisi : Prof. Dr. Şeref SAĞIROĞLU

**DEVELOPING WEB BASED AUTOMATIC LANGUAGE
IDENTIFICATION AND TRANSLATION SYSTEM**

(M.Sc. Thesis)

Uraz YAVANOGLU

**GAZI UNIVERSITY
INSTITUTE OF SCIENCE AND TECHNOLOGY**

MAY 2009

ABSTRACT

Language identification is an important task for web information retrieval services. Automatic language identification and translation have become increasingly important, as more and more documents are being served on internet within many languages. The identification problem can be seen as a specific instance of the more general problem of an item classification though its attributes. In the processes, artificial neural networks were used to identify languages in multi-lingual documents. The results have shown that the approach presented in this work is very successful to meet the expectations in real-time language identification accuracy and translate into 41 different language with the help of a developed platform. It is expected that with the support of this thesis web based documents might be used by internet users as effective as possible.

Science Code : 902.1.014

**Key Words : Language identification, language translation, web platform,
artificial neural network**

Page Number: 147

Adviser : Prof. Dr. Seref SAGIROGLU

TEŞEKKÜR

Çalışmalarım boyunca değerli yardım ve katkılarıyla beni her zaman yönlendiren danışman hocam Sayın Prof. Dr. Şeref SAĞIROĞLU'na Yüksek Lisans eğitimim süresince yardımlarını esirgemeyen hocalarım Sayın Yrd. Doç. Dr. Hasan Şakir BİLGE ve Sayın Doç. Dr. M. Ali AKCAYOL'a, çalışmalarım sırasında manevi desteğini esirgemeyen nişanlım Asuman SAVAŞÇIHABEŞ'e ayrıca maddi manevi her türlü destekleriyle beni hiçbir zaman yalnız bırakmayan çok değerli aileme teşekkür ederim.

İÇİNDEKİLER

Sayfa

ÖZET	iv
ABSTRACT	v

TEŞEKKÜR.....	vi
İÇİNDEKİLER	vii
ÇİZELGELERİN LİSTESİ.....	xi
ŞEKİLLERİN LİSTESİ	xiii
SİMGELER VE KISALTMALAR.....	xv
1. GİRİŞ	1
2. DİL TANIMA SİSTEMLERİ.....	3
2.1. Literatür Çalışması	3
2.2. Metotların Analizi	11
2.2.1. N-gram analiz.....	11
2.2.2. Kısa sözcük metodu.....	12
2.2.3. Bayes metodu	12
2.2.4. Yapay zeka ve istatistiksel metotlar	14
3. YAPAY ZEKA TEKNİKLERİ.....	15
3.1. Yapay Sinir Ağları (YSA) Yapıları	15
3.2. YSA Öğrenme Algoritmaları.....	17
3.2.1. Levenberg-Marquardt Algoritması	17
3.2.2. Geri yayılım algoritması.....	20
3.2.3. Momentumlu geri yayılım algoritması	20
3.2.4. Esnek yayılım algoritması	22
	Sayfa
4. DİL TANIMA BÖLÜMÜ.....	24
4.1. Geliştirilen Birinci Yaklaşım.....	24

4.1.1. Geliştirilen Modüller	26
4.1.2. Yazılım platformu uygulaması	29
4.2. Geliştirilen İkinci Yaklaşım	34
4.2.1. Veri Kümeleri	36
4.2.2. Metot – 1: Birleşim tespit yöntemi	39
4.2.3. Metot – 2: Kesişim tespit yöntemi	40
4.2.4. Geliştirilen modüller	42
4.2.5. YSA modelinin tasarlama aşamaları	47
4.2.6. Geliştirilen platform yazılımı	54
4.3. Test 1	60
4.4. Test 2	78
5. SONUÇ VE ÖNERİLER	96
KAYNAKLAR	101
EKLER	106
EK-1 Almanca dili kesişim veri kümesi	107
EK-2 Arnavutça dili kesişim veri kümesi	109
EK-3 Fransızca dili kesişim veri kümesi	110
EK-4 Galce dili kesişim veri kümesi	111
EK-5 Hırvatça dili kesişim veri kümesi	112
EK-6 İngilizce dili kesişim veri kümesi	113
EK-7 İrlandaca dili kesişim veri kümesi	114
EK-8 İspanyolca dili kesişim veri kümesi	115
EK-9 İtalyanca dili kesişim veri kümesi	116
EK-10 Letonca dili kesişim veri kümesi	117
EK-11 Macarca dili kesişim veri kümesi	118
EK-12 Maltaca dili kesişim veri kümesi	119

Sayfa

EK-13 Portekizce dili kesişim veri kümesi	121
EK-14 Türkçe dili kesişim veri kümesi	122
EK-15 Vietnamca dili kesişim veri kümesi	123
EK-16 Birleşim tespit 1	124

EK-17 Birleşim tespit 2.....	125
EK-18 Kesişim tespit 1	126
EK-19 Kesişim tespit 2	128
EK-20 Genel alfabe kümesi - 1	130
EK-21 Genel alfabe kümesi – 2	131
EK-22 Genel alfabe kümesi – 3	132
EK-23 Genel alfabe kümesi – 4	133
EK-24 Genel alfabe kümesi – 5	135
EK-26 Almanca dili GENIUS test sonuçları	140
EK-27 Arnavutça dili GENIUS test sonuçları	141
EK-28 Fransızca dili GENIUS test sonuçları	142
EK-29 Galce dili GENIUS test sonuçları.....	143
EK-30 Hırvatça dili GENIUS test sonuçları	144
EK-31 İngilizce dili GENIUS test sonuçları	145
EK-32 İrlandaca dili GENIUS test sonuçları	146
EK-33 İspanyolca dili GENIUS test sonuçları.....	147
EK-34 İtalyanca dili GENIUS test sonuçları	148
EK-35 Letonca dili GENIUS test sonuçları	149
EK-36 Macarca dili GENIUS test sonuçları	150
EK-37 Maltaca dili GENIUS test sonuçları	151
EK-38 Portekizce dili GENIUS test sonuçları	152
EK-39 Türkçe dili GENIUS test sonuçları.....	153
EK-40 Vietnamca dili GENIUS test sonuçları	154
EK-41 GENIUS platformu Microsoft Word ortalama başarı sonuçları.....	155
EK-42 GENIUS platformu Adobe PDF ortalama başarı sonuçları	156
EK-43 GENIUS platformu web sitesi HTML ortalama başarı sonuçları	159
ÖZGEÇMİŞ	160

ÇİZELGELERİN LİSTESİ

Çizelge	Sayfa
Çizelge 4.1. Anahtar kelime tablosu	37
Çizelge 4.2. YSA parametrelerinin karşılaştırma	51
Çizelge 4.3. Almanca için doküman bazında başarı oranları	64
Çizelge 4.4. Arnavutça için doküman bazında başarı oranları.....	65
Çizelge 4.5. Fransızca için doküman bazında başarı oranları.....	66
Çizelge 4.6. Galce için doküman bazında başarı oranları.....	67
Çizelge 4.7. Hırvatça için doküman bazında başarı oranları	68
Çizelge 4.8. İngilizce için doküman bazında başarı oranları	69
Çizelge 4.9. İrlandaca için doküman bazında başarı oranları	70
Çizelge 4.10. İspanyolca için doküman bazında başarı oranları.....	71
Çizelge 4.11. İtalyanca için doküman bazında başarı oranları.....	72
Çizelge 4.12. Letonca için doküman bazında başarı oranları	73
Çizelge 4.13. Macarca için doküman bazında başarı oranları	74
Çizelge 4.14. Maltaca için doküman bazında başarı oranları	75
Çizelge 4.15. Portekizce için doküman bazında başarı oranları	76
Çizelge 4.16. Türkçe için doküman bazında başarı oranları.....	77
Çizelge 4.17. Vietnamca için doküman bazında başarı oranları.....	78
Çizelge 4.18. Almanca için web sitesi bazında başarı oranları.....	81
Çizelge 4.19. Arnavutça için web sitesi bazında başarı oranları.....	82
Çizelge 4.20. Fransızca için web sitesi bazında başarı oranları.....	83

Çizelge 4.21. Galce için web sitesi bazında başarı oranları.....	84
Çizelge 4.22. Hırvatça için web sitesi bazında başarı oranları	85
Çizelge	Sayfa
Çizelge 4.23. İngilizce için web sitesi bazında başarı oranları	86
Çizelge 4.24. İrlandaca için web sitesi bazında başarı oranları	87
Çizelge 4.25. İspanyolca için web sitesi bazında başarı oranları.....	88
Çizelge 4.26. İtalyanca için web sitesi bazında başarı oranları.....	89
Çizelge 4.27. Letonca için web sitesi bazında başarı oranları	90
Çizelge 4.28. Macarca için web sitesi bazında başarı oranları	91
Çizelge 4.29. Maltaca için web sitesi bazında başarı oranları	92
Çizelge 4.30. Portekizce için web sitesi bazında başarı oranları	93
Çizelge 4.31. Türkçe için web sitesi bazında başarı oranları.....	94
Çizelge 4.32. Vietnamca için web sitesi bazında başarı oranları.....	95

ŞEKİLLERİN LİSTESİ

Şekil	Sayfa
Şekil 3.1. Genel bir YSA yapısı	16
Şekil 4.1. Yaklaşım aıt Blok Diyagram.....	25
Şekil 4.2. Bilimsel yaklaşım için tasarlanan YSA modeli	28
Şekil 4.3. Yazılım Geliştirme Süreci	30
Şekil 4.4. Geliştirilen yazılım platformu arayüzü	33
Şekil 4.5. Geliştirilen yazılım platformuna aıt blok diyagram.....	35
Şekil 4.6. Birleşim ve Kesişim Yöntemi Blok Diyagramı	39
Şekil 4.7. Kullanılan YSA Modeli	52
Şekil 4.8. YSA Eğitimi Performans vs. Ağırlık Değişim Sayısı Grafiği	53
Şekil 4.9. GENIUS Gazi Mühendislik Zeki Birleşik Servisi (Gazi Engineering Intelligent Unified Service)	55
Şekil 4.10 Word ve PDF dokümanları için sistem modeli blok diyagramı	62
Şekil 4.11 Web sayfaları için sistem modeli blok diyagramı	80

SİMGELER VE KISALTMALAR

Bu çalışmada kullanılmış bazı kısaltmalar, açıklamaları ile birlikte aşağıda sunulmuştur.

Kısaltmalar	Açıklamalar
BTY	Birleşim Tespit Yöntemi
ÇM	Çıkış Modülü
DGÇM	Dil Geçişli Çeviri Modülü
DT	Dil Tanıma
FAM	Frekans Analiz Modülü
GD	Dereceli Azalan Geri Yayılım (Gradient Descent Backpropagation)
HKNS	Her Katmanda Nöron Sayısı
HTML	Yardımlı Metin Biçimleme Dili (Hyper Text Markup Language)
HTTP	Yardımlı Metin Aktarım Protokolü (Hyper Text Transfer Protocol)
KS	Katman Sayısı
KTY	Kesişim Tespit Yöntemi
LM	Levenberg-Marquardt Algoritması
MLP	Multi Layer Perceptron (Çok Katlı Perseptronlar)
MOM	Metin Otomasyon Modülü
MSE	Ortalama Hataların Kareleri Toplamı (Mean Square Error)
RMS	Kare Ortalamalarının Karekökü
SDAM	Doküman Analiz Modülü
SDK	Yazılım Geliştirme Süiti (Software Development Kit)
SSE	Hataların Kareleri Toplamı (Sum Square Error)
TF	Transfer Fonksiyonu

YSA

Yapay Sinir Ağı

ZDTM

Zeki Dil Tanıma Modülü

1. GİRİŞ

İnternetin yaygınlaşması, dillerin ve kültürlerin internete aktarılmasını kolaylaştırır da bu kültürlerin geniş kitleler tarafından algılanması ve öğrenilmesi için o dillerin yaygın olarak kullanılan dillere dönüşümüne ihtiyaç vardır. Bu süreçleri etkileyen en önemli faktör ise doğal dillerin varlığıdır. Bugün farklı coğrafyalar üzerinde yüzlerce dil ve binlerce lehçe konuşulmaktadır. Konuşulan birçok dilin kendi alfabesi ve kuralları bulunmaktadır. Bu kurallar bütünü, dili ayırt edilebilir kılmasının yanında yüzlerce kuralın ve alfabenin bilinmesinin ve anlaşılmasının gerekliliği insanoğlu için büyük bir engeldir. İnternet ortamında farklı kültürlere ait dokümanlar dünya kültürüne katkı sağlıyor olsa da bu dillerin internet ortamını kullanan kişilerin o dili bilmemesi ise bu zengin kültürün anlaşılmasını zorlaştırmakta ve bu zenginliği diğer toplumların kullanımına pekte sunamamaktadır.

İnternet üzerinden her an büyük ölçüde veri akışı bulunması bilgi otobanları üzerinden sağlanmaktadır. Bu otobanlar üzerinden akan bilginin içeriğini saptamak büyük önem arz etmektedir. Bilgiye erişmek için önümüzdeki en büyük engel küreselleşmenin yarattığı çok ulusluluk sürecidir. Bu sürecin sonucu olarak uluslararası kabul gören İngilizcenin yanında internet ortamında her dile ait sayısal belgelerin varlığı günden güne artmaktadır. Buna rağmen milletlerarası iletişim bu artışa paralel olarak azalmaktadır. Bunun sebebi ülkelerin kendi dillerinde ürettikleri bilgiye başka toplumların ulaşmasında yaşadıkları güçlüklerdir. Bu güçlükleri ortadan kaldırmak için internet ortamında bulunan sayısal belgelerin (HTML, PHP, ASP, DOC, PDF vb.) ait oldukları dile ait özniteliklerin bulunması gerekmektedir.

Bir sayısal belgenin dilini tanımak pek çok uygulama alanında yararlıdır. Gördüğü işlevin tabiatı gereği genellikle bir *dil tanıyıcı* (DT) daha büyük bir uygulamanın alt modülü olarak çalışır. Bir ofis uygulama programında özellikle kelime işleme yazılımında DT tüm dokümanın ya da farklı dillerden oluşan bölümlerin dillerini otomatik tanımda kullanılabilir. Bir çeviri yazılımında kaynak ve hedef dil kullanıcı tarafından sisteme girilmektedir. Çevrilmek istenilen dokümanların dillerinin

bilinmemesi çeviri sürecini zorlaştırmaktadır. Bu tez çalışmasında önerilen yöntemle kullanıcılar dillerini hiç bilmedikleri sayısal dokümanları arzu ettikleri dile otomatik olarak dönüştürülebileceklerdir.

İnternette, dönüştürme işlemi için Microsoft, Altavista, Systran, Babylon ve Google gibi firmaların dil dönüştürme çözümleri sunulmaktadır. Bu çözümlerden hızlı bir büyüme süreci içinde olan Google firması tarafından geliştirilen Google Translator sistemi en kapsamlı çeviri çözümüdür. Bu çeviri sistemi sayesinde 35 farklı dili farklı dillere dönüştürebilme desteği vermektedir. Google tarafından geliştirilen yöntem bugüne kadar kullanılan metin bilgisinden dilbilgisi öğeleri elde ederek çevirme yaklaşımını kullanmamaktadır. Google tarafından geliştirilen yöntem sayesinde bir veritabanı diller arasında önceden yapılmış ve halen yapılmakta olan birebir çevirileri cümle yapısıyla saklamaktadır. Google Translator tarafından alınan bir giriş o girişe karşılık gelen veya en çok yakınsayan direkt çeviriye dönüştürülmektedir.

Google Translator ve benzeri çözümlerin sunduğu bu hizmetler pek çok işin yapılması için önemli çözümler olsa da dillerin otomatik tanımı yapılamamakta ve çeviri işlemi sadece html vb. formatları kapsamaktadır. Bu çalışmada web tabanlı olarak geliştirilen bir platform sayesinde Word, Text, PDF ve HTML dokümanlarının dil içeriği otomatik olarak algılanmakta ve dil çevirisi mevcut çözümler kullanılarak otomatik olarak yapılmaktadır.

Bu çalışmanın internetin daha verimli kullanılmasına, farklı dil ve kültürlerde yapılan çalışmaların kolaylıkla okunup öğrenilmesine büyük katkılar sağlayacağı, internette karşılaşılan pek çok problemin çözümüne katkılar sağlayacağı değerlendirilmektedir.

Bu çalışmada Almanca, Arnavutça, Fransızca, Galce, Hırvatça, İngilizce, İrlandaca, İspanyolca, İtalyanca, Letonca, Macarca, Maltaca, Portekizce, Türkçe, Vietnamca dilleri dikkate alınarak önerilen birleşim ve kesişim yöntemleriyle WORD, PDF ve HTML dokümanları dikkate alınarak dokümanların içerik dilinin otomatik olarak algılanması ve istenilen dile dönüştürülmesi için bir sistem geliştirilmiştir.

2. DİL TANIMA SİSTEMLERİ

2.1. Literatür Çalışması

Literatürde pek çok dil tanıma metotları bulunmaktadır [1–15]. Bu metotlar incelendiğinde n-gram analiz gibi istatistiksel metotların sıklıkla kullanıldığı fakat bu metotların ön işlem gibi performans kaybına neden olan süreçler içermesinden dolayı beklenen performansı veremedikleri görüldüğünden yeni metotların araştırılması ve geliştirilmesi gündemde olan bir husustur. Geliştirilen metotların kullanıldığı ve karşılaştırıldığı çalışmalar aşağıda farklı paragraflarda incelenmiştir.

Padro ve Padro tarafından 2004 yılında yapılan çalışmada [1], istatistiksel yöntemlere dayalı üç farklı dil tanıma yöntemi incelenmiştir. Basit parametrelerden bazıları sınıflandırma için gerekli olan kelime miktarları ve tanınması gerekli olan dil sayısı için uygun eğitim kümesi oluşturma önem taşımamaktadır. Yapılan çalışmalar eğitim küme büyüklüğünün 50 kelimeyi geçmesinin yeterli olduğunu göstermiştir. Önemli sayılabilecek parametrelerden birisinin metin uzunluğu olduğu yine bu çalışmanın bulguları arasındadır. İncelenen yöntemlerde ancak 500 karakter ve daha uzun metinlerin tanınması %95 başarıyla gerçekleşmektedir. Yapılan testlerde metin uzunluğu 5000 ve üstüne çıkartıldığında sonuçların %99 başarılı olduğu gösterilmiştir. Başarı oranı metin boyutu kısaldıkça doğru orantılı olarak azalmaktadır. Kısa metinlerde kullanılan yöntemlerin hepsinde başarı oranı bazen %60 seviyesine kadar gerileme göstermiştir [1].

Botha ve arkadaşlarının [2] 2007 yılında yaptığı çalışmada metin tabanlı 11 farklı Güney Afrika dili ele alınarak dil tanıma süreci araştırılmış ve n-gram istatistikleri sınıflandırma yapmak için kullanılmıştır [2]. Bu süreç yönetiminde özellikle destek vektör makineleri (support vector machines) ve yakın komşuluk tabanlı sınıflandırıcılar (likelihood-based classifiers) karşılaştırılmıştır. Tanıma işlemi için giriş olarak uygulanan metinlerden az sayıda kelime ile tanıma işleminin mümkün

olduğu belirlenmiştir. Destek vektör makineleri genel olarak daha başarılı sonuçlar vermesine rağmen sınıflandırma için gerekli olan yüksek işlem karmaşıklığı sebebiyle ancak basit ve kararlı bir sistem tasarımı için uygun olmadığı gösterilmiştir. Yapılan n-gram testlerinde en iyi çalışma performansını veren n-gram analizi için gerekli olan n değeri $n=6$ olarak bulunmuştur. Hollandaca, Danimarka'ca, İngilizce, Fransızca, İtalyanca ve İspanyolca dillerinin incelendiği bir çalışmada çeşitli (2-7 arası) büyüklüklerde n-gram analizlerinin ortalama hata oranları incelendiğinde 100 karakterli bir giriş seti için hata oranının %1.03 olduğu gösterilmiştir. Yapılan bu çalışmada ise 15 karakterli bir giriş seti için hata oranı %0,6 olarak bulunmuştur. Botha ve arkadaşları tarafından yapılan diğer bir çalışmada ise Güney Afrika dillerinin Avrupa dillerine göre tanınma oranının daha düşük olduğu savunulmuş olmasına rağmen hata oranları olmadığı için bu çalışma ile birebir karşılaştırma yapılamamıştır. [2] nolu çalışmada yapılan testlerde daha büyük giriş setleri ile doğrusal olarak daha kesin sonuçlar elde edilmekte ve destek vektör makinelerinin karmaşıklık probleminin sistemi karasızlığa götürdüğü rapor edilmiştir. Aynı çalışmada bahsedilen diğer metotların her dil grubu için oluşturulması gereken eğitim seti büyüklüğü belirtilmiş ve elbette aynı etnik kökeni paylaşan iki farklı dil içinde geçebilecek aynı kelimeler olabileceği için yakın komşuluk tabanlı sınıflandırıcıların büyük miktarda giriş verisi ile elde edilen sonuçlar sunulmuştur [2].

El-Shishiny ve arkadaşları [3] tarafından 2004 yılında yapılan diğer bir çalışmada Arapça kökenli diller ve diğer diller arasında Arapça betiklerin kelime parçalama tabanlı bir metot ile nasıl ayrıştırılabileceği araştırılmıştır. Bu yöntemin geliştirilmesi sırasında Arapça betiklere ait karakter kümesi olarak Arapça dilinde kullanılan kelimeler, ön ve son ekler ile unigramlar kullanılmıştır. Bu çalışmada internet kaynaklarından rastgele seçilen Arapça, Farsça, Urduca, Afganca ve Uygurca gibi Arapça betik tabanlı dillerden oluşan toplam 180 farklı örnek kullanılmıştır. Bu dillerden Arapça betiklerin tanınması ile %94 oranında dil tanıma başarısı elde edilmiştir. Bu her dil için mevcut dilbilgisi kurallarının çözümlenmesi ve sistemin bu kuralları kullanabilecek şekilde yapılandırılması gibi zorluklar taşımaktadır [3].

Kruengkrai ve arkadaşları [4] tarafından 2005 yılında yapılan başka bir çalışmada otomatik dil tanıma için karakter dizisi çekirdeği (string kernels) kavramı tabanlı yeni bir yöntem önerilmiştir. Önerilen yöntemde kodlama şekline rağmen, tanınması istenilen dil metin içinden doğrudan çıkartabilmektedir. Bu işlem yapılırken dikkat edilmesi gereken husus metin bilgisinin kodlama bilgisinin bir karakter dizisinin byteleri şeklinde kodlanmasıdır, çünkü mevcut kodlama örüntüleri incelendiğinde çoğu sayfa için kullanılan ISO-8859-1 kodlama standardı 19 farklı Avrupa dilini desteklemektedir. Bu genellemenin geçerliliğini kaybetmesiyle birlikte dil tanımanın bir sınıflandırma problemi olduğu görülmektedir. Sistemin kararlı bir şekilde çalışması için yapılması gereken ilk işlemin eğitim setinde her dile ait karakteristik yapının çıkartılmasının gerekliliği bu çalışmada vurgulanmıştır. Sonuçlar yeterli sayıda giriş setiyle 17 farklı dil için değerlendirilmiş ve elde edilen test sonuçlarında diller arası yakınsamanın azaltılması için eşleşen sözcük öbeklerine göre ağırlık atama yönteminin kullanılmasının gerektiği sonucuna varılmıştır. Yapılan testler dil tanıma sorunu için kararlı bir metot oluştuğunu göstermesine rağmen performans ve hız konularında sistemin uzaysal karmaşıklığı nedeniyle gerekli sonuçlara bu çalışmada yer verilmemiştir [4].

Zavarsky ve arkadaşları [5] tarafından 2005 yılında yapılan çalışmada kodlama örüntüleri karşılaştırılarak çok yüksek sayıda giriş setlerine ait dil tanıma işlemlerinin sonuçları sunulmuştur. Bu çalışmada kullanılan büyük giriş kümeleri (milyonlarca) Dil İzleme Projesinin bir ürünü olarak ortaya çıkmaktadır. Bu proje düzinelerce milyondan fazla metin tabanlı dokümandan oluşmaktadır. Önerilen yöntem tek bir Linux bilgisayarda her saniye 250 günde 20 milyon dokümanı inceleyebilmektedir. Bu sayede çok işlemcili Linux sunucularda analiz doküman sayısı günlük 100 milyona ulaşabilmektedir. Doğal dillerde yazılan metinlerin tanıma probleminde metinlerin karakter kodlama örüntülerinin belirlenmesi zor bir sorun değildir. Şayet bir doküman birden fazla dil ile yazılmamış ise yeteri kadar uzun olması kabul edilebilir bir sürede belirli sayıda dokümanın analiz edilmesini sağlamaktadır. Bu amaçla sıklıkla kullanılan yöntemler kelime tabanlı metotlardır. Elbette bu metotlar

temelde karakter kodlaması bakımından etkisizdir ve daha önceden sınıflandırma bilgisi eğitilmiş herhangi bir dili tanıyabilirler. Kaydetme yapısı, tanınması beklenen dil ve kodlama örüntüsü için en uygun şekilde ve yeniden kullanılabilirlik ile çok işlemli programlama kavramı göz önüne alınarak önerilmiştir. N-gram ve kelime tabanlı araçları kullanıcılar tanıtmak istedikleri dil için eğitirler. Kullanıcı her dil için gerekli olan materyali toparlayarak eğitim sürecini gerçekleştirir. Bu araçların birçoğu sayısal dokümanlarda sıklıkla tercih edilen Avrupa dilleri ve bazı Asya dilleri için çalışmaktadır. Bu yüzden sisteme giriş olarak uygulanan dokümanların çeşitli kriterleri sağlaması gerekmektedir [5].

Peng ve arkadaşları [6] tarafından 2003 yılında yapılan çalışmada dil bağımsız ve işlem bağımsız metin sınıflandırması ile karakter tabanlı n-gram dil modelleri kullanan bir metot önerilmiştir. Bu yaklaşım en genel haliyle basit teorik ilkeler kullanan ve birçok çeşit dil sınıfında başarıyla çalışan bir yapıdır. Bu yapının kurgulanması için öznitelik vektörlerinin seçimine yada büyük ön işlemlere gerek yoktur. Bu sistemin çalışmasını ve bağımsızlığını sınamak amacıyla Yunanca, İngilizce, Çince ve Japonca dilleri dil tanıma, metin sınıflandırma, yazar nitelme, yazı türü sınıflandırma ve konu tespiti gibi çözülmesi zor sorunlar ile test edilmiştir. Önerilen bu yaklaşım ile Asya kökenli dillerde sık karşılaşılan bir sorun olan kelime parçalama probleminin karakter seviyesinde dil modelleriyle çözülmesi sağlanmıştır. Bu çalışmada işlem bağımsız metin sınıflandırması için karakter seviyesinde n-gram modelleri kullanılmıştır. Önerilen yöntem dört farklı dil için 4 farklı metin sınıflandırma problemine uygulanmıştır. Yapılan testlerde müzik veya DNA gibi ardışık veriler ile çoklu kategori sınıflandırması problemi (Reuters-21578 veri kümesi) gibi sorunların çözümünde önerilen yaklaşımın başarısı ortaya konulmuştur. Bu çalışmanın kötü yanları ise kelimedenden karakter seviyesine kadar parçalama sorunu nedeniyle işlem hızının düşük olmasıdır [6].

Nair ve arkadaşları [7] tarafından 2007 yılında yapılan çalışmada yazı dili İngilizce olmayan ülkelerin kendilerine özgü dil gruplarına karışan Latin kökenli metinlerin ayrıştırılarak gizli Markov Modelleri ile bu metinlere ait dillerin analizleri

incelenmiştir. Bu metinlerin resmi olmayan yazı dili Hintçe ve İngilizce anlamına gelen Hinglish olarak anılmaktadır. Bir başka örnek ise Malezya'dır. Bu ülkenin resmi olmayan yazı dili Manglish yani Malezya'ca ve İngilizce anlamına gelmektedir. Bu İngilizce dil gürültüsü yazı dili İngilizce olmayan ülkelere ait metinlerin ayrıştırılmasında büyük güçlükler neden olmaktadır. Bu çalışmanın konusu gizli Markov Modellerinin İngilizce olan ve olmayan metinleri ayrıştırabilmesi için bir sınıflandırma yapısının geliştirilmesidir. Bu sayede değişik dil modelleri ele alınarak oluşturulan metin editörleri ile dile bağlı renklendirme benzeri bir ayrıştırma aracı geliştirmektir. Geliştirilen yaklaşımın sözlük tabanlı metotlarla karşılaştırıldığında zaman ve uzay karmaşıklığı bakımından üstün olduğu görülmüştür. Sunulan modelin her dil için test edilememesinden dolayı modelin iyileştirilmesi gerekmektedir [7].

Ahmed ve arkadaşları [8] tarafından 2004 yılında yapılan çalışmada etkili dil sınıflandırma için n-gram tasarsız birikimli frekans ekleme metodu kullanılmıştır. Bu yeni sınıflandırma tekniği ile geleneksel Bayesian sınıflandırma metotlarına göre çok daha basit uygulanmaktadır buna rağmen sınıflandırma süreci kısa kelime parçalarında hız ve etkinlik olarak aynı performansı vermektedir. Sınıflandırma hızı öncelik sıralamalı n-gram sınıflandırma metotlarına göre 5-10 kat daha hızlı çalışmaktadır. N-gram sınıflandırma metotlarıyla dil tanıma süreci bölgesel farklılıklardan etkilenmeyen yüksek performansla çalışan literatürde yer edinmiş ve üstüne uzun araştırmalar yapılmış bir sistemdir. Bu sistemin en büyük dezavantajı sınıflandırmada kullanılan öncelik tabanlı metodun hızının düşük olmasıdır. Bu hız düşmesine neden olan şey sınıflandırma için test dokümanında frekans sayma ve sıralama değerlerinin bulunmasının gerekliliğidir. Bu çalışmada birikimli frekans ekleme metodu ile bu performans artışının sağlanması hedeflenmekteyse de uzun kelimeler ve paragraflar içinde geliştirilmesi gerekmektedir [8].

Constable [9] tarafından 2002 yılında çalışmada yeni teknik ve standartlarının geliştirilmesinin ihtiyacı üzerinde durulmuştur. Elbette yeni kullanım senaryolarına tam uyacak bir modelin geliştirilmesinde çeşitli zorluklar yaşanacaktır. Bu çalışma

dil modeli ve dil içerikli kategorilerde gereksinim duyulan varlık bilimsel bir modelin eksikliğini gidermeyi hedeflemektedir. Bu noktada var olan çalışmalarda öne çıkan iki unsur dil ve yerelleştirme konunun özünü oluşturmaktadır. Mevcut problemlerin çözülebilmesi için dil tanıma sistemlerinde olması gereken bazı özelliklerin yazı sistemleri, varlık bilim, alan tanımlı veri kümesi, lehçeler ve diğer alt dil varyasyonları, üst dil bilim kategorileri, tarihsel dil değişimleri, dil bağımlı kategorileştirme ve yerelleştirme ile tüm bu sınıfların tartışıldığı konu başlıklarıyla bir inceleme yapılmıştır [9].

Adams ve arkadaşları [10] tarafından 1997 yılında yapılan çalışmada Java programlama dili platformunda tasarlanmış deneysel bir sistem tanıtılmıştır. Bu sistemin Java platformuyla geliştirilmesinin sebebi kullanıcı tarafı web göstericisinde çalışabilen doğal dil işleme sistemlerinde kullanılması için gerekli uygulama seviyesi dağıtık tabanlı bileşenlerinin Java dili içinde bulunmasıdır. Bu sayede doküman sunucusu yada uzman sistem ajanları tarafından uyumlu çalışabilmektedir. Bu çalışmanın oluşturulmasında dil tanıma sistemlerinde sıklıkla kullanılan n-gram analizleri kullanılmıştır. N-gram analizlerde ihtiyaç duyulan etiketlendirme ve metin parçalama bilgisi dinamik olarak oluşturulmaktadır. Uygulama sonucunda elde edilmek istenilen sonuçlar dil modelleri arasında etiketlendirme performansı ile işlem bütünlüğünü arasındaki köprünün kurulmasıdır. Dil etiketleri tüm doküman için yada doküman grupları için öge tabanlı yapısal bir bileşen olarak ele alınmıştır. Sonuç olarak bu çalışmada karakter tabanlı etiketlendirme ile yapılan n-gram analizlerine yer verilmiştir. Bu modülasyon sayesinde internetten bağımsız çalışan bir yapı ile internet vekil sunucusu tabanlı iki sistem önerilmiştir. Bu çalışma sayesinde konuşma etiketleme, tümce tanımlama, yabancı kelime çevirisi ve konu etiketleme gibi zorlukların aşılmasıyla birlikte web tabanlı zeki arama ve gösterimleri gibi konuların işlenebileceği önerilmektedir [10].

Ölvecky [11] tarafından 2005 yılında yapılan çalışmada internetin büyümesiyle birlikte dil tanıma ve metin sınıflandırmanın öneminin artmasıyla birlikte bu alanda kullanılan n-gram varyasyonları incelenmiştir. İnternet üzerinde artan

dokümantasyona karşılık olarak bu içeriklere özensizlikten meydana gelen metin ve gramer tabanlı hatalar dil tanıma işini zorlaştırmaktadır. Bu metotlar küçük değişiklikler ile metinlerden dil analizini %95 başarı oranıyla gerçekleştirebilmektedirler. Bu çalışmada tartışılan konu n-gram sınıflandırma tekniğinin bilinmeyen metinlerde sınıflandırma yeteneğinin yüksek kesinlikte bilinmesi için yapılması gerekenlerdir. Bu işlemlerden bir tanesi n-gram için kullanılan metin boyutunun azaltılmasıdır, uzun metinler eşit parçalar halinde kesilerek n-gram analiz büyüklüğü azaltılmaktadır. Bu alanda tartışılan bir başka konu ise n-gram giriş profili ile sistem eğitim büyüklüğünün ve yazım kalitesinin nasıl iyileştirilebileceği gösterilmiştir. Bu sayede tanıma performansının %10 kadar artış gösterdiği savunulmaktadır [11].

Bilcu ve arkadaşları [12] tarafından 2006 yılında yapılan çalışmada yazınsal metinlerden dil tanınması yapabilecek yapay sinir ağı tabanlı hibrit bir metot önerilmiştir. Bu problem metinden ses birimine (fonem) bağlı bir çalışmadır. Bu problemde amaç yazılı bir metne ait harflerin kendilerine karşılık gelen ses birimlerine dönüştürülmesidir. Bu süreçte şayet dönüştürülmek istenen tek bir dil varsa dil tanıma işlemine gerek yoktur, ilgili yazılı metin direkt olarak ses birimlerine dönüştürülmesine rağmen giriş dili bilinmeyen metinlerde genellikle ilk aşama metin dilinin tespit edilmesidir. Bu sistemin geliştirilmesi sürecinde çok katmanlı geriye yayılım algoritması ve basit karar verme parametrelerine dayalı kural tabanlı bir sistem kullanılmıştır. Bu uygulama özellikle İngilizce ve Fransızca dillerinin tanınması amacıyla geliştirilmiştir. Bu sayede İngilizce veya Fransızca dillerinde yazılmış metinlerin ayrıştırılarak tanınan dil özelliklerine göre metinden ses birimlerine çevrim gibi modüler bir yapı tasarlanmıştır. Bu çalışma sayesinde iki dile ait metinlerden ses birimi hecelerine geçiş sürecinin sağlanabileceği sistemin bu doğrultuda geliştirilmesi ve mevcut sorunların çözümlenmesiyle sistemin daha kararlı bir şekil alacağı savunulmuştur [12].

Liu ve arkadaşları [13] tarafından 2005 yılında yapılan çalışmada önerilen sistem dil tanıma problemini karakterlerin resimlerinden tanımayı hedefleyen, birden çok dil

barındıran dokümanlarda dil sınırlarının belirlenmesi için ilgili karakterlere ait resim verilerinin sınıflandırılması için yeni bir yöntem önerilmiştir. Bu amaçla prototip bir sınıflandırma yöntemi ve destek vektör makineleri kullanılmıştır. Büyük boyutlu eğitim veri kümesi nedeniyle her iki metot için eğitim hızını artıracak yeni bir teknik önerilmiştir. Bu metotlar kullanılarak Çince, İngilizce ve Japoncanın (Hiragana ve Katakana dahil) dillerinin tanınması hedeflenmiştir. Elde edilen sonuçlar, oldukça kararlı ve literatürde kabul görmüş yöntemlerle karşılaştırılabilecek düzeydedir [13].

Zhu ve arkadaşları [14] tarafından 2008 yılında yapılan çalışmada el yazısı ve makine çıktısı yazılardan dil tanınmasına üzerine bir yöntem önermişlerdir. Bu yöntemin gerçekleşmesi için çok sayıda gerçek içerikli dokümandan oluşan resim koleksiyonları kullanılmıştır. Bu dokümanlar 8 dilde 1500 kadardır. Bu diller Arapça, Çince, İngilizce, Hintçe, Japonca, Korece, Rusça ve Tayca ile bu dillere ait hem el yazısı hem de makine çıktısı içeriklerden oluşmaktadır. Yapılan testler sistemin yüksek karmaşıklıkta dokümanlar için iyi bir tanımlayıcı olduğunu göstermesine rağmen kod çizelgelerinin oluşturulmasında karşılaşılan uzaysal karmaşıklık gibi sorunlar nedeniyle sistemin uygulama aşamasında bazı sorunların aşılması gerekmektedir [14].

Baykan ve arkadaşları [15] tarafından 2008 yılında yapılan çalışmada internet ortamında girilen tek bir adresten dil tanınmanın mümkün olduğuna ilişkin öneriler sunulmaktadır. Bu sayede belirli bir dile ait içeriklerin getirilmesini saplamayı amaçlayan internet tarama sistemleri ve arama motorları tarafından böyle bir sistemin yararlı olacağı savunulmuştur. Bu sistem gereksiz sayfaların taranmasını da engellemektedir. Sunulan, çalışmada birçok makine öğrenme algoritması değerlendirilmiş ve İngilizce, Fransızca, Almanca, İspanyolca ve İtalyanca için karşılaştırmalı testler yapılmıştır. Bu testler neticesinde en iyi metot F-ölçümü olmuştur. Bu yöntem ile tüm dillerde web tarayıcılarından elde edilen 25K boyutunda 1260 web sayfasında ortalama %90 oranında başarı elde edilmiştir. 5 K büyüklüğündeki internet arama motorlarından elde edilen başarı oranı ise. 96 olmuştur. Bu çalışmada kullanılan performans kriteri F-Ölçümü'dür. Bu ölçüm

makine dil tanıma başarı kriteri olarak anılmaktadır. Bu test süreçlerinde kullanılan sayfalar ODP verileridir. Bu internetin kategorilere göre taranması amacıyla geliştirilen ve dil bağımlı olarak yayınlanan bir veritabanıdır. Karşılaştırmalı algoritmalar olarak geleneksel Bayes metodu, maksimum entropi metodu kullanılmıştır. Bu çalışmada karşılaşılan en büyük sorun İngilizce gözüken web sayfası adreslerine bağlı içeriklerin İngilizce olmayan dillerde yazılmış olmasıdır. Kullanılan yöntem önceliklere ülkelere ait ccTLD kayıtlarının belirlenmesidir, bu her ülkeye ait olan ve alan adının sonuna o ülkeye ait tanımlayıcı karakter dizisinin gelmesi işlemidir. Bu aşamadan sonra her ülke grubu için 10 farklı durma kelimesi belirlenmektedir, bu durma kelimeleri her dil içinde en sık geçen kelimelerden seçilmektedir. Bu durma kelimelerinin 8 ila 10 tanesi 1'den 10'a kadar kombinasyonlarla arama motorlarında aratılarak sonuçlar işlenmektedir. Bu sisteme son olarak 97 milyon tekrarlanmayan sonuç içeren bir internet arama motorunda 1260 rastgele sayfa eklenmektedir. Bu testlerde kişilerden ODP veri kümesinde kategorize edilmemiş 100 sayfayı tanımlamaları istenmiş ve ancak %50 başarı elde edilmiştir, sistem testlerinde ise bu başarı oranı %85 olarak gözlenmiştir [15].

2.2. Metotların Analizi

2.2.1. N-gram analiz

Bir n-gram bir dizinin n karakter uzunluğundaki alt dizesidir. Karakterler kelimenin içerisinde mutlaka ardışık olarak sıralanmalıdırlar. Bununla birlikte literatürde, ender kullanılmakla birlikte karakterlerin keyfi sırayla seçilebildiği n-gram tanımlarına da rastlanabilir. N-gram metodu dokümandaki kelimeleri uç uca çakışan alt dizelere ayrılarak bu alt dizelerin bulunma olasılıkları ile dilin ayırt edilmesi için bir bilgi elde etmeye dayanmaktadır. Bir kelimedenden pek çok n-gram elde etmek mümkündür. Örneğin “araba” sözcüğünde örnek vererek gösterecek olursak;

Bigram → _a, ar, ra, ab, ba, a_

Trigram → __a, _ar, ara, rab, aba, ba_, a__

Quadgram $\rightarrow$ __a, __ar, _ara, arab, raba, aba_, ba_, a__

Genel olarak n karakter uzunluğundaki bir kelimeden $n+1$ adet bigram, $n+2$ adet trigram ve $n+3$ adet quadgram elde edilebileceğini söyleyebiliriz. Bu metot bir dilde, bazı kelimelerin diğerlerinden daha sık kullanıldığı çıkarımından faydalanır, bunlar da elde edilen alt dizilerle örtüşmektedir. Her harf grubunun görülme olasılığı bu harf gruplarının oluşturulduğu kelimelerin görülme olasılığı ile direkt olarak bağıntılıdır.

2.2.2. Kısa sözcük metodu

Kısa sözcük metodu (küçük kelime metodu da denir) n -gram analizine kelimelerin bütün olarak tanımlamada kullanılmasının dışında benzer. Adından da anlaşılabilirliği üzere bu kelimeler kısa kelimelerdir, genelde dört harften az karakter içerirler. Sıfatlar, bağlaçlar gibi her dokümanda mutlaka bulunan kelimeleri içerirler.

2.2.3. Bayes metodu

Bu metotta koşutlu olasılık dili tanımlamada belirleyici özelliktir. Gerçekleştirilen uygulamalarda, öğretim evresinde uzun dokümanlar mümkün olan en fazla kelimeyi içerebilecek bir veri tabanı oluşturmak üzere kullanılırlar. Bu veritabanlarındaki kelimelerin görülme sıklıkları da elbette dil için ayırt edici bir özelliktir. Öğrenme sürecinden geçtikten sonra bilinmeyen bir dilin analizi yapılacağı zaman veritabanında bulunan tüm diller için Bayes teoremini kullanarak bir olasılık hesaplaması yapılır. Bayesian Olasılık Kuramı, matematiksel istatistik kuramının bir dalıdır. Bu kuram; belirsizlik taşıyan herhangi bir durumun modelinin oluşturularak, bu durumla ilgili evrensel doğrular ve gerçekçi gözlemler ışığında belli sonuçlar üretilmesini sağlar. Olasılık kuramında en önemli kavramlardan birisi de koşullu olasılıktır [18].

Bayes teoremi bir süreç sırasında ortaya çıkan bir rastgele olay A ile bir diğer rastgele olay B (eğer B için kaybolmamış olasılık varsa) için koşullu olasılıkları ve uç olasılıkları arasındaki ilişkidir, dolayısıyla

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)}$$

(2.1)

Bayes teoremi formülü içinde bulunan her bir terime özel isimler verilmektedir:

- (i) $P(A)$ terimine A için önsel olasılık veya uç olasılık adı verilir. Bu önseldir, çünkü B olayı hakkında önceden herhangi bir bilgiyi içermemektedir.
- (ii) $P(A|B)$ terimi verilmiş B için A'nın koşullu olasılığı adını alır.
- (iii) $P(B|A)$ terimi verilmiş A için B'nin koşullu olasılığı adını taşır.
- (iv) $P(B)$ terimi B olayı için 'önsel' olasılıktır veya B'nin uç olasılığıdır ve matematiksel rolü normalize eden bir sabittir.

Açıklandığı gibi, Bayes teoremi eğer B gözlemlenmiş ise, A gözlemi hakkındaki inançların ne şekilde güncelleştirilebileceğini ortaya çıkartır [18].

Bayes teoremi olabilirlik terimleri ile de şöyle ifade edilebilir,

$$P(A|B) \propto L(A|b)P(A) \quad (2.2)$$

Burada $L(A|b)$ terimi verilmiş sabit b için A'nın olabilirliği ve $L(A|B)/P(B)$ orantısına bazen standardize edilmiş olabilirlik veya normalize edilmiş olabilirlik adı da verilir [18].

$$P(B|A) \propto L(A|B) \quad (2.3)$$

Eşitlik kullanarak Bayes teoremi ortaya çıkartılır.

Bu teoremi ispat etmek için koşullu olasılık tanımından başlanır. B olayı bilirse A olayının olasılığı şöyle verilir [18].

$$P(A|B) = \frac{P(A \cap B)}{P(B)} \quad (2.4)$$

Aynı şekilde A olayı verilmiş ise B olayının olasılığı şudur:

$$P(B|A) = \frac{P(B \cap A)}{P(A)} \quad (2.5)$$

Bu iki eşitlik yeniden düzenlenip birbirlerine birleştirilirse,

$$P(A|B)P(B) = P(A \cap B) = P(B|A)P(A) \quad (2.6)$$

Eşitliği bulunur. Bu lemma bazen olasılıklar için çarpım kuralı olarak anılmaktadır. Her iki taraf da P(B) (eğer sıfır değilse) ile bölünürse, ortaya çıkan şu ifade Bayes teoremidir [18].

$$P(B|A) = \frac{P(B|A)P(A)}{P(B)} \quad (2.7)$$

2.2.4. Yapay zeka ve istatistiksel metotlar

Bu metotla dokümanın kendisinden elde edilen bilgilerden çok karakter sıklığı analizi gibi basit istatistikî metotlardan yararlanılmaktadır. Metodun belirlenmesinde harf sıklığı analizinin dili tanımlayıcı bir unsur olarak oluşturulması planlanan yapay sinir ağı yapısının eğitiminde kullanılabileceği düşünülmüştür. Konuyla ilgili literatür taraması yapıldığında ise bu konuda daha çok yukarıda belirtilen metotların kullanıldığı, harf sıklığı analizi ile dil tanıma sürecinin incelenmediği görülmektedir.

3. YAPAY ZEKA TEKNİKLERİ

Yapay zeka araştırmacılarının baştan beri ulaşmak istediği hedef, insan gibi düşünen ve davranan sistemler geliştirmektir. Ancak buna ulaşmanın güçlüğü anlaşılınca çalışmanın yönü rasyonel düşünen ve davranan sistemlerin tasarlanmasına çevrilmiştir. Geleneksel yöntemlerle çözümü zor veya imkansız olan problemlerin çözümünde kullanılan, yapay zeka teknikleri, yapay sinir ağları (YSA), bulanık mantık, sezgisel (genetik, tabu arama, karınca koloni, ısıt işlem benzetimi, bağışıklık sistemi, arı) algoritmalar olarak sıralanabilir. Bu çalışmada YSA tabanlı sistemler geliştirildiği için bu bölümde sadece YSA'larla ilgili yapılar ve öğrenme algoritmaları ile temel bilgiler verilmiştir.

3.1. Yapay Sinir Ağları (YSA) Yapıları

YSA'lar yapılarına göre ileri beslemeli (feed forward) ve geri beslemeli (feed back) olarak ikiye ayrılır. İleri beslemeli ağlarda işaretler, giriş katmanından çıkış katmanına doğru tek yönde iletilirler. Geri beslemeli ağlarda, çıkış ve ara katman çıkışları kendinden önceki katmanlara ya da girişe geri beslenir. İleri beslemeli ağlara Çok Katlı Perseptronlar (MLP), Radyal Tabanlı Fonksiyon Ağı (RBFN) ve LVQ (Learning Vector Quantization) örnek verilebilirken, ART (Adaptive Resonance Theory), SOM (Self Organizing Maps) ve Elman ve Jordan ağları geri beslemeli ağlara örnek olarak verilebilir [16]. Bu YSA yapıları Şekil 3.1'de verildiği gibi ileri beslemeli standart birleşmeli yapıda olabileceği gibi geri beslemeli dinamik yapıda da olabilir.

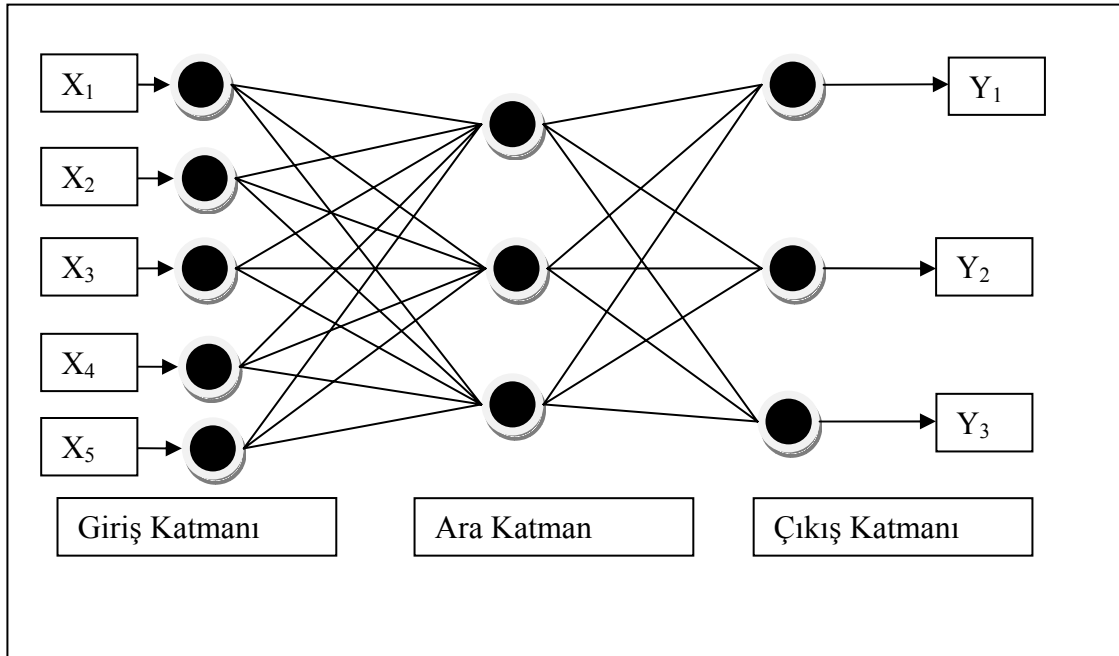
Bu çalışmada, ileri beslemeli ağ yapılarından olan MLP yapısı kullanılmıştır. MLP yapısının tercih edilmesinin nedeni, bilinen en eski YSA modellerinden olması çok farklı uygulamalarda başarılı olması ve sınıflandırma problemlerinde başarılı sonuçlar üretmesidir. Bunların yanında, farklı öğrenme algoritmaları ile kullanıma

uygun olması MLP'nin sağladığı diğer bir üstünlüktür. Şekil 3.1'de genel olarak gösterilen MLP modeli, bir giriş, bir veya daha fazla ara katman ve bir de çıkış katmanı içerir. Bir katmandaki bütün nöronlar bir sonraki katmandaki bütün nöronlara bağlıdır. Giriş katındaki nöronlar tampon gibi davranırlar ve giriş sinyalini ara katmandaki nöronlara dağıtırlar. Ara katmandaki her bir nöronun çıkışı, kendine gelen bütün giriş sinyallerini takip eden bağlantı ağırlıkları ile çarpımlarının toplanması ile elde edilir. Elde edilen bu toplam, çıkışın toplam bir fonksiyonu olarak hesaplanabilir. MLP'de giriş katmanı hariç bir nöronun çıkışı

$$y_k = f \left(\sum w_{kj} x_j \right)$$

(3.1)

ile hesaplanır. Burada, w nöronlar arasındaki ağırlık değerini, f ise kullanılan transfer fonksiyonunu gösterir.



Şekil 3.1. Genel bir YSA yapısı

3.2. YSA Öğrenme Algoritmaları

İleri beslemeli ağlarda kullanılan çok sayıda öğrenme algoritması bulunmaktadır [16]. Bu algoritmalar genel olarak performans fonksiyonunun minimizasyonu için ağırlık ve bias değerlerinin değiştirilmesinde performans fonksiyonunun eğimini kullanmaktadır. Eğim (gradyan) değeri ise geri yayılım adı verilen bir öğrenme algoritmasıyla belirlenmektedir. Geri yayılım algoritması ağırlık ve bias değerlerinin düzeltilmesi için ağ çıkışından ağ girişine doğru bir hesaplama yolu oluşturmaktadır [16].

MLP’de giriş ve çıkış sayıları, uygulanacak olan probleme göre belirlenirken, ara katman sayısı ile ara katman nöron sayıları “deneme yanılma” yolu ile bulunur [16]. YSA’larda diğer bir önemli etken ise öğrenme algoritmalarıdır. Öğrenme algoritmaları danışmanlı, danışmansız ve takviyeli olmak üzere üç farklı gruba ayrılabilir [16]. Geri Yayılım (BP), Levenberg-Marquardt (LM), Esnek Yayılım (RP), Delta-Bar-Delta (DBD) ve Hızlı Yayılım (QP) gibi algoritmalar danışmanlı öğrenme algoritmalarına, Adaptif Rezonans Teorisi (ART) ve Kendi Kendini Düzenleyen Haritalar (SOM) gibi yapılarda kullanılan algoritmalar ise danışmansız öğrenme algoritmalarına, Genetik Algoritmalar (GA) ise takviyeli öğrenme algoritmalarına örnek olarak verilebilir [16]. Bu tez çalışmasında literatürde pek çok çalışmada başarıyla uygulanan MLP YSA modelini eğitmek için LM (Levenberg-Marquardt), GDM (Gradient Descent with Momentum), GDX (Gradient descent with momentum and adaptive learning rate backpropagation) gibi öğrenme algoritmaları kullanılmıştır. Bu algoritmalar aşağıda kısaca açıklanmıştır.

3.2.1. Levenberg-Marquardt Algoritması

LM algoritması, maksimum komşuluk fikri üzerine kurulmuş bir en az kareler hesaplama metodudur [16]. Bu algoritma, Gauss-Newton ve En Dik İniş (Steepest Descent) algoritmalarının en iyi özelliklerinden oluşur ve bu iki metodun

kısıtlamalarını ortadan kaldırır. Genel olarak bu metot, yavaş yakınsama probleminden etkilenmez.

LM algoritmasında $E(w)$ 'nin bir amaç hata fonksiyonu olduğu düşünülürse, m adet çıkış nöronu için hata terimi $e_i^2(w)$ [16]:

$$(3.2)$$

olarak verilir.

Eş. 3.2'de w ağırlıkları ifade edilirken,

$$(3.3)$$

Burada, amaç fonksiyonu $f(.)$ ve onun Jakobiyesi J 'nin bir noktada w bilindiği farz edilir.

LM öğrenme algoritmasında hedef, parametre vektörü w 'nin, $E(w)$ 'yi minimum yapacak şekilde optimize edilmesidir. LM algoritmasının kullanılmasıyla yeni vektör w_{k+1} , farz edilen vektör w_k 'dan Eş. 3.3 yardımıyla aşağıdaki gibi hesaplanabilir.

$$w_{k+1} = w_k + \delta w_k \quad (3.4)$$

burada δw_k ifadesi

$$(J_k^T J_k + \lambda I) \delta w_k = -J_k^T f(w_k) \quad (3.5)$$

eşitliğinden faydalanılarak hesaplanır. Eş. 3.5'de;

J_k : f 'nin w_k değerlendirilmiş Jakopyeni,

λ : Marquardt parametresi, ve

I: birim veya tanımlama matrisidir.

Levenberg-Marquardt algoritmasında hesaplama akışı aşağıdaki şekilde özetlenebilir [16].

- (i) $E(w_k)$ 'yi hesapla
- (ii) küçük bir λ değeri ile başla mesela ($\lambda = 0.01$)
- (iii) δw_k için Eş. 3.5'i çöz ve $E(w_k + \delta w_k)$ değerini hesapla,
- (iv) şayet $E(w_k + \delta w_k) \geq E(w_k) \lambda$ 'yı 10 kat artır ve (iii)'e git
- (v) şayet $E(w_k + \delta w_k) < E(w_k) \lambda$ 'yı 10 kat azalt, $w_k : w_k \leftarrow w_k + \delta w_k$ 'yı güncelleştir ve (iii) git.

Hedef çıkışı hesaplamak için YSA ağırlıklarının LM öğrenme algoritması kullanılarak eğitilmesi, ağırlık dizisi w_0 'a bir başlangıç değerinin atanması ile başlar

E

ve hataların karelerinin toplamı E 'nin hesaplanmasıyla devam eder. Her

e_i^2

terimi, hedef çıkış (y) ile gerçek çıkış (y_d) arasındaki farkın karesini ifade

e_i^2

eder. Bütün veri seti için E hata terimlerinin tamamının elde edilmesiyle,

ağırlık dizileri (i)'den (v)'ye kadar olan hesaplama akışı içerisinde, daha önce de açıklandığı gibi LM öğrenme algoritması adımlarının uygulanmasıyla adapte edilir.

3.2.2. Geri yayılım algoritması

Bu algoritma özetle, ağırlıklar ve biaslar için aynı şekilde bir hesaplama yöntemi tanımlamaktadır.

$$x_{k+1} = x_k - \alpha_k g_k \quad (3.6)$$

ile tanımlanır. Burada x_k mevcut andaki ağırlık değerini, x_{k+1} düzeltilmiş ağırlık değerlerini, α_k öğrenme katsayısını, g_k eğimin mevcut değerini ifade eder. Standart geri yayılım algoritması eğimin azaltılması için 2 farklı yöntem sunmaktadır. İlk yöntem olan artırmalı şekilde eğitimlerin hesaplanarak ağırlıkların güncelleştirilmesi işlemi her bir girişin ağırlık uygulanmasıyla gerçekleştirilir, diğer yöntem olan grup şeklinde ise ağırlıkların güncelleştirilmesi için tüm eğitim kümesinin ağırlık uygulanması gerekmektedir.

3.2.3. Momentumlu geri yayılım algoritması

Uygulamalarda en yaygın ve en sık kullanılan yöntem standart geri yayılım algoritmasıdır. Bu algoritmanın işlem yeteneği rahat anlaşılabilir ve matematiksel olarak ifade edilmesi kolaydır. Bu nedenle en çok tercih edilen öğrenme algoritmasıdır. Bu algoritma, hataları çıkıştan girişe doğru azaltmaya çalışmasından dolayı geri yayılım olarak adlandırılmıştır. Geri yayılım algoritması MLP'leri eğitmede en çok kullanılan temel bir öğrenme algoritmasıdır [16]. Ağırlıkları değiştirme işlemi veya öğrenme işlemi aşağıdaki adımlarda verilmiştir.

- (i) Başlangıç ağırlıklarını rasgele seç
- (ii) Öğrenmeye başla
- (iii) Giriş kümesini giriş katına uygula
- (iv) İşlemci elemanlarının üzerinden çıkışı hesapla
- (v) Hata Bul ve Hata kabul edilemez ise Eğim (Gradient) azaltma ile ağırlıkları düzenle ve (iii) işlemine git.
- (vi) Test işlemine başla
- (vii) Test kümesini yapay sinir ağının giriş katına uygula
- (viii) İşlemci elemanlarının üzerinden çıkışı hesapla
- (ix) Ağırlık gerçek çıkışını hesapla
- (x) Test kümesi tamamlandıysa dur, yoksa (vii) nolu işlemine git.

Eğitim ve test işlemleri aynı şekilde yapılmaktadır. Bu algoritma ile i ve j elemanları arasındaki ağırlıklardaki değişim $\Delta w_{ji}(t)$ Eş. 3.7'deki gibi verilir.

$$\Delta w_{ji}(t+1) = \eta \delta_j x_i + \alpha \Delta w_{ji}(t) \quad (3.7)$$

η : Öğrenme katsayısı

α : Momentum Katsayısı

δ_j : Ara veya çıkış katmanındaki herhangi bir j nöronuna ait faktördür.

Çıkış katmanı için ilgili denklem Eş. 3.8'de verilmiştir.

$$\delta_j = \frac{\partial f}{(\partial net_j) (y_j^{(t)} - y_j)} \quad (3.8)$$

Bu denklemde; $net_j = \sum x_i w_{ji}$ ile ifade edilirken, $y_j^{(t)}$ ise j işlemci elemanının hedef çıkışıdır. Ara katmanlardaki nöronlar için bu faktör Eş. 3.9'de verilmiştir.

$$\delta_j = \frac{\partial f}{\partial net_j} \sum w_{qi} \delta_q$$

(3.9)

Ara katmanlarda nöronlar için herhangi bir hedef çıkış bulunmadığından Eş. 3.8 yerine Eş. 3.9 kullanılır. Bu algoritma temel olarak istenilen çıkış ile yapay sinir ağının hesaplanan çıkışı arasında hatanın ağırlıklara bağlı olarak düşürülmesini hedeflemektedir [16].

3.2.4. Esnek yayılım algoritması

MLP yapılarında genel olarak ara katman transfer fonksiyonları sigmoid tercih edilmektedir. Sonsuz aralıkta olabilen giriş değerleri sınırlı bir aralığa sıkıştırıldığı için bu fonksiyonlar sıkıştırıcı fonksiyonlar olarak anılmaktadır. Sigmoid transfer fonksiyonları büyük giriş değerlerini sıfıra yakınsayacak şekilde işlem görürler. Bu sebeple bias ve ağırlık değerleri henüz istenilen düzeye erişmeden yaşanan eğim değerinin çok yavaş değişmesi olasılığı öğrenme sürecinde sorunlara neden olmaktadır. Bu öğrenme algoritmasının amacı kısmi türevlerin olumsuz sonuçlara neden olmasını engellemektir. Ağırlık değerini güncelleştirmek için sadece türev değerlerine ait işaretler kullanılır. Bu öğrenme algoritması ile diğer algoritmalar arasındaki en belirgin farklılık türev değerinin öneminin olmamasıdır. Bu özellik esnek yayılım algoritmasına hızlı çözüm yeteneği kazandırmaktadır [16].

$$\sum w_{ji}(k) = \begin{cases} -A_{ji}(k) & \text{ise } B(k) > 0 \\ +A_{ji}(k) & \text{ise } B(k) < 0 \\ 0, & \text{ise değişim yok} \end{cases} \quad (3.10)$$

ile hesaplanır. Burada $B(k) = \frac{\partial E}{(\partial w_{ji})(k)}$ hesaplanır. $A_{ji}(k)$ ise,

$$A_{ji}(k) = \begin{cases} \eta A_{ji}(k-1) & \text{ise } B(k-1)B(k) > 0 \\ \mu A_{ji}(k-1) & \text{ise } B(k-1)B(k) < 0 \\ A_{ji}(k-1) & \text{ise deęişim yok} \end{cases} \quad (3.11)$$

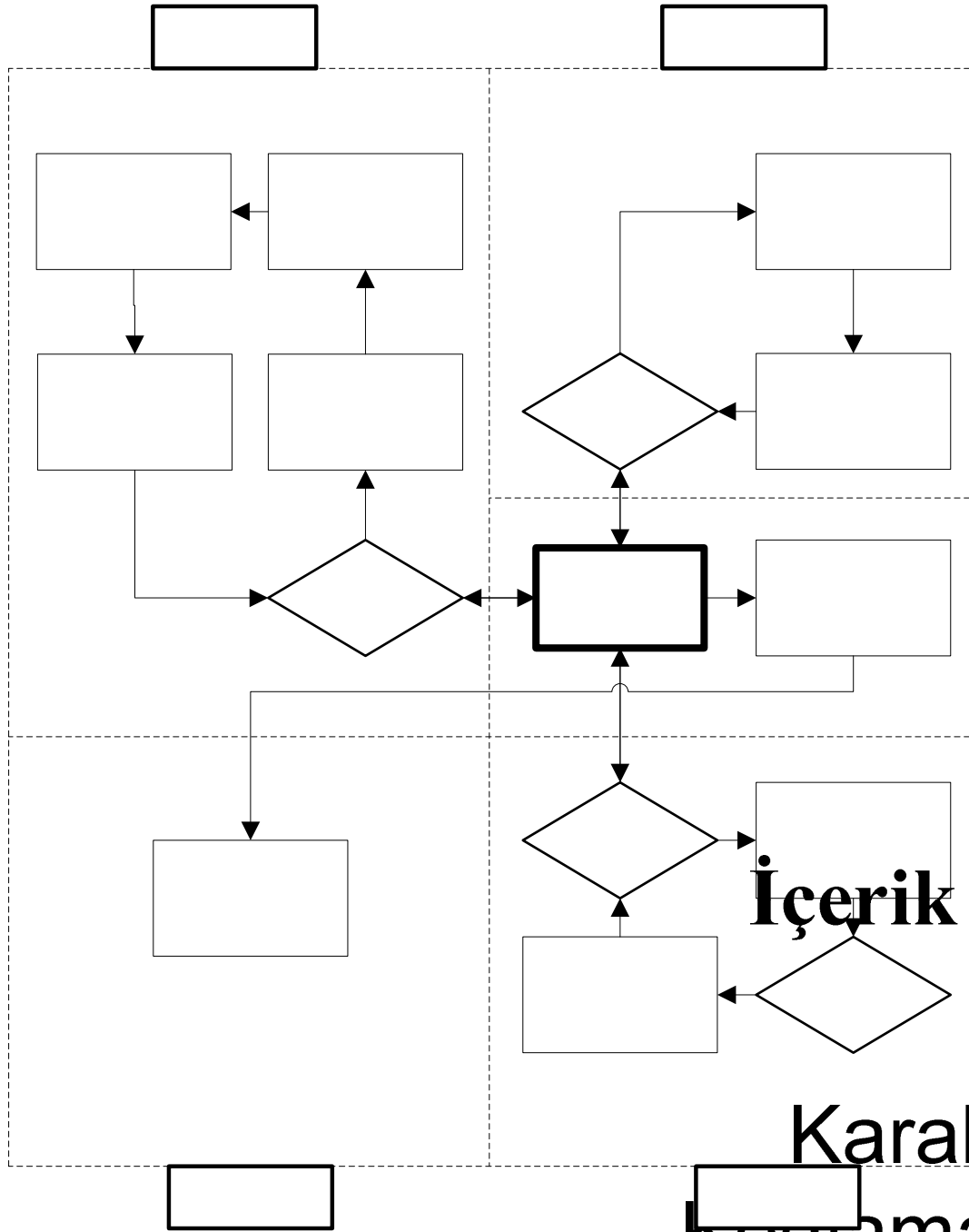
Eşitlik $B(k-1) = \frac{\partial E}{(\partial w_{ji})(k-1)}$, η ve μ sırasıyla $0 < \mu < 1 < \eta$ artma ve azalma faktörleridir. Bu hesaplama süreçleri içerisinde türev değeri sıfır ise güncelleme değeri sabit kalır. Ağırlıklar salınım yaptığıında, ağırlık deęişimi azalacağı gibi birkaç iterasyon boyunca deęişim aynı yönde gerçekleşirse ağırlık deęişimi artar. Bu algoritma sayesinde hafıza eski deęerler saklanmadığından sistem performansını artırmaktadır [16].

4. DİL TANIMA BÖLÜMÜ

Bu tez çalışmasının amacı sayısal dokümanların ve web sitesi içeriklerinin dilini tanımak ve tanınan dile göre içeriğin istenilen dile otomatik olarak çevrilmesini sağlamaktır. Bu çerçevede iki farklı yaklaşım geliştirilmiştir. Bu tez kapsamında geliştirilen iki yaklaşım ile bunlara ait modüller, çalışma mekanizmaları ve çalışmanın performansını test etmek için iki test yöntemi aşağıda detaylı açıklanmıştır.

4.1. Geliştirilen Birinci Yaklaşım

Bu çalışmada farklı dillerdeki web sayfası içeriklerinin tanınması için yapay sinir ağları kullanılmıştır. Sonuçlar Türkçe, Fransızca, İtalyanca, Hollandaca ve Almanca dilleri için test edilmiştir. Geliştirilen dil tanıma sistemi yapay sinir ağları sayesinde tanıma performansı oldukça yüksektir. Sistem tasarlanırken tanınması istenilen alfabelerin birleşim kümesi kullanılmıştır. Bu sebeple sadece tek bir dil sınıfına özgü harflerde eğitim kümesinde yer almaktadır. Bu işlem karmaşıklığını artırdığı gibi, dile özgü harflerin sistemde bulunmasından dolayı yapay sinir ağları öğrenme süreci ezbere kayarak sistemden istenmeyen davranışlara neden olmaktadır. Bu çalışma uluslararası bir konferansta 2007 yılında sunulmuştur [17]. Bu bildiri kapsamın da geliştirilen yazılım web sayfası içeriklerini inceleyerek sayfanın dilini tespit etmekte ve ilgili sayfayı bahsedilen çeviri servisleri yardımıyla İngilizce diline dönüştürmektedir. Bu çalışmaya ait akış diyagramı Şekil 4.1’de verilmiş ve detayları sunulmuştur



Şekil 4.1. Yaklaşımına ait Blok Diyagram

Şekil 4.1’de verilen blok diyagramın akışı aşağıda özetlenmiştir.

- (1) Modül 1 de Kullanıcı Girişine Web Sayfası Adresi Uygula
- (2) Varsayılan Karakter Kodlaması ile Sayfa Kaynak Kodunu Sisteme getir
- (3) Olması Gereken Karakter Kodlama Kümesi ile Sayfa Kodunu Sisteme Getir

URL Karakter
Kodlaması Bu

- (4) Modül 2 de Sayfa Kodu Üzerinde İngilizce Zorunluluk Kelimeleri ile HTML Kodlarını Temizle
- (5) Modül 3 de Önerilen Alfabe Kümesi ile Karakter Sıklık Analizi Yap
- (6) Karakter Frekans Analizi Sonucunu Önceden Eğitilmiş Yapay Sinir Ağına Uygula
- (7) Yapay Sinir Ağı Çıkışını Karşılaştı ve Dil Tanıma Sonucu Bul
- (8) Modüle 4 de Web Sayfasını Tanıma Sonucundan İngilizceye Çevir

Bir metine ait harf sıklık yüzdesinden doğal dilleri tanıyabilmek için eğitim esnasında; sisteme girişler, hangi dile ait olduğu bilinen metinlerin harf sıklık analizleri, çıkışlar ise bu dilleri temsil etmeye yarayan sayı değerleridir. Böylece sistem harf sıklık analizleri ile tanınması istenilen dil kodu arasındaki ilişkiyi öğrenmektedir. Bu zeki sistemde sunulan yöntem ile aradaki ilişkiyi öğrendikten sonra sisteme sadece tanınması istenilen dile ait metin bilgisi web sayfası üzerinden çevrimiçi olarak uygulanır, sistem öğrendiği bilgileri kullanarak metin bilgisine ait harf sıklığını modeller ve bu modelin özelliklerine göre uygulanan doğal dile ait dil sınıfını belirler. Bu sayede uygun koda karşılık gelen dil çözümlenerek açığa çıkar.

4.1.1. Geliştirilen Modüller

Alt başlıklarda Şekil 4.1’de verilen modüller detaylı açıklanmıştır.

İçerik Getirme Modülü

Bu çalışmada geliştirme platform olarak ASP.Net C# programlama altyapısı kullanılmıştır. Web sayfalarıyla uğraşmanın en sıkıntılı yönleri içerdikleri Hypertext Markup Language (HTML) etiketleri ve karakter kodlama kümeleridir. Bir sayfanın karakter kodlama kümesi anlaşılmadan metinsel işlemlerde başarı sağlamak mümkün değildir. Sayfalarda genellikle UTF-8 karakter kodlama kümesi kullanılmakla birlikte ISO-8859-9 Türkçe gibi bölgesel karakter kodlama kümeleri de kullanılmaktadır. Geliştirilen uygulama karakter kodlama kümesini anlamak için

çeşitli “String” operasyonları gerçekleştirmektedir. Sayfalara ait HTML etiketleri içeren kodlar birçok durumda karakter kodlama kümesini de içinde barındırmaktadır. Bu karakter kodlama kümesi kelime arama ve ayraç kullanarak sınıflandırma ile elde edilmektedir. Metinlerin uygun karakter kodlama kümelerine göre alınması için sayfanın karakter kodlama kümesi UTF-8 olarak alınmakta UTF-8 verisinden elde edilen karakter kodlama kümesi değeri ile sayfa yeniden sunucudan çağırılmaktadır. Bu sayede düzgün karakter kodlama kümesi kullanılarak kodlanmış HTML etiketli veri elde edilmektedir.

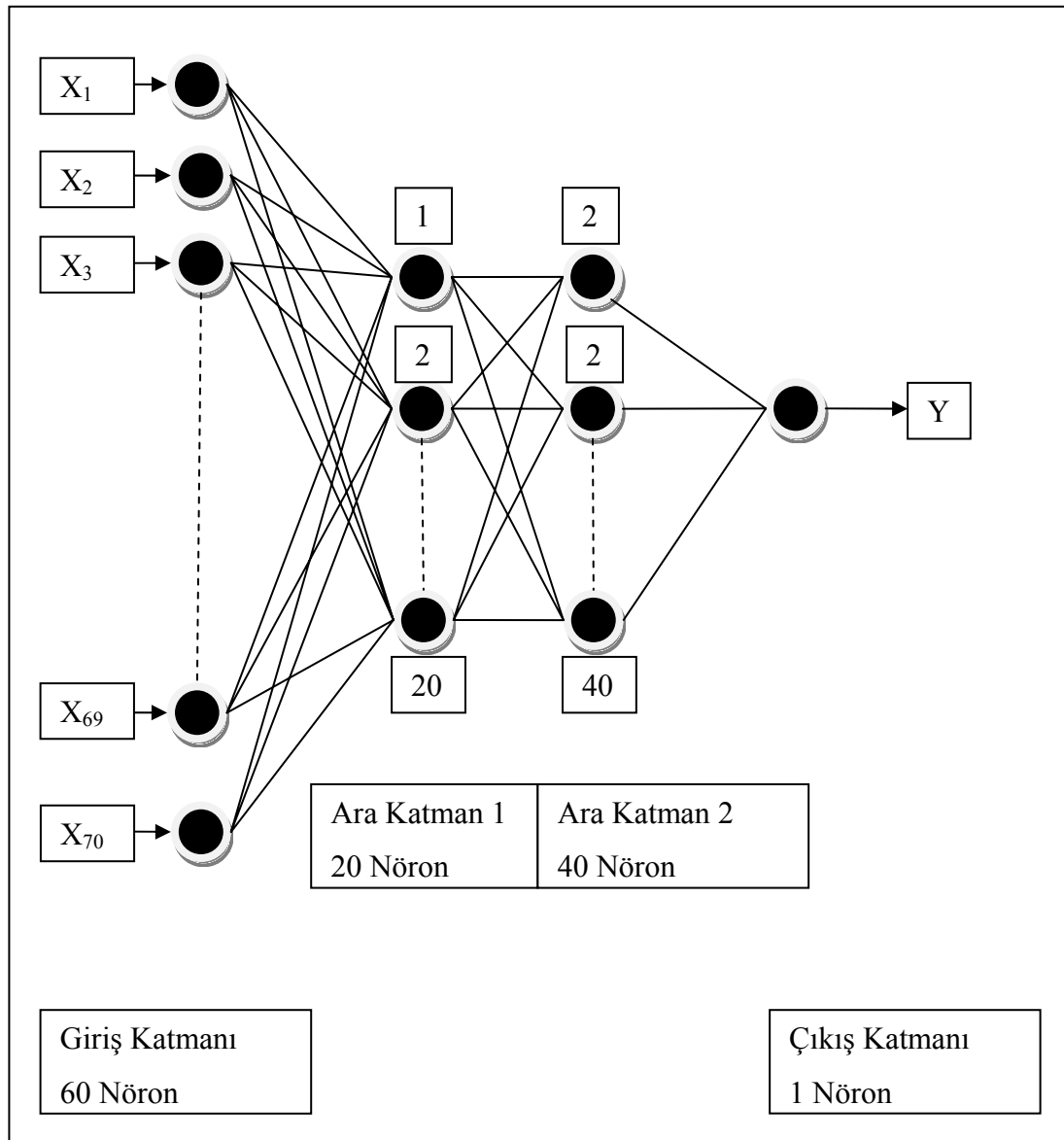
İçerik Temizleme Modülü

HTML etiketlerinin temizlenmesi ve sayfaya ait metinlerin çıkartılması önemli bir süreçtir. Çünkü sayfanın dili HTML etiket dili olarak kullanılan İngilizce ile karıştırılabilir. Bu karışıklığa sebebiyet vermemek için sayfanın sayfadaki HTML etiketlerinden arındırılması gerekmektedir. Bu işlem için “Regular Expression” denilen Otomat yapıları kullanılmıştır. Bir HTML etiketi < ile başlayıp > etiketi ile bitmektedir. Bu etiketlerin başlama ve bitiş formülleri "<(.\n)*?>" örneğinde olduğu gibi Regex adlı C# programlama dili komutuna verildiği zaman bu formülün arasını sağlayan tüm etiketler çıkartılmaktadır. Onlarca HTML etiketini çıkartmak için hazır bir sınıf yapısı kullanılmıştır. Sonuç metin HTML etiketlerinden arındırılmış olmasına rağmen boşluklar, web adresleri, yorum satırları gibi ek bileşenler “Regular Expression” ile ayrıştırılamayacak bağlamdan bağımsız deyimlerinde çıkartılması gerekmektedir. Bu yüzden elde edilen metinden “com”, “tr”, “org” veya “document”, “read”, “write” gibi betik komutları çıkartılmaktadır. Böylece yapay sinir ağlarına giriş olarak uygulanabilecek karakter frekans analizleri yapmak mümkün olmaktadır.

Dil Tanıma Modülü

İçerik temizleme modülünden elde edilen içerik bilgisinin önceden aynı yöntem kullanılarak eğitilen Şekil 4.2’de verilen YSA modeline giriş olarak verilebilmesi için sistem tarafından tanınması istenilen tüm dillere ait alfabe kümesi ile karakter

frekans analizinin yapılması gerekmektedir. Bu sayede YSA'lar karakter frekans analizi bilgisi ile giriş olarak verilen ve dili bilinmeyen web sayfası arasında matematiksel bir ilişki kurmaktadır. Bu matematiksel ilişkinin sağlıklı işlemesi ise giriş olarak uygulanan sayfanın HTML kaynak kodunun iyi bir şekilde analiz edilerek temizlenmesidir.



Şekil 4.2. Bilimsel yaklaşım için tasarlanan YSA modeli

Çeviri Modülü

Yapay sinir ağlarından elde edilen sonuçlara göre sayfanın dili sistem tarafından eğitim esnasından çıkartılan referans sonuçlar ile karşılaştırılarak bulunmaktadır. Bu sonuç sınırlı sayıda dil için sayfanın anlaşılabilir ortak bir dil platformuna dönüştürülmesi zorunluluğunu doğurmaktadır. Geliştirilen sistem mevcut dil çeviri yazılımlarıyla uyum içinde çalışabilmektedir. Bu çalışmanın yapıldığı sırada en çok kullanılan online çeviri aracı olan Altavista firmasının Babelfish yazılımı kullanılarak sayfaların sadece İngilizce diline çevirileri yapılmıştır.

4.1.2. Yazılım platformu uygulaması

Bu yazılım mevcut çeviri yazılımları ile entegre çalışabilmesi için C# ve ASP dillerinde Visual Studio platformunda kodlanmıştır. Bu sebeple internet ortamında erişime açık bir servis olarak çalışmaktadır. Kullanıcılardan sayfa girişi alarak bu sayfaya ait dili ve sayfanın İngilizce diline dönüşüm işlemini otomatik olarak gerçekleştirmektedir. Bu uygulamanın detaylı açıklaması ve test süreçlerini aşağıda verilmiştir.

Tasarım ve Test Süreçleri

Yazılıma ekran çıktıları Şekil 4.3’de sunulmuştur. Yazılım internet ortamında çalışabilecek bir servis olarak tasarlanmıştır. Kullanıcıların bulundukları istekler ana sunucu makine tarafından işlenerek sonuçlar tekrar kullanıcıya ulaştırılmaktadır. Bu nedenle sistemin çok kullanıcılı çalışması için sunucu makine ve internet bant genişliği kapasitelerinin yeterli düzeyde olması gerekmektedir.

Şekil 4.3. Yazılım Geliştirme Süreci

Şekil 4.3’de verilen ekran çıktısı yazılım geliştirme süreçlerinden alınmıştır. Bu ekran üzerinde verilen sıfırla butonu kullanılan penceredeki parametreleri sıfırlayarak yeniden kullanılabilir hale getirmektedir. Bu butonun hemen altında sayfa adresi etiketli bir metin kutusu bulunmaktadır. Kullanıcı dil tanıma sistemine giriş olarak uygulamak istediği sayfa adreslerini bu kutu alalına yazmaktadır. Bu kutu alanının hemen sağında bulunan içerik getir butonu çift yönlü çalışan bir mimari yapıya sahiptir ve kullanıcı için işlem karmaşıklığını azaltmaktadır. İçerik getir butonu basıldığında kullanıcı makinesine kendisine giriş olarak uygulanan web sitesinin adresini ana sunucu makineye göndermektedir. Ana sunucu bu web sayfasını varsayılan UTF-8 karakter kodlama kümesi ile bulunduğu alan barındırma hizmeti veren sunucudan almaktadır. Bu işlem ile kaynak koduna ulaşılan sayfaya

çeşitli örüntü tanıma filtreleri uygulanarak sayfanın gerçek karakter kodlama örüntüsü bulunmaktadır. Ana sunucu istenilen sayfanın karakter kodlama örüntüsünü bulduktan sonra UTF-8 ile tanınamayan karakterleri de sisteme almak amacıyla bulunan karakter kodlama kümesiyle sayfayı ilgili alan barındırma sunucusundan tekrar almaktadır. Sunucudan sorunsuz şekilde ana sunucuya alınan sayfaya ait kaynak koduna HTML etiketleri temizleme işlemleri uygulandıktan sonra sayfa içeriği ile etiketlenmiş olan metin kutusuna gönderilmektedir.

Bu sayede kullanıcılar dil tanınması amacıyla giriş olarak uyguladıkları internet adresinin HTML etiketleri temizlenmiş halini kendi makinelerinde görüntüleyebilmektedir. Bu aşamada halen işlemler ana sunucu üzerinde yapılmaktadır, kullanıcılar sadece bu işlemlerin sonuçlarını kendi makinelerinde görüntülemektedir. Dil tanıma sürecinin en önemli aşaması daha önceden eğitimi yapılmış olan YSA modeline giriş veri kümesinin uygulanmasıdır. Bu uygulama işlemi için metin bilgisinin frekans analizi olarak andığımız yöntem tabi tutularak ön işlemden geçirilmesi gerekmektedir. Bu nedenle konunun daha iyi özümlemesi için sistem tarafından otomatik yapılan bir süreç butona bağlanmıştır. Şekil 4.3’de verilen ekran çıktısının sağ üst kısmında görülen frekans analizi butonuna tıklandığı zaman sistem tarafından tanınması beklenen dillere ait alfabe kesişim kümesi baz alınarak harf frekans sıklık yüzdeleri hesaplanmaktadır. Hesaplanan bu yüzde değerler kullanıcı makinelerine gönderilerek sonuçlar bu butonun altında bulunan metin kutucuklarında gösterilmektedir.

Bu işlemden sonra frekans analizi yapılan metin dolayısıyla internet sitesi için dil tanıma işlemi yapılmasını sağlayan YSA yapısına giriş olarak uygulanması için sayfanın orta sağında bulunan yapay sinir ağlarına uygula butonuna basılmaktadır. Bu buton sayesinde ana sunucu kullanıcı tarafından gönderilen frekans analizi bilgisini ana sunucu üzerinde bulunan ve Matlab aracı tarafından oluşturulan YSA modeline giriş olarak uygulamaktadır. Bu işlem sonucunda ms mertebesinde bir hız ile ana sunucu YSA tarafından çıkış katmanında üretilen sonucu kullanıcı makinesine göndermektedir. Bu sonuçlar her dile eğitim esnasında verilen referans

değerler ile karşılaştırılarak en çok yakınsadığı yani aralarındaki hatanın en düşük olduğu dil kullanıcıya bir açılır pencere içinde gösterilmektedir. Kullanıcılar sayfanın alt kısmında yer alan İngilizceye çevir butonu sayesinde bu çalışma yapıldığı sırada internet sayfalarının çevirisinde en çok kullanılan Altavista Babelfish servisi kullanılarak sayfanın çevirisi ana sunucu üzerinden kullanıcılara aktarılmaktadır.

Bu çalışmaya ait www.gazi.edu.tr internet adresinin dil tanıma süreci ve geliştirilen yazılım Şekil 4.4’de ekran görüntüsünde sunulmaktadır. Bu ekran görüntüsünde sayfanın HTML’den temizlenmiş kaynak kodu, sayfaya ait sistem tarafından bulunan karakter kodlama kümesi, frekans analizi sonucunda ortak alfabe harflerine karşılık gelen yüzdelik değerler ve yapay sinir ağlarına uygulana frekans analizi neticesinde elde edilen dil tanıma sonucu verilmiştir.

Address

Page Content

Gazi Üniversitesi Yabancı Dil Destek Programı Devam Ediyor
25.06.2007 Bahar Dönemi İletişim Site Haritası --
İngilizce Satejik Plan: TÖMER: GÜADEK: GAZİSEM: İda
re Faaliyet Raporu: Akademik Yüksel. Ölçütleri: Tanıtı
m Kataloğu: İhale Bilgi Sistemi Gazi'de Ara Personelara
-
Radyo Gazi Yerelnet G.Ü. Spor Kulübü Öğrenci İşleri Bilgi S
istemi Ü Personel Ara İdari Bülten Etkinlikler Kurumsal De
ğerlendirme Süreci Yabancı Dil Destek Programı Devam Ediy
or Yapı İşleri Dairesi Başkanlığı Web Sayfası Faal Durumda
2007-
2008 Yabancı Uyruğu Öğrenci Kabulü/ Admission Of Foreign St
udents SHÇEK Sınav Sonuçları Ünvan Değişikliği Sınavı Baş
vuru Sonuçları Akademik Personel İlanı Yaz Okulu Ücretler

A-10,44
B-2,24
C-0,90
Ç-0,56
D-4,15
E-9,88
F-0,90
G-1,80
Ğ-1,01
H-0,56
I-3,70
İ-9,88
J-0,11
K-4,26
L-6,40
M-3,48
N-5,61
O-2,92
Ö-1,12
P-1,23
R-7,18
S-5,27
Ş-1,01
T-3,82
U-3,48
Ü-2,69
V-1,46
Y-2,47
Z-1,23
W-0,11
X-0,11
Q-0,00
Ç-0,00

Character Set

Neural Network R

Windows Internet Explorer

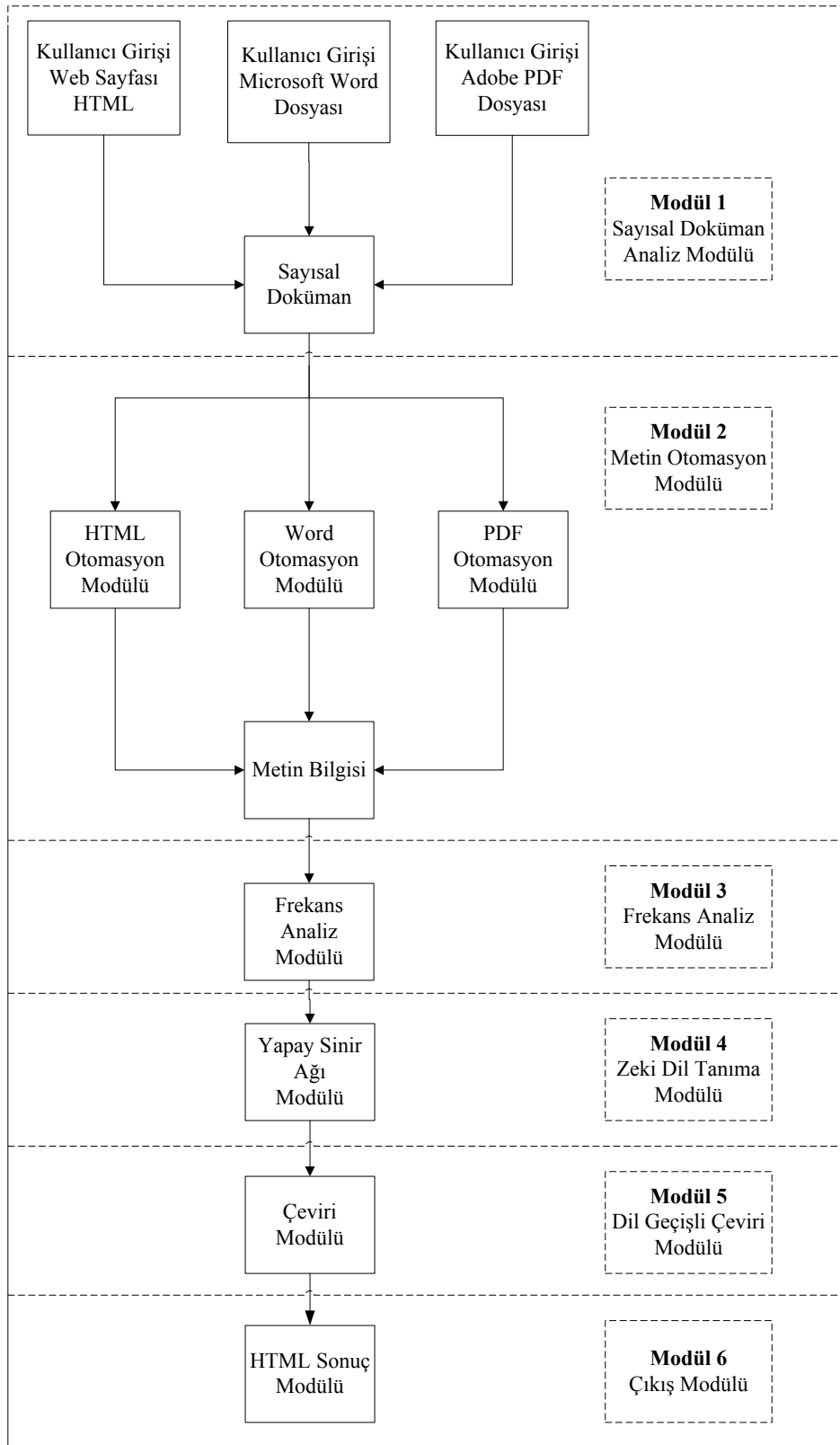
Türkçe-Turkish

Şekil 4.4. Geliştirilen yazılım platformu arayüzü

4.2. Geliştirilen İkinci Yaklaşım

Bu teze konu olan ikinci yöntem Bölüm 4.1’de verilen çalışmanın geliştirilmiş ve iyileştirilmiş halini içermektedir. Yeni önerilen yöntemler kullanılarak sayısal dokümanlardan elde edilen giriş verilerini kullanarak istenilen dile otomatik dönüşüm gerçekleştirilebilen zeki bir dil tanıma uygulaması geliştirilmiştir. YSA kullanılarak gerçekleştirilen bu zeki dil tanıyıcı ve dönüştürücü yazılımının blok şeması Şekil 4.5’de verilmiştir. Şekil 4.5’de sunulan blok şemasında görülebileceği gibi geliştirilen bu uygulama yazılımı, ön işlem, eğitim, test, sonuç-başarı oranı, otomatik dil çevirisi ve ana menü olmak üzere beş modülden oluşmaktadır. Bu modüller Şekil 4.5’de kesikli çizgiler ile birbirinden ayrılmış olan Sayısal Doküman Analiz Modülü, Metin Otomasyon Modülü, Frekans Analiz Modülü, Zeki Dil Tanıma Modülü, Dil Geçişli Çeviri Modülü ve Çıkış Modülü’dür

Bu tez kapsamında gerçekleştirilen, 15 dilde (Almanca, Arnavutça, Fransızca, Galce, Hırvatça, İngilizce, İrlandaca, İspanyolca, İtalyanca, Letonca, Macarca, Maltaca, Portekizce, Türkçe, Vietnamca) zeki dil tanıyıcı ve dönüştürücü tasarımı için yapılan işlemler bu bölümde anlatılmıştır. Öncelikle eğitim ve test için kullanılan veri kümeleri hakkında detaylı bilgi verilmiştir. Zeki dil tanıyıcı ve dönüştürücünün en önemli kısmını oluşturan YSA yapısı açıklanmış ve elde edilen en iyi YSA modelinin test sonuçları ise bu bölümün sonunda sunulmuştur.



Şekil 4.5. Geliştirilen yazılım platformuna ait blok diyagram

4.2.1. Veri Kümeleri

Bu tez çalışmasında doküman içeriğinin daha iyi ayrıştırılması için Bölüm 4.1’de önerilen ortak alfabe kümesi yaklaşımı YSA üzerindeki performans düşüklüğünü ortadan kaldırmak sebebiyle patenti tarafımızca alınmış olan Birleşim ve Kesişim tespit yöntemleri geliştirilmiştir. Bu yöntemlerin uygulanabilmesi için bir dile ait alfabe örüntüsünün hangi kurallar çerçevesinde birleştirileceğinin belirlenmesi gerekmektedir. Bu tez çalışmasında Latin Alfabesi ve bu alfabeden türeyen alfabelerin bağlı oldukları diller dikkate alınarak çalışmalar yapılmıştır.

Eğitim veri kümeleri oluşturulurken Latin alfabesi grubuna dahil 15 farklı dil için giriş amacıyla kullanılacak sayısal (WORD ve PDF) dokümanları kullanılmıştır. Geliştirilen bu yaklaşımda test verilerinin üretilmesi için Çizelge 4.1’de verilen anahtar kelime tablosu ve Google arama motoru kullanılmıştır. Bu tablo da verilen Erkek, Kadın, İnsan, Dünya, Ülke, Bilim, Kültür, Sanat, Politika, Hayat kelimeleri bir dil içerisinde en sık kullanılan kelimeler olarak önerilmiştir. Veri kümeleri oluşturulurken bu kelimelerin tanınması istenilen diğer dillerde çevirileri yapılmıştır. Bu çeviriler bir arama motorundan rasgele seçilen sayfalarda yer alan sonuçların her dilde benzer sayısal dokümanlara ulaşmayı sağlanacağı düşüncesiyle yapılmıştır. Bu sayede farklı kişiler tarafından yazılan ama benzer özellikler taşıyan toplumsal, bilimsel, politik vb. yazılar barındıran sayısal dokümanlar veri kümelerinin oluşturulmasında kullanılmıştır. Bu anahtar kelimelerin kullanılmasıyla elde edilen sayısal dokümanların farklı dillerde aynı konuları içermesi ile önişlemler sonucunda elde edilecek eğitim veri kümelerinin rastlantısal sonuçlar içermemesini de hedeflenmektedir. Eğitim işlemi için sayısal dokümanların önişlemden geçirilmesi gerekmektedir. Bu önişlem sonucunda Latin Alfabesi grubuna ait dillerin sınıflandırılması gerçekleşmektedir.

Çizelge 4.1. Anahtar kelime tablosu

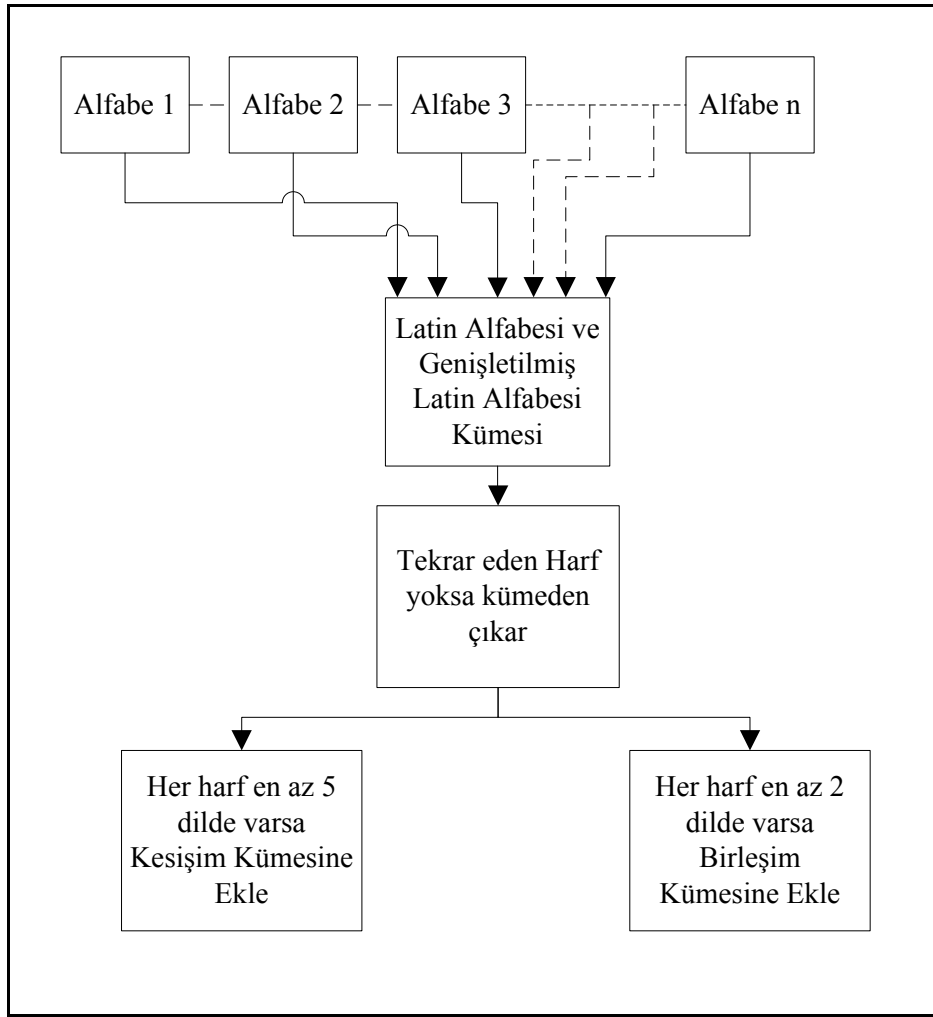
	Anahtar 1	Anahtar 2	Anahtar 3	Anahtar 4	Anahtar 5	Anahtar 6	Anahtar 7	Anahtar 8	Anahtar 9	Anahtar 10
Almanca	Männlich	Frau	Mensch	Welt	Land	Wissenschaft	Kultur	Kunst	Politik	Leben
Arnavutça	marr pozitë	femëror	njerëzor	planeti	fshatar	shkence	kulturë	zanat	Politikë	jetëdhënës
Fransızca	Homme	Femme	Humain	Terre	Pays	La Science	Culture	Art	Politique	La Vie
Galce	Dyn	Benyw	Dynol	Daeir	Gwledig	Gwyddoniaeth	Diwylliant	Celfyddyd	Gwleidyddiaeth	Bywyd
Hırvatça	čovjek	žena	čeljad	kopno	domovina	grana	gajenje	likovni	politička	osoba
İngilizce	Man	Woman	Human	Earth	Country	Science	Culture	Art	Politics	Life
İrlandaca	fear	bean	daonna	talamh	tír	eolaíocht	cultúr	ealaín	pholaitíocht	beatha
İspanyolca	Hombre	Mujer	Humano	Tierra	País	Ciencia	Cultura	Arte	Política	Vida
İtalyanca	Uomo	Donna	Umano	Terra	Paese	Scienza	Coltura	Arte	Politica	Vita
Letonca	cilvēks	sieva	cilvēka	zeme	valsts	zinātne	kultūra	māksla	Politika	dzīvība
Macarca	ember	nő	emberi	szárazföld	ország	tudomány	kultúra	művészeti	politikai	élet
Maltaca	Ragel	Mara	Uman	Dinja	Pajjiz	Xjenza	Kultura	Arti	Pulitka	Hajja
Portekizce	Homem	Mulher	Humano	Terra	País	Ciência	Cultura	Arte	Política	Vida
Türkçe	Erkek	Kadın	İnsan	Dünya	Ülke	Bilim	Kültür	Sanat	Politika	Hayat
Vietnamca	người đàn ông	người phụ nữ	con người	trái đất	nước	khoa học	văn hóa	nghệ thuật	chính trị	cuộc sống

Latin alfabesi 80’den fazla farklı dil için temel teşkil etmektedir. Alfabetik birleşim kümesi bir dile ait özel karakterlere duyarlıdır. Latin alfabesi zaman içinde diller arasında gelişerek farklılık göstermektedir. Klasik Latin alfabesi a, b, c, d, e, f, g, h, i, k, l, m, n, o, p, q, r, s, t, v, x, y, z [wiki latin] harflerinden oluşmaktadır. Bu harflere ek olarak günümüzde kullandığımız modern Latin alfabesi “j, u, w” harflerini de içermektedir. Bu çalışmaya konu olan 15 dil içerisinde İngilizce, Almanca, Fransızca, Macarca, İspanyolca dilin özel durumları dışında modern Latin alfabesindeki 26 harfi kapsamaktadır. Bu çalışmaya konu olan diğer diller ise modern Latin alfabesinin bazı harflerini içermemektedir. Bu dillere ait karşılaştırma Çizelge 4.1’de verilmiştir.

Latin alfabesinin genişletilmiş grubunda 161 harf bulunmaktadır. Bu harflerden bazıları Almanca’da bulunan “ß” gibi sadece dile özgü olmakla birlikte Galce ve Türkçe’de olan “Ü” gibi birden fazla dilde ortak olarak bulunmaktadır. Bu aşamada sadece dile özgü harflerin kullanımı YSA’nın öğrenme sürecini etkileyerek, sürecin ezbere kaymasına neden olmaktadır. Bu yüzden Birleşim Tespit Yönteminde Yapay Sinir Ağının karar mekanizmasını iyileştirmek için tek dile özgü harfler kullanılmamıştır. Birleşim Tespit Kümesi oluşturulurken her harfin en az 2 dilde bulunması esas alınmıştır. Kesişim Tespit Yönteminde ise kullanılan harfler hem klasik Latin alfabesinde hem de genişlemiş Latin Alfabesinde yer alan en az 5 alfabede ortak olan harflerdir. Bu sayede işlem uzayı küçültülerek sistemin Latin Alfabesinden türemiş her dil için çalışması esas alınmıştır. Bu çalışmada incelediğimiz dillere ait Birleşim Tespit Yöntemi alfabe kümesi Tablo ve de Kesişim Tespit Yöntemi alfabe kümesi Tablo ve verilmiştir. Eğitim verisi olarak kullanılacak dokümanların Birleşim Tespit Yöntemi (BTY) ve Kesişim Tespit Yöntemi (KTY) kullanılarak önışleme tabi tutulması gerekmektedir.

Önışlemler sonucunda önerilen yöntemlere göre harf sıklıklarının giriş metni içinde bulunma oranları hesaplanmaktadır. YSA harf sayıları ile doğal diller arasında bir ilişki kurulmasını sağlamaktadır. Alfabetik harf sıklığından dil tanıyan sistemin genel yapısı Şekil 4.6’da verilmiştir. Sistem alfabelere ait ortak veritabanı ile alfabeleri

oluşturan diller arasındaki ilişki modelini öğrenir veya mevcut dil desteklerinden faydalanır. Bunun için, sisteme uygun sayıda veri kümesi oluşturulur ve sistem eğitilir. Bu bölümde önerilen yöntemlere ait detaylar açıklanmıştır.



Şekil 4.6. Birleşim ve Kesişim Yöntemi Blok Diyagramı

4.2.2. Metot – 1: Birleşim tespit yöntemi

Bu yöntem YSA'nın öğrenme performansını artırmak için önerilmiştir. Yöntem, tanınması istenilen dillere ait alfabelerin tek bir ortak alfabeye entegre edilme fikrine yeni bir bakış açısı getirmektedir. Bu yöntem kullanılarak oluşturulan ortak alfabe kümesinde EK-20 - EK-25'de verilen standart ve genişletilmiş Latin alfabe

kümesinin tanınması istenilen 15 dil için analizlerinden EK-16 ve EK-17 'de verilen tekrar eden en az 2 harf seçilerek oluşturulmuş Birleşim Tespit Yöntemi Kümeleri oluşturulmuştur. Bu sayede bahsedilen 15 dil için her dil alfabesi içinde 2 kez ve daha fazla tekrar eden harflerden ortak alfabe kümesi oluşturulmuştur. Bu ortak alfabe kümesi ile tasarlanan sistem bütünü aşağıda adımları verilen Birleşim Tespit Yöntemi (BTY) olarak anılmaktadır.

1. Tanınması istenilen dillere ait alfabe karakterleri çıkartılır.
2. Bu alfabelerin karakterlerine ait birleşim kümesi bulunur.
3. Giriş metni (web sayfası, Word, PDF vs.) sisteme giriş olarak uygulanır.
4. Sistem tarafından alfabelerin kesişim kümesinde bulunan karakterlere otomatik olarak bir sayı atanır.
5. Girilen metin içinde yer alan harfler sayaç tarafından sayılır, alfabe birleşim kümesiyle karşılaştırarak tekrar eden harflerin sayıcı değerini bir arttırılır. Bu sayede alfabe kesişim kümesinde bulunan harflerin metin içinde kaç kez geçtiği bulunur.
6. Metin analizi tamamlandıktan sonra, harf sayaç sonuçları toplanarak yinelenen toplam harf sayısı elde edilir.
7. Sistem her harfin girilen metin üzerindeki yüzdelik dağılımını hesaplar.
8. Eğitim kümesi oluşturulurken bu yüzdelik dağılımlara, incelen dile ait bir sıra numarası ağına çıkışına verilir.
9. Eğitilen ağ, yöntem neticesinde ağa uygulanan yüzdelik değerler neticesinde ağın ürettiği çıkış, eğitimde kullanılan sıra numaraları ile karşılaştırılarak test metninin hangi dile en çok yakınsadığı ve buna bağlı olan yüzdelik hatası bulunur.
10. Elde edilen sonuçlar çeviri programları, arama motorları, ofis programları gibi otomasyon yazılımlarında kullanılarak kullanıcılara sunulur.

4.2.3. Metot – 2: Kesişim tespit yöntemi

Bu yöntem YSA'nın öğrenme performansını artırmak için önerilmiştir. Yöntem, tanınması istenilen dillere ait alfabelerin tek bir ortak alfabeye entegre edilme fikrine

yeni bir bakış açısı getirmektedir. Bu yöntem kullanılarak oluşturulan ortak alfabe kümesinde EK-20 – EK-25’de verilen standart ve genişletilmiş Latin alfabe kümesinin tanınması istenilen 15 dil için analizlerinden EK-18 ve EK-19’da verilen tekrar eden en az 5 harf seçilerek oluşturulmuş Kesişim Tespit Yöntemi Kümeleri oluşturulmuştur. Bu sayede bahsedilen 15 dil için her dil alfabesi içinde 5 kez ve daha fazla tekrar eden harflerden ortak alfabe kümesi oluşturulmuştur. Bu ortak alfabe kümesi ile tasarlanan sistem bütünü aşağıda adımları verilen Kesişim Tespit Yöntemi (KTY) olarak anılmaktadır.

1. Tanınması istenilen dillere ait alfabe karakterleri çıkartılır.
2. Bu alfabelerin karakterlerine ait kesişim kümesi bulunur.
3. Giriş metni (web sayfası, Word, PDF vs.) sisteme giriş olarak uygulanır.
4. Sistem tarafından alfabelerin kesişim kümesinde bulunan karakterlere otomatik olarak bir sayı atanır.
5. Girilen metin içinde yer alan harfler sayaç tarafından sayılır, alfabe kesişim kümesiyle karşılaştırarak tekrar eden harflerin sayıcı değerini bir arttırılır. Bu sayede alfabe kesişim kümesinde bulunan harflerin metin içinde kaç kez geçtiği bulunur.
6. Metin analizi tamamlandıktan sonra, harf sayaç sonuçları toplanarak yinelenen toplam harf sayısı elde edilir.
7. Sistem her harfin girilen metin üzerindeki yüzdelik dağılımını hesaplar.
8. Eğitim kümesi oluşturulurken bu yüzdelik dağılımlara, incelen dile ait bir sıra numarası ağına çıkışına verilir.
9. Eğitilen ağ, yöntem neticesinde ağa uygulanan yüzdelik değerler neticesinde ağın ürettiği çıkış, eğitimde kullanılan sıra numaraları ile karşılaştırılarak test metninin hangi dile en çok yakınsadığı ve buna bağlı olan yüzdelik hatası bulunur.
10. Elde edilen sonuçlar çeviri programları, arama motorları, ofis programları gibi otomasyon yazılımlarında kullanılarak kullanıcılara sunulur.

4.2.4. Geliştirilen modüller

Bu sayfada Şekil 4.5’de verilmiş olan yazılım platformunun blok diyagramında bahsedilen Sayısal Doküman Analiz Modülü (SDAM), Metin Otomasyon Modülü (MOM), Frekans Analiz Modülü (FAM), Zeki Dil Tanıma Modülü (ZDTM), Dil Geçişli Çeviri Modülü (DGÇM) ve Çıkış Modülü (ÇM) için detaylı açıklamalar aşağıda sırasıyla verilmiştir.

Sayısal Doküman Analiz Modülü

Sayısal Doküman Analiz Modülü (SDAM) sistemin girişine uygulanan sayısal dokümanların sınıflandırılması ve kullanıma hazır hale getirilmesi sürecindeki işlemleri gerçekleştirir. Sistem İnternet sayfası, Düz Metin, Microsoft Word dosya formatı ve Adobe PDF dosya formatı ile toplam 4 veri tipiyle çalışabilmektedir. Bu geliştirme sürecinde İnternet sayfaları için Bölüm 4.1’de verilen yaklaşım tercih edilmiştir. Bu yaklaşıma göre dili tanınması istenilen internet sitesinin varsayılan karakter kodlama kümesi ile bir kaydı alınmaktadır. Bu kayıt üzerinden ilk analiz yapıldıktan sonra sayfanın yayınlandığı karakter kodlama kümesi tespit edilmektedir. Bu karakter kodlama kümesi ile sayfa bir kez daha sunucudan istenmektedir. Bu sayede sayfaya ait kaynak kodu kaynağa ait karakter örüntülerini koruyarak sisteme ulaşmaktadır. Sisteme ulaşan kaynak koduna ek bir analiz daha yapılarak sayfaya ait bağlantı derinlikleri bulunmaktadır. Bu bağlantı derinlikleri sayesinde istenilen sayfaya ait kaç alt bağlantı olduğu kullanıcının bilgisine sunulmaktadır. Kullanıcılar bu modül sayesinde sayfaya ait mevcut bağlantı kadar sisteme giriş sayfası verebilmektedir.

Bu modül içerisinde kullanıcıların DOC uzantılı Microsoft Word dosyalarını ve PDF uzantılı Adobe Acrobat dosyalarını sisteme giriş olarak uygulamalarını sağlayan iki alt sistem bulunmaktadır. Bu sistemler dosyaları sonraki byte tabanlı olarak işleme olarak akışları (stream) sonraki modüle iletirler. Ayrıca, bu modül de kullanıcı tarafından girilen düz metinler ise metin otomasyon modülüne doğrudan aktarılır.

Metin Otomasyon Modülü

Bu modül kullanıcı tarafından girilen sayısal verilerden metin bilgisi elde etmek amacıyla kullanılmaktadır. Sayısal doküman analiz modülü tarafından elde edilen internet sayfalarına ait kaynak kodları, Microsoft Word ve Adobe Reader dosya akışları bu modülde otomasyon sistemine girerken, kullanıcı tarafından girilen düz metin bilgisi sonraki modüle aktarılmaktadır. Bir bilgisayar sisteminde bahsi geçen dosya formatlarının ve internet sayfa dosya formatı olan HTML'nin bir yazılım tarafından işlenebilmesi için sistem tarafından özel bileşenlere ihtiyaç duyulmaktadır. Bu özel bileşenler genel olarak dosya formatlarını üreten firmalar tarafından yazılım geliştiricilerine sunulmaktadır. Yazılım geliştirme araçları olarak adlandırılan özel paketler SDK (Software Development Kit) sayesinde yazılımlar ilgili dosya formatlarını işleyebilen yapıya gelmelerinin yanında SDK üreticisi firmaya bağımlı hale gelmektedir. Bu yüzden kullanılan SDK'ların kullanıcı sistemlerinde muadili olan ve bu dosya formatlarının okunması, yazılması ve işlenmesini sağlayan Adobe Reader yada Microsoft Ofis gibi yazılımların kurulmasını gerektirmektedir. Bu yazılım bağımlılığı kullanıcıları sadece ürünü değil, çalışması için gerekli yazılımları almasına da neden olmaktadır.

Bu sebeple Metin Otomasyon Modülü (MOM) tasarlanırken Microsoft Ofis ve Adobe Reader SDK paketleri yerine açık kaynak kodlu dağıtımı yapılan Cellbi Microsoft Word Otomasyon Kütüphanesi [22] ve PDFBox Adobe Reader Otomasyon Kütüphanesi [23] kullanılmıştır. Bu sayede kullanıcıların sistemlerine farklı yazılımlar kurması yada bu yazılımları satın almaları gerekmemektedir. Bu ara otomasyon yazılımları sayesinde ilk modülden akış dizisi olarak alınan dosyalar ilgili yapılarda işlendikten sonra düz metin bilgisi elde edilmektedir.

Frekans Analizi Modülü

Bu modül ile birleşim ve kesişim tespit yöntemiyle oluşturulan alfabe ile frekans analizi yaklaşımları gerçekleştirilir. Frekans Analiz Modülü'nde (FAM) kullanılan Birleşim ve Kesişim yöntemlerine ait detaylı bilgi önceki bölümlerde bahsedilmektedir. Bu modül her iki yöntemi birbirinden bağımsız ele alarak çalışmaktadır. Kesişim tespit yöntemi bahsi geçen dillere ait alfabelerde 5 kez ve daha fazla tekrar eden harflerden oluşan bir alfabe kullanırken Birleşim tespit yöntemi aynı alfabelerde 2 kez ve daha fazla tekrar eden harf kümesini belirlemektedir. Kesişim tespit yönteminde 43, Birleşim tespit yönteminde ise 60 karakter örüntüsü bulunmaktadır. MOM'den (Metin Otomasyon Modülü) elde edilen metin bilgisi bu modülde YSA tarafından işlenebilecek bir hale gelmektedir. Bu tez kapsamında geliştirilen yapay sinir ağı modeli metin harf frekans bilgisi ile dil arasındaki matematiksel ilişkiyi öğrenmektedir. Bu matematiksel ilişkinin girişine uygulanacak harf frekans analizi ise bu modül tarafından yapılarak giriş metni üzerinden birleşim ve kesişim yöntemleri alfabe kümesi esas alınarak harflerin yüzdelik frekans dağılımları hesaplanmaktadır. Bu modül sayesinde hesaplanan yüzdelik frekanslar sonraki modüle aktarılmaktadır.

Zeki Dil Tanıma Modülü

Bu tez kapsamında önerilen yöntemler kullanılarak daha önceden dilleri bilinen Microsoft Word ve Adobe Reader dosyalarının metin otomasyon modülünden elde edilen metin bilgileri kullanarak yapılan frekans analizleri ile ilgili dil arasındaki matematiksel ilişkiyi öğrenen YSA modeli bu modülün temelini oluşturmaktadır. Bu YSA modeli sayesinde sisteme uygulanan bir internet sayfası, Microsoft Word veya Adobe Reader dosyası ile düz metin bilgisinin dil tanıma sürecini aktarılması yapılabilmektedir. Bu YSA modeli Mathworks firmasının Matlab isimli mühendislik yazılımı tarafından sonraki bölümlerde ifade edildiği gibi oluşturulmaktadır. Bu model sayesinde Matlab yazılımı üzerinde YSA eğitim ve test süreçleri takip

edilebilmektedir. Bu yüzden geliştirilen yazılımın Matlab yazılımı ile eşzamanlı şekilde çalışması sağlanmalıdır. Bu çalışma şekli Matlab tarafından üretilen bir dinamik bağlantı kütüphanesi tarafından sağlanmaktadır. Bu sayede C# gibi programa dillerine Matlab yazılımı entegre edilebilmektedir. Bu YSA modelinin eğitim ve test sürecinin dinamik bağlantı kütüphanesi yapısına kavuşması Matlab Deployment Aracı sayesinde olmaktadır. Frekans Analiz Modülü (FAM) tarafından bu modüle gönderilen dil tanınması istenilen sayısal dokümana yüzdelik metin frekans bilgisi dinamik bağlantı kütüphanesi üzerinden daha önceden eğitilmiş ağ yapısına ulaşarak giriş olarak uygulanmakta ve tekrar aynı yolu izleyerek Matlab tarafından oluşturulan yapay sinir ağının çıkışından elde edilen sonucu kullanıcı ara yüzüne iletmektedir.

Bu sonuç bilgisi kullanıcı ara yüzünde sistem tarafından tanınabilecek mevcut diller arasında yüzdelik başarı oranıyla ekrana yansıtılmaktadır. Bu sayede kullanıcılar giriş olarak uyguladıkları sayısal dokümanın 1'den yaklaşıktan n yaklaşık sonuca kadar olabileceği dili yüzdelik olabilme oranlarıyla elde edebilmektedirler.

Dil Geçişli Çeviri Modülü

Bu modül sayesinde kullanıcılar dillerini hiç bilmedikleri sayısal dokümanları literatürde önerilen dil dönüşümü gerçekleştiren sistemler ile dönüştürebilmektedir. Bu sistemlerden en çok kullanılan Bölüm 4.1'de sunulan yaklaşımda kullanılan Altavista Babelfish sistemidir. Bu sistem sınırlı sayıda dil için metin ve internet tabanlı dönüşümler gerçekleştirebilmektedir. Bu tez çalışması yapılırken Google firması tarafından Google Translator sistemi geliştirilmiştir. Bu sistem sayesinde 35 farklı dil kendi içlerinde dönüşüme tabi olabilmektedir. Google tarafından geliştirilen yöntem bugüne kadar kullanılan metin bilgisinden dilbilgisi öğeleri elde ederek çevirme yaklaşımını kullanmamaktadır. Google tarafından geliştirilen yöntem sayesinde bir veritabanı diller arasında önceden yapılmış ve halen yapılmakta olan birebir çevirileri cümle yapısıyla saklamaktadır. Google Translator tarafından alınan bir giriş o girişe karşılık gelen veya en çok yakınsayan direkt çeviriye

dönüştürülmektedir. Bu sayede birebir çeviri sistemi oluşturma çalışmaları hız kazanmıştır.

Bu modül Google Translator tarafından geliştirilen yöntemi baz almaktadır. Google tarafından parametrik olarak kullanıma açılan sistem sayesinde Zeki Dil Tanıma Modülün'den (ZDTM) elde edilen dil bilgisi kullanıcı ara yüzüne yansıtılarak, kullanıcıların metin bilgisini Google Translator tarafından desteklenen 35 farklı dile dönüştürülmesi otomatik olarak yapılmaktadır. Bu otomatik çeviri işlemi sırasında Google kullanıcı tarafından sisteme girilen internet sayfalarını düzeni bozulmaksızın kullanıcılara sunmaktadır. Google Çeviri arayüzü Microsoft Word DOC ve Adobe Acrobat PDF dosyaları akış düzenleri bozulmadan metin bilgisi olarak çeviri işlemine tabi tutulmaktadır. Bu sayede istenilen dosya formatının evrensel bir sayısal doküman yapısında çevirisi elde edilmektedir.

Çıkış Modülü

Bu modül sayesinde kullanıcılar geliştirilen yazılım sayesinde dil tanıma işlemi gerçekleştirilen internet sayfalarını, Microsoft Word ve Adobe Reader dokümanları ile düz metin dosyalarına ait çeviri metinlerini yazılıma bütünleşik internet tarayıcısında görüntüleyebilir, kayıt edebilir ve yazıcı çıktılarını alabilirler. Bu modül kullanıcı bağımsız işlem yapabilme yeteneğine sahiptir. Bu sayede kullanıcılar her çeviri işlemi gerçekleştirdiklerinde sistem üretilen sonuç çıktıları kullanıcı bilgisayarlarına kayıt ederek kullanıcıların çevirisi yapılan sayısal dokümanlara kolayca ulaşmalarını sağlamaktadır. Çevirisi yapılan sayısal dokümanlar farklı ortamlara aktarılmasını kolaylaştırmak ve çeşitli metin işleme programlarına kolayca aktarılabilmelerini sağlamak amacıyla internet sayfası ve metin formatı olarak iki farklı dosya formatı ile kayıt edilmektedir. Bu kayıtlar karmaşıklığı azaltmak kullanıcı tarafından için giriş olarak uygulanan dosya adı veya internet sayfası adı ile yapılmaktadır. Bu tez çalışmasının sonraki bölümlerinde girilen sayısal dokümanların çevirisinin kayıt altına alındığı ÇM'ne (Çıkış Modülü) gelinceye kadar

gerçekleşen dil tanıma sürecinin matematiksel altyapısı ve YSA eğitim ve test süreçleri anlatılmaktadır.

4.2.5. YSA modelinin tasarlama aşamaları

YSA uygulamalarının başarısı uygulanacak yaklaşımlar ile yakından ilgilidir. YSA'larla çözümlenemeyecek problem sayısı çok az olmasına karşın başarılı bir tasarımı etkileyen pek çok faktör bulunmaktadır [16]. Bu tasarımlar genel olarak,

1. Uygun yapının seçimi
2. Bu yapıya uygun öğrenme algoritması ve uygulanan algoritmanın uygun parametrelerinin seçimi
3. Seçilen yapıya uygun giriş, ara katman ve çıkış sayılarının belirlenmesi
4. Ara katman nöron sayılarının yeteri kadar seçilmesi
5. Seçilen nöronlarda kullanılacak olan aktivasyon fonksiyonunun belirlenmesi
6. Eğitim ve test kümelerinin belirlenmesi ve bu kümelerde kullanılacak normalizasyon seviyelerinin belirlenmesi

gibi zor ve zaman alıcı bir süreç gerektirmektedir. Bu süreç doğru şekilde işletilmediği durumlarda karmaşıklık artacağından YSA'nın performansı ciddi oranda etkilenecektir. Bu sebeple ağırlık büyüklüğü, öğrenme seviyesi ve ağırlıkların sınıflandırma seviyeleri de bir YSA'nın performansını ve genelleme kabiliyetini etkileyen başlıca faktörler arasındadır. Bir YSA uygun parametrelerle tasarlanırsa kararlı bir yapı ortaya çıkacağından kararlı sonuçlar elde edilecektir [16].

YSA sistemlerinin başarısını etkileyen çok sayıda yapısal parametre vardır. Bu parametrelerin en uygun olanını seçmek başlı başına bir problemdir. Bu parametrelerin doğru seçimi teorik olarak mümkün gibi gözükse de pratikte bu parametreleri önceden kestirmek kolay bir iş değildir. Bunun için literatürdeki benzer çalışmalar incelenerek uygulanacak problem için belirlenmiş olan parametrelerle yapıyı tasarlamaya ve eğitmeye çalışmak en akılcı yaklaşımdır. Bu

sayede daha önceden belirli bir probleme bulunmuş çözüm iyi bir başlangıç noktası sağlamaktadır. Elbette, uygun YSA mimarisinin veya yapısının seçimi, uygulama problemine bağlı olarak seçilmesi gereken bir kriterdir. Bir YSA'nın karmaşıklığının azaltılmasında en etkin araç, ağın mimari yapısını değiştirmektedir. Bunun nedeni gereğinden fazla işlemci elemanı içeren ağlarda daha düşük genelleme kabiliyetinin oluşmasıdır [16].

Bir YSA yapısının seçilmesinden sonra karşılaşılan en büyük güçlük ise öğrenme algoritmasıdır. YSA'ların az veriyle öğrenebilmeleri ve genelleme yapabilmeleri YSA'ların birçok problemin çözümünde kullanılmalarının başlıca sebeplerindendir. Bunun nedeni deneysel olarak fazlaca veri elde etmenin maliyeti artırmasıdır. YSA sisteminin problemi öğrenme başarısı kullanılabilirliğini belirleyen öğrenme başarısını etkiler. En uygun, öğrenme seviyesi öğrenme fonksiyonunun kontrolü için kullanılan performans fonksiyonunun önceden amaçlanan bir değere ulaşması ile sağlanamayabilir. Uygulamalarda, eğitim süresi boyunca performans fonksiyonunun izlenmesi ile birlikte sık sık genelleme testlerinin gerçekleştirilmesi yolu ile en uygun öğrenme seviyesi elde edilir. Seçilecek olan transfer fonksiyonları ise öğrenmede önemli yaklaşımlardan biridir. Problemin davranışına uygun olan fonksiyon tipinin belirlenmesi başarı yüzdesini yüksek oranda etkilemektedir. Bu sebeple bazı problemlerde tek tip transfer fonksiyonu yerine birden fazla fonksiyon bir arada kullanılmaktadır [16].

YSA'larda verilerin sisteme giriş olarak verilebilmesi için ön işlemden geçirilerek normalize edilmesi performansı etkileyen faktörlerden biridir. Normalizasyon, giriş verileri transfer edilirken fonksiyonun aktif olan bölgesinden aktarılmasını sağlar. Veri normalizasyonu aşırı değerlenmiş kümülatif toplamaların oluşmasını engellediği için işlemci elemanlarını korumaktadır. Genellikle verilerin $[0,1]$ veya $[-1,+1]$ aralıklarından birine ölçeklenmesi yapılmaktadır. Ölçekleme, verilerin geçerli aralıkta sıkıştırılması anlamına gelmektedir. Veri ölçeklemede kullanılan normalizasyon araçlarından biriside transfer fonksiyonlardır. Bu sebeple aslında

fonksiyonun kendisinde bir ölçekleme aracı olarak kullanılabilir. Ön işleme aşaması dinamik yapıya sahip olan YSA uygulamalarında son derece önem taşımaktadır [16].

YSA'larında gerekli parametrelerin uyum içinde çalışması için ara katman ve her bir ara katmanda bulunacak olan nöron sayılarının belirlenmesi gerekmektedir. Literatürde bazı yaklaşımlar sunulmuş olmasına karşın bazı özel problemlerde bu yaklaşımlar yetersiz kalmaktadır. Bunun en basit ama en uygun kısa yolu birkaç deneme yaptıktan sonra en uygun yapının ve nöron sayılarının en olması gerektiğinin belirlenmesidir. Literatürde, her türlü problemin en az iki ara katmanla çözülebileceği yönünde varsayımlarda bulunmaktadır. Özel problemler için bu kavram her zaman doğru olmayabilir ve katmanlar arasında bazı özel bağlantılara ihtiyaç duyulabilir. Bir YSA tasarımında temel amaç, YSA sisteminin mümkün olduğunca basit yapıda tasarlanmasıdır. Ara bağlantıların, kısmen veya tamamen, ileri beslemeli veya çift yönlü olmaları tasarım için önemli yaklaşımlardandır [16].

YSA hesaplama işlemlerini yaparken bazı hata fonksiyonları ile öğrenme sürecini yönetmektedir. Bu hata fonksiyonlarının tipi problemin cinsine göre farklılık göstermektedir. Bu husus öğrenme performansını etkileyen önemli kriterlerden birisidir. Bu hata fonksiyonlarından en sık kullanılan karesel ortalama hatadır [16].

$$MSE = \frac{1}{N} \sum_{i=1}^N (t_i - td_i)^2 \quad (4.1)$$

ile ifade edilir. İleri beslemeli ağlarda kullanılan hata fonksiyonlarından bir diğeri ise kare ortalamalarının karekökü

$$SSE = \sum_{i=1}^N (t_i - td_i)^2 \quad (4.2)$$

eşitliği ile hesaplanır. YSA yapılarında en sık kullanılan bir diğer hata fonksiyonu ise karesel ortalama hata karekökü fonksiyonudur. Bu fonksiyonda,

$$RMS = \sqrt{\frac{1}{N} \sum_{i=1}^N (t_i - td_i)^2} \quad (4.3)$$

eşitliği ile ifade edilir.

Test kümesinde hata değeri eğitim süresince kullanılmaz ama bu hata değeri farklı modellerin kıyaslanmasında faydalı olabilir. Test kümesinin hata fonksiyonunun grafiksel gösterimi yapısal tasarım aşamasında oldukça faydalıdır.

Bu tez çalışmasında geliştirilen Zeki Dil Tanıyıcı YSA modülü eğitimi gerçekleştirilirken takip edilen adımlar (Bkz. Şekil 4.5’de) aşağıda sırayla sunulmaktadır.

- (1) YSA yapısının oluşturulması
- (2) Eğitim veri setinin okunması ve normalize edilmesi
- (3) İlk ağırlık değerlerinin atanması ve
- (4) Eğitime başlanması,
- (5) Belirlenen epok sayısına kadar eğitimin devam ettirilmesi
- (6) Sonuçların değerlendirilmesi

Önerilen yöntemler incelenen dillerin alfabelerinin farklı sayılarda birleşmesinden oluşmaktadır. Bu sebeple YSA her yöntem için farklı giriş sayılarıyla yeniden tasarlanmıştır. Bu tasarımlar yapılırken Çizelge 4.2’de verildiği gibi farklı ağ yapıları deneme yanılma yoluyla test edilerek ortalama hata değerlerinin en aza indirgenmesi hedeflenmiştir.

Çizelge 4.2. YSA parametrelerinin karşılaştırma

YSA #	AKS	HKNS	TF	ÖA	MSE
1	2	5,5,1	S,TH,L	LM	3.5×10^0
2	2	10,30,1	S,TH,L	LM	8.9×10^{-22}
3	2	20,40,1	S,TH,L	LM	1.04×10^{-29}
4	2	30,40,1	S,TH,L	LM	9.97×10^{-24}
5	2	5,5,1	S,TH,L	GD	18.65×10^0
6	2	20,40,1	S,TH,L	GD	3.4×10^0
7	2	30,80,1	S,TH,L	GD	8.46×10^0
8	2	10,10,1	S,TH,L	GDM	18.66×10^0
9	2	20,20,1	S,TH,L	GDM	7.7×10^0
10	2	10,10,1	S,TH,L	GDA	8.25×10^0
11	2	30,50,1	S,TH,L	GDA	8.11×10^0
12	2	20,40,1	TH,S,L	LM	3.38×10^{-22}
13	3	20,30,40,1	S,TH,TH,L	LM	7.40×10^{-16}
14	3	20,30,40,1	S,S,TH,L	LM	2.28×10^{-26}
15	3	20,30,40,1	TH,S,TH,L	LM	3.54×10^{-19}

ÖA: Öğrenme Algoritması

AKS: Ara Katman Sayısı

HKNS: Her Katmanda Nöron Sayısı

TF: Transfer Fonksiyonu

TH: Tanjant Hiperbolik

S: Sigmoid

L: Doğrusal

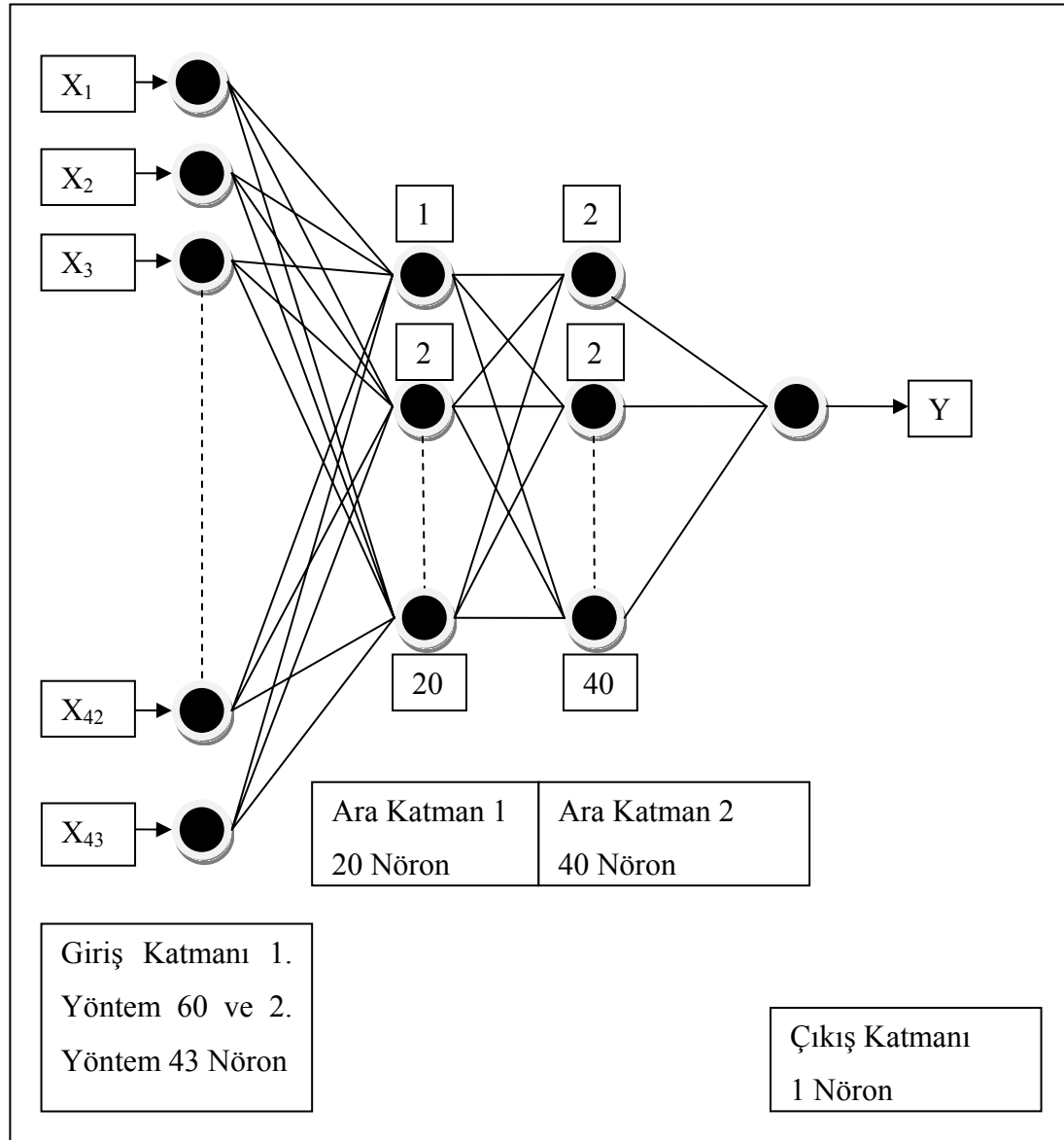
MSE: Ortalama Hataların Karekökü Toplamı

LM: Levenberg-Marquardt backpropagation

GD: Dereceli Azalan Geri Yayılım

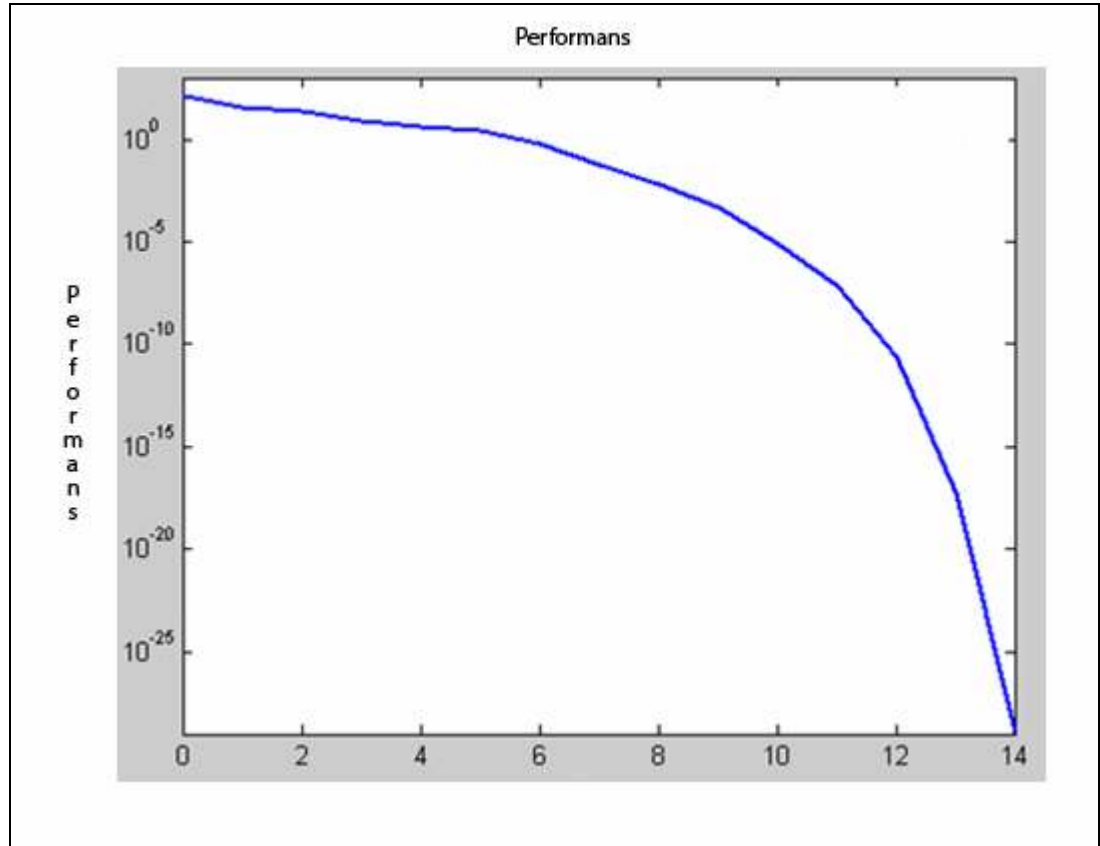
GDA: Adaptif Öğrenme Oranlı Dereceli Azalan Geri Yayılım

Bu tez çalışması sırasında kullanılan YSA yapılarından MLP tercih edilmiş ve 3 farklı YSA öğrenme algoritması ile YSA sınıflandırma performansı test edilmiştir. YSA performansında etkili olan epok sayısı, nöron sayısı ve geçiş fonksiyonu parametreleri değiştirilerek uygun parametreler bulunmaya çalışılmıştır.



Şekil 4.7. Kullanılan YSA Modeli

Şekil 4.7’de verilen YSA modeli Kesişim tespit yöntemi için kullanılmaktadır. Birleşim Tespit Yöntemi için Şekil 4.7’de verilen YSA modeli 43 giriş yerine 60 giriş olarak tasarlanmıştır.



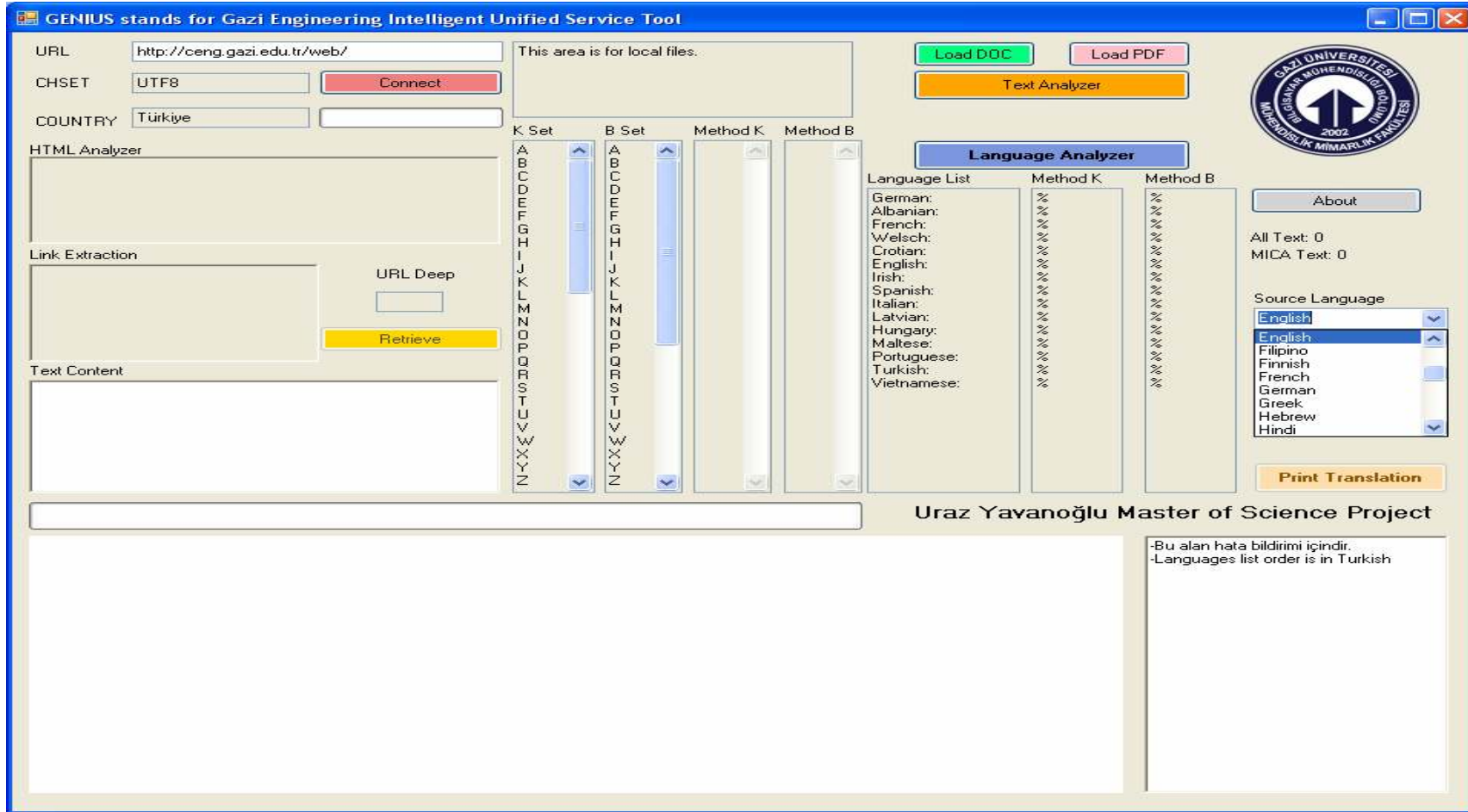
Şekil 4.8. YSA Eğitimi Performans vs. Ağırlık Değişim Sayısı Grafiği

YSA modelinin eğitilmesi için yapılan testler sonucunda 1. Ara katmanda sigmoid transfer fonksiyonuna sahip 20 nöronlu yapı ile 2. Ara katmanda tanjant hiperbolik transfer fonksiyonuna sahip 40 nöronlu yapı kullanılmıştır. Sistem sıfır hata oranına ulaşmak için 50 epok seviyesine kadar eğitime alınmıştır. Eğitim 0 epok seviyesinde ortalama hata kareleri toplamı 129.463, eğim ise $336.114/1e-010$ olarak ölçülmüştür. Eğitim 5. Epok seviyesine ulaştığı zaman ortalama hata kareleri toplamı 2.453 ve eğim $15.0804/1e-010$ olarak ölçülmüştür. Eğitim 10. Epok seviyesine ulaştığı zaman ortalama hata kareleri toplamı $8.11853e-006$ ve eğim $0.00743012/1e-010$ olarak ölçülmüştür. Eğitimin bu aşamasında hata oranının giderek düştüğü ve YSA modelinin uygulanan eğitim veri kümelerinden genelleme çıkartabildiği

görülmektedir. Hata oranı sıfıra yaklaşırken sistem varsayılan eğitim minimum eğitim düzeyine ulaşarak 14. Epokta seviyesinde ortalama hata kareleri toplamı $1.04089e-029$ ve eğitim $9.32831e-015/1e-010$ olarak ölçülmüştür. Alınan bu iyi sonuçlardan sonra eğitim işlemi sonlandırılmış ve tespit edilen bu parametrelerle sistemin eğitilmesi gerçekleştirilmiştir.

4.2.6. Geliştirilen platform yazılımı

Bu yazılım platformu C# dilinde Visual Studio 2005 ortamında sayısal dokümanları işlemek amacıyla açık kaynak kodlu otomasyon yazılımları kullanılarak tasarlanmıştır. Bu nedenle Adobe Reader PDF veya Microsoft Word DOC formatlarının işlenmesi amacıyla her iki yazılımında kullanıcının makinesinde bulunmasına ihtiyaç duymamaktadır. Yazılım genel olarak önceki bölümlerde anlatılan 6 modülden oluşmaktadır. Bu modüller Bölüm 4.5’de akış diyagramında verildiği gibi Sayısal Doküman Analiz Modülü (SDAM), Metin Otomasyon Modülü (MOM), Frekans Analiz Modülü (FAM), Zeki Dil Tanıma Modülü (ZDTM), Dil Geçişli Çeviri Modülü (DGÇM) ve Çıkış Modülü (ÇM) oluşmaktadır. Bu yazılımın adını İngilizce deha manasına gelen genius kelimesinden almaktadır. GENIUS (Gazi Engineering Intelligent Unified Service) bir diğer deyişle Gazi Mühendislik Zeki Birleşim Servisi uluslar arası çapta geliştirilmiş bir araçtır. Bu araç sayesinde kullanıcılar dillerini bilmedikleri internet sayfalarını, düz metinleri, Word ve PDF dokümanlarını kendi dillerine otomatik çevirebileceklerdir.



Şekil 4.9. GENIUS Gazi Mühendislik Zeki Birleşik Servisi (Gazi Engineering Intelligent Unified Service)

Bu bölümde ekran görüntüsü Şekil 4.9’da sunulan yazılım platformunun ara yüzleri ve kullanımı detaylı olarak açıklanmıştır.

GENIUS kullanıcı dostu olarak geliştirilmiş ve kullanıcıların üretilen sonuçları ekranda en uygun gösterecek şekilde tasarlanmıştır. Her kullanıcının sorunsuz çalıştırabileceği ve internet hizmetlerinden maksimum düzeyde yarar sağlayabileceği bir araç olarak geliştirme süreçlerinde ikinci geliştirilen yaklaşım kullanılmıştır. Bu yazılımı ilk yaklaşımdan ayıran en büyük farklılık geliştirilen platformun bir sunucu üzerinde değil kullanıcı makinesi üzerinde çalışmasıdır. Bu sayede ana sunucu gereksinimi ve bant genişliği ihtiyaçları ortadan kalkmıştır. Bu ihtiyaçların ortadan kalkmasıyla birlikte en sık kullanılan sayısal dokümanlarla entegrasyon ve çevirilerin bilgisayar sistemlerine kayıt edilerek saklanmaları gibi sunucu üzerinden gerçekleştirilmesi büyük maliyet taşıyan yeni özellikler eklenmiştir.

Şekil 4.9’da verilen servis yazılımı internet hizmet kısmıyla başlamaktadır. Bu kısım sayesinde kullanıcılar filleri hiç bilmedikleri internet sayfalarını 35 farklı dil desteği Google Translator yazılımı sayesinde kendi dillerine dönüştürebilmektedirler. Bu dönüşüm için gerekli olan altyapı desteği GENIUS tarafından sağlanmaktadır.

Kullanıcılar Şekil 4.9’da URL metin kutucuğuna dil tanıma sistemi tarafından tanınması sağlandıktan sonra otomatik çevirisi yapılacak olan sisteme giriş amacıyla internet sayfa adresi yazdıktan sonra tek yapmaları gereken bağlan (connect) butonuna basmalarıdır. Bu buton sayesinde çift yönlü çalışan bir mimari yapı kullanıcı için işlem karmaşıklığını azaltmaktadır. Bağlan butonuna basıldığında kullanıcı makinesine kendisine giriş olarak uygulanan internet sitesinin adresini varsayılan UTF-8 karakter kodlama kümesi ile bulunduğu alan barındırma hizmeti veren sunucudan almaktadır. Bu işlem ile kaynak koduna ulaşılan sayfaya çeşitli örüntü tanıma filtreleri uygulanarak sayfanın gerçek karakter kodlama kümesi bulunmaktadır. Çevirisi yapılmak istenilen sayfanın karakter kodlama kümesi bulunduğundan sonra UTF-8 ile tanınamayan karakterleri de sisteme almak amacıyla bulunan karakter kodlama kümesiyle sayfa ilgili alan barındırma sunucusundan

tekrar alınmaktadır. Sunucudan sorunsuz şekilde kullanıcı makinesine alınan sayfaya ait kaynak koduna HTML etiketleri temizleme işlemleri uygulandıktan sonra Sayfa Analizi (HTML Analyzer) ile etiketlenmiş olan metin kutusuna gönderilmektedir. Bu yaklaşım ilk yazılım yaklaşımından farklı olarak Bağlantı Çıkart (Link Extraction) etiketiyle işaretlenmiş metin kutusu içine çeviri yapılmak istenilen sayfaya ait bağlantıları getirmektedir. Kullanıcı yukarıda bulunan Bağlantı Derinliği (URL Deep) metin kutusuna sayfanın içerdiği bağlantı sayısı kadar bağlantı derinliği girebilmektedir. Bu sayede sayfanın dilinin tanınması için gerekli olan metin boyutu büyüdüğünden daha kararlı sonuçlar almak mümkün olmaktadır. Son olarak getir (Retrieve) butonu ile sayfa kullanıcı tarafından girilen bağlantı sayısı derinliği kadar ilk başta bulunan karakter kodlama kümesi ile kullanıcı makinesine çağırılarak ilk yazılım yaklaşımda benimsenen HTML etiket filtreleme işlemine tabi tutulmaktadır. Bu işlem sayfanın sol alt köşesinde bulunan durum çubuğu çalışarak istenilen sayfa derinliği kadar metin verisi elde edilmektedir. Bu yazılımın internet sayfalarıyla çalışan bölümüdür. Bu yazılım ayrıca Microsoft Word ve Adobe Reader sayısal dokümanlarıyla da çalışabilmektedir.

Kullanıcı penceresinin sol tarafında DOC Yükle (Load DOC) ve PDF Yükle (Load PDF) butonları görülmektedir. DOC Yükle (Load DOC) butonu Microsoft Word yazılımının dosya formatı olan DOC dosyalarını, PDF Yükle (Load PDF) ise Adobe Reader programının dosya formatı olan PDF dosyalarını sisteme yüklemektedir. Bu butonlara basıldığı zaman Göster (Browse) penceresi açılarak kullanıcıdan giriş olarak uygulanabilecek DOC ve PDF uzantılı dosyalar istenmektedir. Kullanıcı sisteme DOC yada PDF dosyası seçimi yaptığı zaman daha önceki bölümlerde anlatılan Microsoft Word Dosya Otomasyon modülü ile Adobe Reader Dosya Otomasyon modülü devreye girmektedir. Bu modüller açık kaynak kodlu otomasyon yazılımlarıdır. Bu ara katman yazılımları sayesinde DOC ve PDF uzantılı dosyalar kullanıcılara ilgili programları yükleme zorunluluğu getirmemektedir. Bu alt kütüphaneler kullanılarak geliştirdiğimiz otomasyon yazılımları PDF ve DOC uzantılı dosyalardan metin verisi çekebilmektedir. Bu sayede dokümanlardan çekilen metin bilgisi aynı internet sayfası metin bilgisi sonucunda olduğu gibi Şekil 4.9'da

sayfanın sol alt kısmında bulunan Metin İçeriği (Text Content) etiketiyle verilmiş olan metin kutusuna basılmaktadır. Elbette, internet sayfa bağlantısı ile DOC veya PDF uzantılı olmayan düz metin bilgisi de sisteme giriş olarak uygulanabilmektedir. Buraya kadar anlatılan bölüm Doküman Analiz Modülü (SDAM), Metin Otomasyon Modülü (MOM) kapsamına girmektedir.

Geliştirilen otomasyon modüllerinden elde edilen ham verinin YSA modeline uygulanabilmesi için çeşitli ön işlemlerden geçmesi gerekmektedir. Bu ön işlemler sonucunda metin bilgisi dili oluşturan yapı taşlarına bölünmektedir. Bir dili oluşturan temel yapı taşı reel olarak harflerdir. EK-16 ve EK-17 ile EK-18 ve EK-19 de verilen harf kümesi çizelgeleri bu yazılım tarafından tanınması istenilen 15 farklı dil için ortak bir alfabe desteği sunmaktadır. Bu ortak alfabe desteği önerilen birleşim ve kesişim yöntemleri sayesinde ortaya çıkmaktadır. Bu yöntemlere ait detaylı bilgi önceki bölümlerde verilmiştir. Bu bölümlerin yazılım üzerinde görüntülenmesi amacıyla 15 dilin birleşim ve kesişim yöntemleri için elde edilen alfabe kümeleri sırasıyla 60 ve 43 harfli kümeler yazılım üzerinde kesişim (K) ve birleşim (B) kümeleri (Set) şeklinde kullanıcıya sunulmaktadır. İnternet sayfası girişiyle sayısal doküman girişi arasında bulunan metin kutusu sayısal doküman yüklemesi yapıldığı zaman dokümana ait dosya yolunun görüntülenmesini sağlamakta bu sayede kullanıcı mevcut çeviri dosyalarının hangi alt dizine kayıt edildiğini görebilmektedir.

Sayısal doküman otomasyon modülün hemen altında Metin Analizi (Text Analyzer) butonu bulunmaktadır. Bu buton sayesinde geliştirdiğimiz metotlar ile giriş metni üretilen alfabe kümeleri üzerine uygulanarak giriş metni içinde alfabe kümesindeki harflerin geçme yüzdeleri kısaca harflerin frekans analizleri yapılmaktadır. Frekans analizi neticesinde elde edilen yüzdeler K ve B harf kümesi olarak etiketlenen metin kutularının karşılıklarına gelen Method K ve Method B frekans çıktısı kutucuklarına her harf bir yüzdeler dilim olarak basılmaktadır. Bu sayede YSA modeline giriş olarak uygulanacak eğitim seti veya test seti ortaya çıkmaktadır. Önceki bölümlerde anlatıldığı gibi elde edilen sayısal dokümanlar bu yöntemle işlenerek EK 1-15 arasında Kesişim Giriş kümesi örneklemeyle verilen, Kesişim ve

Birleşim giriş veri kümeleri oluşturulmuştur. Bu Kesişim ve Birleşim metotlarıyla eğitilen YSA'nın test aşamasında girişine hesaplanan yüzdelik değerlerin gönderilmesi için kullanıcı tarafından Dil Analizi (Language Analyzer) butonuna basılması gerekmektedir. Bu esnada C# programlama dili Matlab programıyla dinamik bağlantı kütüphanesi düzeyinde haberleşerek daha önceden Matlab programı üzerinde oluşturulan ve eğitilen YSA modeline giriş olarak uygulanmaktadır. Windows işletim sistemlerinde alışık olduğumuz nokta net çalıştırma kütüphanesi (dot net framework) yapısında Matlab çalıştırma zamanı kütüphaneleriyle (matlab compact framework) YSA test metotlarını çağırarak giriş karşılık YSA'nın çıkışını üretmekte yine aynı yolu izleyerek C# programlama ara yüzüne ulaşmaktadır. C# ara yüzüne ulaşan çıkış sonucu olması gereken hata sonuçları ile karşılaştırılarak yüzdelik hata değerleri hesaplanmaktadır. Bu hesaplama işleminde Almanca (1), Arnavutça (2), Fransızca (3), Galce (4), Hırvatça (5), İngilizce (6), İrlandaca (7), İspanyolca (8), İtalyanca (9), Letonca (10), Macarca (11), Maltaca (12), Portekizce (13), Türkçe (14), Vietnamca (15) dilleri parantezler içinde verilen referans numara değerlerinin sayısal karşılıkları kullanılmaktadır.

Bu yazılım tarafından Latin alfabesine mensup 15 farklı dil kullanılmıştır. Bu diller Dil Listesi (Language List) etiketi altında bulunan metin kutusunda eğitim sırasıyla verilmiştir. Latin Alfabesi 80'den fazla dile sahip çok geniş bir kitleye hitap eden yapıya sahiptir. Bu sebeple hazırlanan yazılımın esnekliği belirli bir coğrafi konumda konuşulan dillerin sisteme entegrasyon süresinin kısılmasını sağlamaktadır. Dil Listesi solunda verilen 2 farklı metin kutusu Birleşim ve Kesişim metotlarını temel alarak YSA'nın sonucunu varsayılan dil çıkış sonuçları ile karşılaştırmaktadır. Bu alanlarda her metot için varılan sonuçların hata oranları hesaplanarak başarı sıraları ekrana otomatik olarak basılmaktadır.

Dil tanıma işlemi neticesinde dil tanıyıcı tarafından tanınan sayısal dokümanın dili en solda bulunan Kaynak Dil (Source Language) ile etiketlenmiş açılır listeden otomatik olarak seçilmektedir. Bu kısım direkt olarak Google Çevirici (Google Translator) ile entegre çalışmaktadır. Kullanıcıların herhangi bir Google hesabı olması

gerekmemektedir. Yazılım gerekli çeviri içeriklerini kendi alt dönüşümleriyle kullanıcılara iletmektedir. Kaynak Dil etiketinin altında Hedef Dil (Destination Language) açılan kutu nesnesi bulunmaktadır. Kullanıcılar kaynak dil bilindikten sonra hedef dili seçerek Çevir (Translate) butonuna tıkladıklarında bir internet sayfası açılarak sisteme giriş olarak uygulanan sayısal dokümanın çeviri bir internet sayfasında kullanıcılara sunulmaktadır. Çevir ve Kaydet (Translate and Save) butonu tıklanırsa ilgili çeviri metinleri dil tanınması ve çevirisi istenilen sayısal dokümanlar ile aynı alt klasöre HTML dosyası olarak kopyalanmaktadır. Bu sayede DOC yada PDF uzantılı dosyalar her bilgisayar sistemi tarafından okunabilen HTML dosyasına kullanıcı tarafından anlaşılan dil çevirisi ile kayıt altına alınmaktadır.

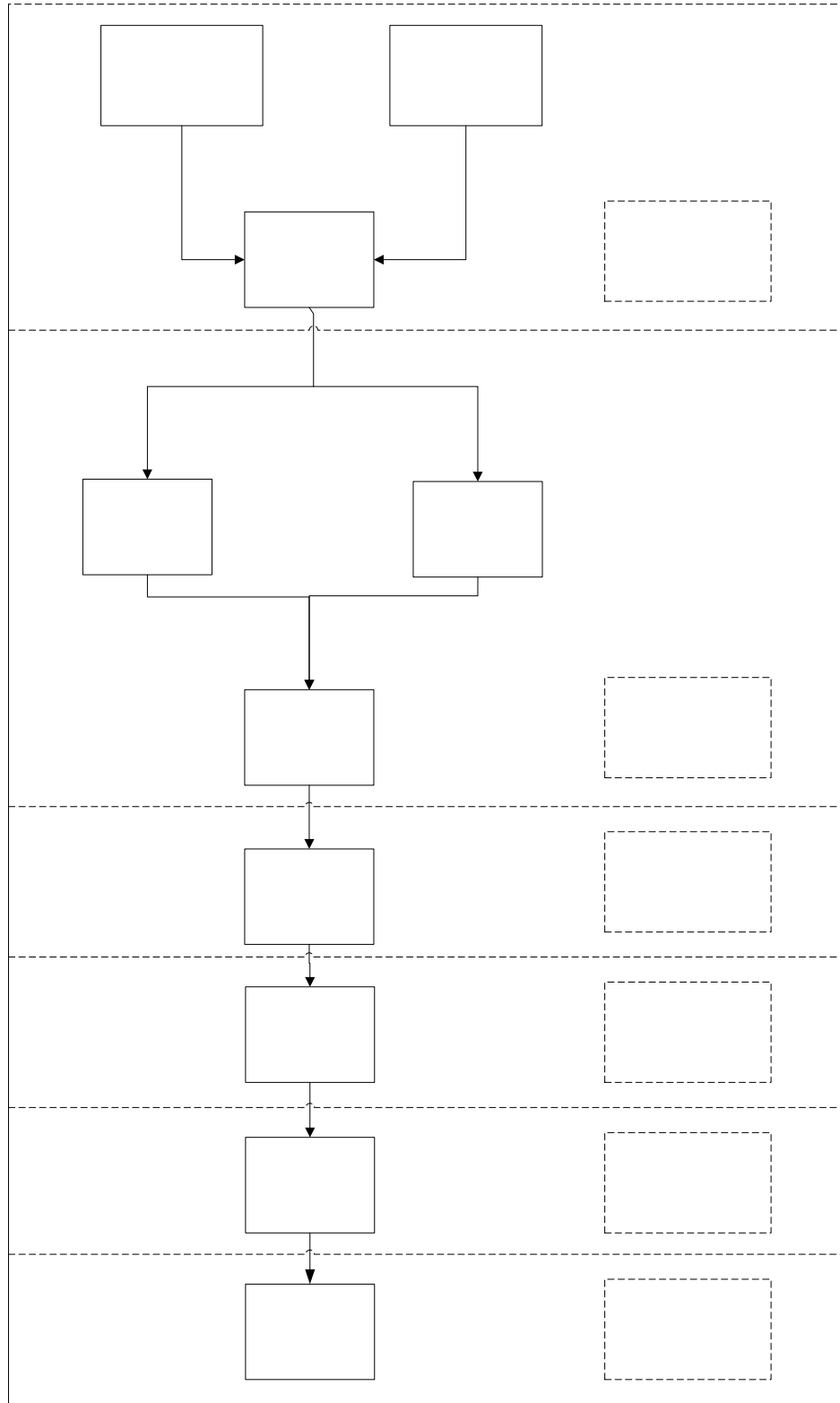
Böylece kullanıcılar hiç dillerini bilmedikleri dokümanların dillerine ve istedikleri dile çevirilerine sadece tek bir tıklama ile ulaşabilmektedir. Bu yazılımın çalışması betimlendiği için tek tıklama işlemi farklı parçalar halinde sunulmaktadır. Bu yazılım internet sitelerini, format bozulması yaşanmadan yeni bir web göstericisi açarak kullanıcılara sunmaktadır.

Geliştirilen GENIUS isimli yazılımın doküman içeriklerinin tespit yüzdelerini ortaya koymak için iki farklı test çalışması yapılmıştır. Testin birinci bölümünde (Test #1) Microsoft Word ve Adobe Reader tabanlı dokümanlar test edilmiştir. İkinci kısımda (Test #2) ise web sayfalarının testleri yapılmıştır. Bu testler farklı başlıklar altında açıklanmıştır. Bu testte ise HTML, PHP, JSP ve ASP tabanlı dokümanlar temel alınmıştır.

4.3. Test 1

Bu testin yapılması için geliştirilen GENIUS yazılım platformu kullanılmıştır. Sistemin herhangi bir içerikten bağımsız veriler ile çalışabilmesi için Google arama motorunda sayısal dokümanlarda herhangi bir konudan bağımsız geçebilecek olan dünya, insan, bilim, politika vb. kelimeler ile bölgesel arama seçenekleri kullanılarak Adobe Reader ve Microsoft Word dosyaları araştırılmıştır. Bulunan sayısal

dokümanlar hem içerik olarak hem de kelime sayısı, kodlama türü, tablo, resim vb. özellikler bakımından benzerlik bulunmamaktadır. Bu çalışmada iki önemli doküman formatı olan DOC ve PDF dikkate alınmıştır.



Şekil 4.10. Word ve PDF dokümanları için sistem modeli blok diyagramı

Kulla
Micr
D

Testlerin sağlıklı yapılması için, Almanca, Arnavutça, Fransızca, Galce, Hırvatça, İngilizce, İrlandaca, İspanyolca, İtalyanca, Letonca, Macarca, Maltaca, Portekizce, Türkçe ve Vietnamca dilleri için web ortamından rasgele seçilmiş 5 Adobe Reader dosyası ve 5 Microsoft Word dosyası kullanılmıştır. Sonuçlar her dil için alfabetik olarak sıralanmış ve elde edilen en yüksek başarı oranları altı çizili olarak yazılmıştır. Yüzde başarı oranları YSA çıkış değeri ve olması gereken sayısal sonuç arasındaki hata farkı ile hesaplanmaktadır. Bu sebeple her dil için seçilen dokümanlar üzerinde 300 farklı test yapılmıştır. Yapılan testlerde elde edilen 4500 sonuç EK-1 – EK-40 arasında sunulmuş olmakla birlikte her dile özel sonuçlar takip eden satırlarda örneklendirilerek anlatılmıştır. İlgili çizelgelerde verilen başarı oranları daha önceki bölümlerde anlatılan ve 1. Yöntem (#1) olan Kesişim Tespit yöntemi ve 2. Yöntem (#2) olan Birleşim Tespit yöntemi dikkate alınarak elde edilen sonuçları yansıtmaktadır.

Geliştirilen yazılım sayısal dokümanlar için Şekil 4.5’de verilen akış diyagramını çalışma zamanında Şekil 4.10’da PDF ve DOC dosyaları için Şekil 4.11’de HTML tabanlı dosyalar için özelleştirmektedir. Şekil 4.10’da verilen sayısal doküman tanıma modülü sisteme giriş olarak uygulanan Microsoft Word ve Adobe PDF dosyalarını takip eden modüllerde işledikten sonra tespit edilen dili istenilen dile dönüşümünü yaparak yazılım içine gömülmüş olan internet göstericisinde ilgili Microsoft Word yada Adobe PDF sayısal dokümanına ait çeviri metnini göstermektedir. Kullanıcılar bu çevirilerin yazıcı çıktılarına da erişebilmektedir.

Almanca

Almanca olan dokümanlarla ilgili olarak yapılan testlerde elde edilen sonuçlar Çizelge 4.3’de verilmiştir. Elde edilen sonuçlar her iki doküman tipi ve her iki tespit yöntemi kullanılarak yapılmıştır. Elde edilen sonuçlardan Microsoft Word DOC ve Adobe PDF doküman tipleri için 1. Yöntem (#1) ile %95 ve 2. Yöntem (#) ile %90 oranında bir başarı elde edilmiştir. Aynı tablodan görülebileceği gibi sonuçların Almanca olması beklenirken elde edilen sonuçlarda Almancadır. Bu sebeple almanca

için her iki dosya formatı ve her iki yönteminde %100 başarılı olduğu görülmüştür. Almanca dokümanlar ile diğer dillere ait olan sonuçların detayları Çizelge 4.3’de verilmiştir.

Çizelge 4.3. Almanca için doküman bazında başarı oranları

Dil	Almanca			
Format	DOC Dosya Tipi		PDF Dosya Tipi	
Yöntem	#1	#2	#1	#2
Almanca	92%	83%	97%	96%
Arnavutça	53%	58%	51%	48%
Fransızca	35%	38%	34%	32%
Galce	26%	29%	25%	24%
Hırvatça	21%	23%	20%	19%
İngilizce	17%	19%	17%	16%
İrlandaca	15%	16%	14%	13%
İspanyolca	13%	14%	12%	12%
İtalyanca	11%	12%	11%	10%
Letonca	10%	11%	10%	9%
Macarca	9%	10%	9%	8%
Maltaca	8%	9%	8%	8%
Portekizce	8%	8%	7%	7%
Türkçe	7%	8%	7%	6%
Vietnamca	7%	7%	6%	6%

Arnavutça

Arnavutça olan dokümanlarla ilgili olarak yapılan testlerde elde edilen sonuçlar Çizelge 4.4’de verilmiştir. Elde edilen sonuçlar her iki doküman tipi ve her iki tespit yöntemi kullanılarak yapılmıştır. Elde edilen sonuçlardan Microsoft Word DOC ve Adobe PDF doküman tipleri için 1. Yöntem (#1) ile %99 ve 2. Yöntem (#2) ile %99 oranında bir başarı elde edilmiştir. Aynı tablodan görülebileceği gibi sonuçların Arnavutça olması beklenirken elde edilen sonuçlarda Arnavutçadır. Bu sebeple Arnavutça için her iki dosya formatı ve her iki yönteminde %100 başarılı olduğu görülmüştür. Arnavutça dokümanlar ile diğer dillere ait olan sonuçların detayları Çizelge 4.4’de verilmiştir.

Çizelge 4.4. Arnavutça için doküman bazında başarı oranları

Dil	Arnavutça			
	DOC Dosya Tipi		PDF Dosya Tipi	
Yöntem	#1	#2	#1	#2
Almanca	1%	0%	0%	0%
Arnavutça	99%	99%	99%	99%
Fransızca	66%	66%	66%	66%
Galce	49%	49%	49%	50%
Hırvatça	39%	39%	39%	40%
İngilizce	33%	33%	33%	33%
İrlandaca	28%	28%	28%	28%
İspanyolca	24%	24%	24%	25%
İtalyanca	22%	22%	22%	22%
Letonca	19%	19%	19%	20%
Macarca	18%	18%	18%	18%
Maltaca	16%	16%	16%	16%
Portekizce	15%	15%	15%	15%
Türkçe	14%	14%	14%	14%
Vietnamca	13%	13%	13%	13%

Fransızca

Fransızca olan dokümanlarla ilgili olarak yapılan testlerde elde edilen sonuçlar Çizelge 4.5’de verilmiştir. Elde edilen sonuçlar her iki doküman tipi ve her iki tespit yöntemi kullanılarak yapılmıştır. Elde edilen sonuçlardan Microsoft Word DOC ve Adobe PDF doküman tipleri için 1. Yöntem (#1) ile %93 ve 2. Yöntem (#2) ile %90 oranında bir başarı elde edilmiştir. Aynı tablodan görülebileceği gibi sonuçların Fransızca olması beklenirken elde edilen sonuçlarda Fransızcadır. Bu sebeple Fransızca için her iki dosya formatı ve her iki yönteminde %100 başarılı olduğu görülmüştür. Fransızca dokümanlar ile diğer dillere ait olan sonuçların detayları Çizelge 4.5’de verilmiştir.

Çizelge 4.5. Fransızca için doküman bazında başarı oranları

Dil	Fransızca			
Format	DOC Dosya Tipi		PDF Dosya Tipi	
Yöntem	#1	#2	#1	#2
Almanca	0%	0%	0%	0%
Arnavutça	47%	33%	31%	61%
Fransızca	98%	89%	87%	92%
Galce	76%	83%	84%	69%
Hırvatça	60%	66%	67%	55%
İngilizce	50%	55%	56%	46%
İrlandaca	43%	47%	48%	39%
İspanyolca	38%	41%	42%	34%
İtalyanca	33%	36%	37%	30%
Letonca	30%	33%	33%	27%
Macarca	27%	30%	30%	25%
Maltaca	25%	27%	28%	23%
Portekizce	23%	25%	25%	21%
Türkçe	21%	23%	24%	19%
Vietnamca	20%	22%	22%	18%

Galce

Galce olan dokümanlarla ilgili olarak yapılan testlerde elde edilen sonuçlar Çizelge 4.6'da verilmiştir. Elde edilen sonuçlar her iki doküman tipi ve her iki tespit yöntemi kullanılarak yapılmıştır. Elde edilen sonuçlardan Microsoft Word DOC ve Adobe PDF doküman tipleri için 1. Yöntem (#1) ile %99 ve 2. Yöntem (#2) ile %98 oranında bir başarı elde edilmiştir. Aynı tablodan görülebileceği gibi sonuçların Galce olması beklenirken elde edilen sonuçlarda Galce'dir. Bu sebeple Galce için her iki dosya formatı ve her iki yönteminde %100 başarılı olduğu görülmüştür. Galce dokümanlar ile diğer dillere ait olan sonuçların detayları Çizelge 4.6'da verilmiştir.

Çizelge 4.6. Galce için doküman bazında başarı oranları

Dil	Galce			
Format	DOC Dosya Tipi		PDF Dosya Tipi	
Yöntem	#1	#2	#1	#2
Almanca	0%	0%	0%	0%
Arnavutça	0%	0%	0%	0%
Fransızca	65%	61%	66%	66%
Galce	99%	96%	100%	100%
Hırvatça	80%	82%	80%	80%
İngilizce	67%	69%	66%	66%
İrlandaca	57%	59%	57%	57%
İspanyolca	50%	51%	50%	50%
İtalyanca	44%	46%	44%	44%
Letonca	40%	41%	40%	40%
Macarca	36%	37%	36%	36%
Maltaca	33%	34%	33%	33%
Portekizce	30%	31%	30%	30%
Türkçe	28%	29%	28%	28%
Vietnamca	26%	27%	26%	26%

Hırvatça

Hırvatça olan dokümanlarla ilgili olarak yapılan testlerde elde edilen sonuçlar Çizelge 4.7’de verilmiştir. Elde edilen sonuçlar her iki doküman tipi ve her iki tespit yöntemi kullanılarak yapılmıştır. Elde edilen sonuçlardan Microsoft Word DOC ve Adobe PDF doküman tipleri için 1. Yöntem (#1) ile %97 ve 2. Yöntem (#2) ile %97 oranında bir başarı elde edilmiştir. Aynı tablodan görülebileceği gibi sonuçların Hırvatça olması beklenirken elde edilen sonuçlarda Hırvatçadır. Bu sebeple Hırvatça için her iki dosya formatı ve her iki yönteminde %100 başarılı olduğu görülmüştür. Hırvatça dokümanlar ile diğer dillere ait olan sonuçların detayları Çizelge 4.7’de verilmiştir.

Çizelge 4.7. Hırvatça için doküman bazında başarı oranları

Dil	Hırvatça			
Format	DOC Dosya Tipi		PDF Dosya Tipi	
Yöntem	#1	#2	#1	#2
Almanca	0%	0%	0%	0%
Arnavutça	0%	0%	0%	0%
Fransızca	37%	42%	28%	33%
Galce	78%	82%	71%	74%
Hırvatça	97%	94%	97%	99%
İngilizce	81%	78%	85%	83%
İrlandaca	69%	67%	73%	71%
İspanyolca	60%	58%	64%	62%
İtalyanca	54%	52%	57%	55%
Letonca	48%	47%	51%	50%
Macarca	44%	42%	46%	45%
Maltaca	40%	39%	42%	41%
Portekizce	37%	36%	39%	38%
Türkçe	34%	33%	36%	35%
Vietnamca	32%	31%	34%	33%

İngilizce

İngilizce olan dokümanlarla ilgili olarak yapılan testlerde elde edilen sonuçlar Çizelge 4.8’de verilmiştir. Elde edilen sonuçlar her iki doküman tipi ve her iki tespit yöntemi kullanılarak yapılmıştır. Elde edilen sonuçlardan Microsoft Word DOC ve Adobe PDF doküman tipleri için 1. Yöntem (#1) ile %98 ve 2. Yöntem (#2) ile %98 oranında bir başarı elde edilmiştir. Aynı tablodan görülebileceği gibi sonuçların İngilizce olması beklenirken elde edilen sonuçlarda İngilizcedir. Bu sebeple İngilizce için her iki dosya formatı ve her iki yönteminde %100 başarılı olduğu görülmüştür. İngilizce dokümanlar ile diğer dillere ait olan sonuçların detayları Çizelge 4.8’de verilmiştir.

Çizelge 4.8. İngilizce için doküman bazında başarı oranları

Dil	İngilizce			
Format	DOC Dosya Tipi		PDF Dosya Tipi	
Yöntem	#1	#2	#1	#2
Almanca	0%	0%	0%	0%
Arnavutça	0%	0%	0%	0%
Fransızca	0%	0%	0%	4%
Galce	47%	50%	50%	53%
Hırvatça	78%	80%	80%	82%
İngilizce	98%	99%	99%	97%
İrlandaca	86%	85%	85%	83%
İspanyolca	76%	74%	74%	73%
İtalyanca	67%	66%	66%	65%
Letonca	60%	59%	59%	58%
Macarca	55%	54%	54%	53%
Maltaca	50%	49%	49%	48%
Portekizce	46%	46%	46%	45%
Türkçe	43%	42%	42%	41%
Vietnamca	40%	39%	39%	39%

İrlandaca

İrlandaca olan dokümanlarla ilgili olarak yapılan testlerde elde edilen sonuçlar Çizelge 4.9’da verilmiştir. Elde edilen sonuçlar her iki doküman tipi ve her iki tespit yöntemi kullanılarak yapılmıştır. Elde edilen sonuçlardan Microsoft Word DOC ve Adobe PDF doküman tipleri için 1. Yöntem (#1) ile %98 ve 2. Yöntem (#2) ile %98 oranında bir başarı elde edilmiştir. Aynı tablodan görülebileceği gibi sonuçların İrlandaca olması beklenirken elde edilen sonuçlarda İrlandaca’dır. Bu sebeple İrlandaca için her iki dosya formatı ve her iki yönteminde %100 başarılı olduğu görülmüştür. İrlandaca dokümanlar ile diğer dillere ait olan sonuçların detayları Çizelge 4.9’da verilmiştir.

Çizelge 4.9. İrlandaca için doküman bazında başarı oranları

Dil	İrlandaca			
Format	DOC Dosya Tipi		PDF Dosya Tipi	
Yöntem	#1	#2	#1	#2
Almanca	0%	0%	0%	0%
Arnavutça	0%	0%	0%	0%
Fransızca	0%	0%	0%	0%
Galce	28%	26%	25%	21%
Hırvatça	62%	61%	60%	56%
İngilizce	85%	84%	83%	80%
İrlandaca	98%	99%	99%	97%
İspanyolca	85%	86%	87%	89%
İtalyanca	76%	77%	77%	79%
Letonca	68%	69%	69%	71%
Macarca	62%	63%	63%	65%
Maltaca	57%	57%	58%	59%
Portekizce	52%	53%	53%	55%
Türkçe	49%	49%	49%	51%
Vietnamca	45%	46%	46%	47%

İspanyolca

İspanyolca olan dokümanlarla ilgili olarak yapılan testlerde elde edilen sonuçlar Çizelge 4.10'da verilmiştir. Elde edilen sonuçlar her iki doküman tipi ve her iki tespit yöntemi kullanılarak yapılmıştır. Elde edilen sonuçlardan Microsoft Word DOC ve Adobe PDF doküman tipleri için 1. Yöntem ile %97 ve 2. Yöntem ile %98 oranında bir başarı elde edilmiştir. Aynı tablodan görülebileceği gibi sonuçların İspanyolca olması beklenirken elde edilen sonuçlarda İspanyolca'dır. Bu sebeple İspanyolca için her iki dosya formatı ve her iki yönteminde %100 başarılı olduğu görülmüştür. İspanyolca dokümanlar ile diğer dillere ait olan sonuçların detayları Çizelge 4.10'da verilmiştir.

Çizelge 4.10. İspanyolca için doküman bazında başarı oranları

Dil	İspanyolca			
Format	DOC Dosya Tipi		PDF Dosya Tipi	
Yöntem	#1	#2	#1	#2
Almanca	0%	0%	0%	0%
Arnavutça	0%	0%	0%	0%
Fransızca	0%	0%	0%	0%
Galce	0%	4%	0%	0%
Hırvatça	39%	43%	32%	40%
İngilizce	65%	69%	60%	66%
İrlandaca	85%	88%	80%	85%
İspanyolca	99%	97%	95%	99%
İtalyanca	89%	87%	92%	88%
Letonca	80%	78%	83%	79%
Macarca	73%	71%	76%	72%
Maltaca	67%	65%	69%	66%
Portekizce	61%	60%	64%	61%
Türkçe	57%	55%	59%	57%
Vietnamca	53%	52%	55%	53%

İtalyanca

İtalyanca olan dokümanlarla ilgili olarak yapılan testlerde elde edilen sonuçlar Çizelge 4.11’de verilmiştir. Elde edilen sonuçlar her iki doküman tipi ve her iki tespit yöntemi kullanılarak yapılmıştır. Elde edilen sonuçlardan Microsoft Word DOC ve Adobe PDF doküman tipleri için 1. Yöntem (#1) ile %96 ve 2. Yöntem (#2) ile %99 oranında bir başarı elde edilmiştir. Aynı tablodan görülebileceği gibi sonuçların İtalyanca olması beklenirken elde edilen sonuçlarda İtalyanca’dır. Bu sebeple İtalyanca için her iki dosya formatı ve her iki yönteminde %100 başarılı olduğu görülmüştür. İtalyanca dokümanlar ile diğer dillere ait olan sonuçların detayları Çizelge 4.11’de verilmiştir.

Çizelge 4.11. İtalyanca için doküman bazında başarı oranları

Dil	İtalyanca			
Format	DOC Dosya Tipi		PDF Dosya Tipi	
Yöntem	#1	#2	#1	#2
Almanca	0%	0%	0%	0%
Arnavutça	0%	0%	0%	0%
Fransızca	0%	0%	0%	0%
Galce	0%	0%	0%	0%
Hırvatça	10%	19%	23%	19%
İngilizce	42%	49%	52%	49%
İrlandaca	64%	70%	73%	71%
İspanyolca	81%	87%	89%	87%
İtalyanca	94%	99%	98%	99%
Letonca	94%	90%	88%	90%
Macarca	86%	82%	80%	81%
Maltaca	78%	75%	73%	75%
Portekizce	72%	69%	67%	69%
Türkçe	67%	64%	63%	64%
Vietnamca	63%	60%	58%	60%

Letonca

Letonca olan dokümanlarla ilgili olarak yapılan testlerde elde edilen sonuçlar Çizelge 4.12’de verilmiştir. Elde edilen sonuçlar her iki doküman tipi ve her iki tespit yöntemi kullanılarak yapılmıştır. Elde edilen sonuçlardan Microsoft Word DOC ve Adobe PDF doküman tipleri için 1. Yöntem (#1) ile %96 ve 2. Yöntem (#2) ile %99 oranında bir başarı elde edilmiştir. Aynı tablodan görülebileceği gibi sonuçların Letonca olması beklenirken elde edilen sonuçlarda Letonca’dır. Bu sebeple Letonca için her iki dosya formatı ve her iki yönteminde %100 başarılı olduğu görülmüştür. Letonca dokümanlar ile diğer dillere ait olan sonuçların detayları Çizelge 4.12’de verilmiştir.

Çizelge 4.12. Letonca için doküman bazında başarı oranları

Dil	Letonca			
Format	DOC Dosya Tipi		PDF Dosya Tipi	
Yöntem	#1	#2	#1	#2
Almanca	0%	0%	0%	0%
Arnavutça	0%	0%	0%	0%
Fransızca	0%	0%	0%	0%
Galce	0%	0%	0%	0%
Hırvatça	0%	0%	0%	0%
İngilizce	26%	33%	28%	33%
İrlandaca	51%	57%	52%	57%
İspanyolca	69%	74%	71%	74%
İtalyanca	84%	88%	85%	88%
Letonca	95%	99%	96%	99%
Macarca	94%	90%	93%	90%
Maltaca	86%	83%	85%	83%
Portekizce	80%	76%	79%	76%
Türkçe	74%	71%	73%	71%
Vietnamca	69%	66%	68%	66%

Macarca

Macarca olan dokümanlarla ilgili olarak yapılan testlerde elde edilen sonuçlar Çizelge 4.13’de verilmiştir. Elde edilen sonuçlar her iki doküman tipi ve her iki tespit yöntemi kullanılarak yapılmıştır. Elde edilen sonuçlardan Microsoft Word DOC ve Adobe PDF doküman tipleri için 1. Yöntem (#1) ile %98 ve 2. Yöntem (#2) ile %97 oranında bir başarı elde edilmiştir. Aynı tablodan görülebileceği gibi sonuçların Macarca olması beklenirken elde edilen sonuçlarda Macarca’dır. Bu sebeple Macarca için her iki dosya formatı ve her iki yönteminde %100 başarılı olduğu görülmüştür. Macarca dokümanlar ile diğer dillere ait olan sonuçların detayları Çizelge 4.13’de verilmiştir.

Çizelge 4.13. Macarca için doküman bazında başarı oranları

Dil	Macarca			
Format	DOC Dosya Tipi		PDF Dosya Tipi	
Yöntem	#1	#2	#1	#2
Almanca	0%	0%	0%	0%
Arnavutça	0%	0%	0%	0%
Fransızca	0%	0%	0%	0%
Galce	0%	0%	0%	0%
Hırvatça	0%	0%	0%	0%
İngilizce	22%	12%	16%	21%
İrlandaca	47%	39%	42%	46%
İspanyolca	66%	59%	62%	65%
İtalyanca	81%	74%	77%	80%
Letonca	93%	87%	90%	92%
Macarca	96%	97%	99%	97%
Maltaca	88%	93%	91%	89%
Portekizce	82%	86%	84%	82%
Türkçe	76%	80%	78%	76%
Vietnamca	71%	75%	73%	71%

Maltaca

Maltaca olan dokümanlarla ilgili olarak yapılan testlerde elde edilen sonuçlar Çizelge 4.14’de verilmiştir. Elde edilen sonuçlar her iki doküman tipi ve her iki tespit yöntemi kullanılarak yapılmıştır. Elde edilen sonuçlardan Microsoft Word DOC ve Adobe PDF doküman tipleri için 1. Yöntem (#1) ile %100 ve 2. Yöntem (#2) ile %100 oranında bir başarı elde edilmiştir. Aynı tablodan görülebileceği gibi sonuçların Maltaca olması beklenirken elde edilen sonuçlarda Maltaca’dır. Bu sebeple Maltaca için her iki dosya formatı ve her iki yönteminde %100 başarılı olduğu görülmüştür. Maltaca dokümanlar ile diğer dillere ait olan sonuçların detayları Çizelge 4.14’de verilmiştir.

Çizelge 4.14. Maltaca için doküman bazında başarı oranları

Dil	Maltaca			
Format	DOC Dosya Tipi		PDF Dosya Tipi	
Yöntem	#1	#2	#1	#2
Almanca	0%	0%	0%	0%
Arnavutça	0%	0%	0%	0%
Fransızca	0%	0%	0%	0%
Galce	0%	0%	0%	0%
Hırvatça	0%	0%	0%	0%
İngilizce	0%	0%	0%	0%
İrlandaca	28%	28%	28%	28%
İspanyolca	49%	49%	50%	50%
İtalyanca	66%	66%	66%	66%
Letonca	79%	79%	80%	80%
Macarca	90%	90%	90%	90%
Maltaca	99%	99%	100%	100%
Portekizce	92%	92%	92%	92%
Türkçe	85%	85%	85%	85%
Vietnamca	80%	80%	80%	80%

Portekizce

Portekizce olan dokümanlarla ilgili olarak yapılan testlerde elde edilen sonuçlar Çizelge 4.15’de verilmiştir. Elde edilen sonuçlar her iki doküman tipi ve her iki tespit yöntemi kullanılarak yapılmıştır. Elde edilen sonuçlardan Microsoft Word DOC ve Adobe PDF doküman tipleri için 1. Yöntem (#1) ile %95 ve 2. Yöntem (#2) ile %99 oranında bir başarı elde edilmiştir. Aynı tablodan görülebileceği gibi sonuçların Portekizce olması beklenirken elde edilen sonuçlarda Portekizce’dir. Bu sebeple Portekizce için her iki dosya formatı ve her iki yönteminde %98 başarılı olduğu görülmüştür. Portekizce dokümanlar ile diğer dillere ait olan sonuçların detayları Çizelge 4.15’de verilmiştir.

Çizelge 4.15. Portekizce için doküman bazında başarı oranları

Dil	Portekizce			
Format	DOC Dosya Tipi		PDF Dosya Tipi	
Yöntem	#1	#2	#1	#2
Almanca	0%	0%	0%	0%
Arnavutça	0%	0%	0%	0%
Fransızca	0%	0%	0%	0%
Galce	0%	0%	0%	0%
Hırvatça	0%	0%	0%	0%
İngilizce	0%	0%	0%	0%
İrlandaca	6%	16%	5%	12%
İspanyolca	30%	39%	29%	35%
İtalyanca	49%	56%	48%	53%
Letonca	64%	71%	63%	68%
Macarca	76%	82%	75%	80%
Maltaca	86%	92%	86%	90%
Portekizce	95%	99%	95%	98%
Türkçe	96%	91%	97%	93%
Vietnamca	90%	85%	90%	87%

Türkçe

Türkçe olan dokümanlarla ilgili olarak yapılan testlerde elde edilen sonuçlar Çizelge 4.16’da verilmiştir. Elde edilen sonuçlar her iki doküman tipi ve her iki tespit yöntemi kullanılarak yapılmıştır. Elde edilen sonuçlardan Microsoft Word DOC ve Adobe PDF doküman tipleri için 1. Yöntem (#1) ile %99 ve 2. Yöntem (#2) ile %99 oranında bir başarı elde edilmiştir. Aynı tablodan görülebileceği gibi sonuçların Türkçe olması beklenirken elde edilen sonuçlarda Türkçe’dir. Bu sebeple Türkçe için her iki dosya formatı ve her iki yönteminde %100 başarılı olduğu görülmüştür. Türkçe dokümanlar ile diğer dillere ait olan sonuçların detayları Çizelge 4.16’da verilmiştir.

Çizelge 4.16. Türkçe için doküman bazında başarı oranları

Dil	Türkçe			
Format	DOC Dosya Tipi		PDF Dosya Tipi	
Yöntem	#1	#2	#1	#2
Almanca	0%	0%	0%	0%
Arnavutça	0%	0%	0%	0%
Fransızca	0%	0%	0%	0%
Galce	0%	0%	0%	0%
Hırvatça	0%	0%	0%	0%
İngilizce	0%	0%	0%	0%
İrlandaca	0%	0%	0%	0%
İspanyolca	24%	25%	25%	23%
İtalyanca	44%	44%	44%	43%
Letonca	59%	60%	60%	59%
Macarca	72%	73%	72%	71%
Maltaca	83%	83%	83%	82%
Portekizce	92%	92%	92%	91%
Türkçe	99%	99%	99%	99%
Vietnamca	93%	93%	93%	93%

Vietnamca

Vietnamca olan dokümanlarla ilgili olarak yapılan testlerde elde edilen sonuçlar Çizelge 4.17’de verilmiştir. Elde edilen sonuçlar her iki doküman tipi ve her iki tespit yöntemi kullanılarak yapılmıştır. Elde edilen sonuçlardan Microsoft Word DOC ve Adobe PDF doküman tipleri için 1. Yöntem (#1) ile %99 ve 2. Yöntem (#2) ile %99 oranında bir başarı elde edilmiştir. Aynı tablodan görülebileceği gibi sonuçların Vietnamca olması beklenirken elde edilen sonuçlarda Vietnamca’dır. Bu sebeple Vietnamca için her iki dosya formatı ve her iki yönteminde %100 başarılı olduğu görülmüştür. Vietnamca dokümanlar ile diğer dillere ait olan sonuçların detayları Çizelge 4.17’de verilmiştir.

Çizelge 4.17. Vietnamca için doküman bazında başarı oranları

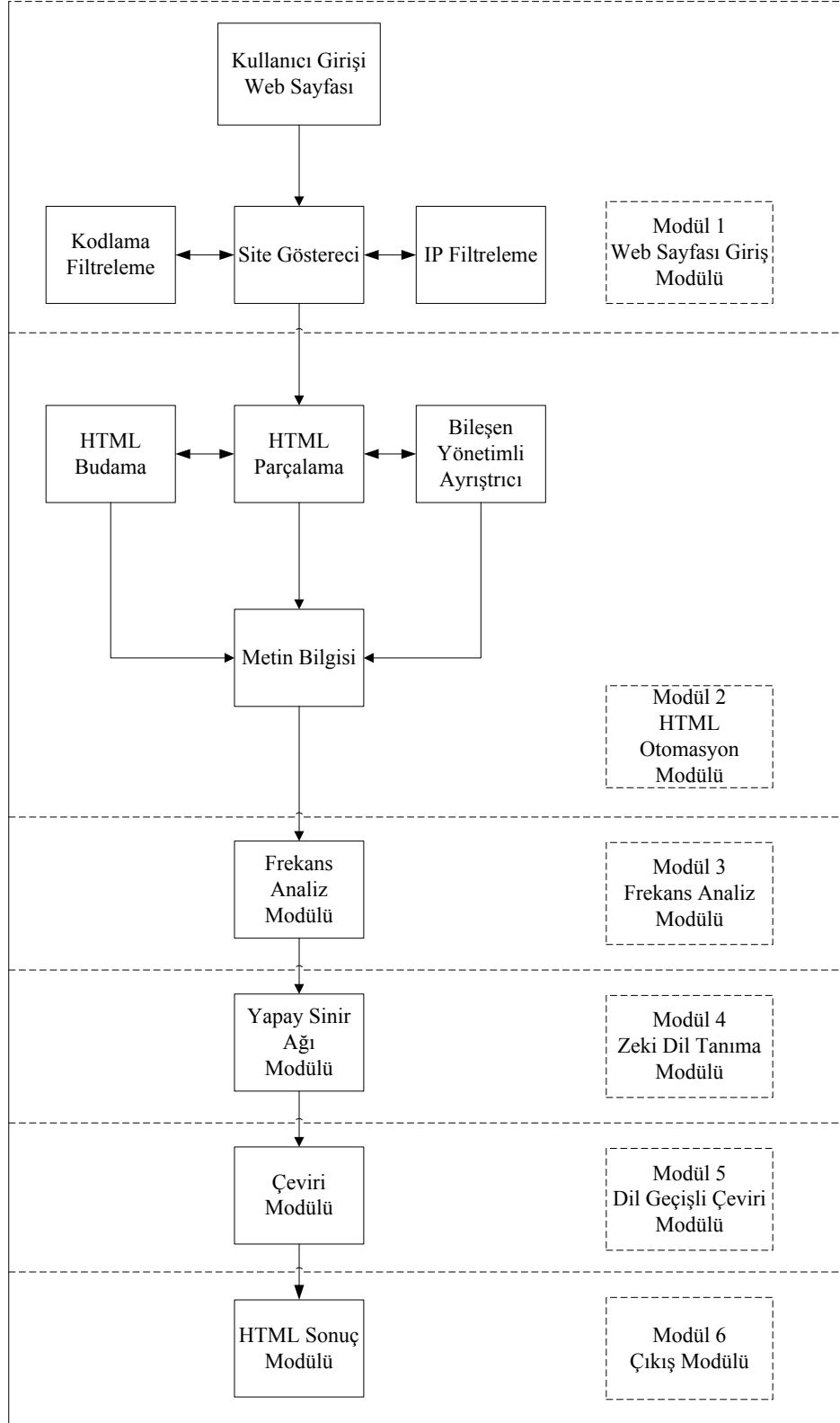
Dil	Vietnamca			
Format	DOC Dosya Tipi		PDF Dosya Tipi	
Yöntem	#1	#2	#1	#2
Almanca	0%	0%	0%	0%
Arnavutça	0%	0%	0%	0%
Fransızca	0%	0%	0%	0%
Galce	0%	0%	0%	0%
Hırvatça	0%	0%	0%	0%
İngilizce	0%	0%	0%	0%
İrlandaca	0%	0%	0%	0%
İspanyolca	13%	10%	16%	11%
İtalyanca	34%	31%	36%	32%
Letonca	50%	48%	52%	49%
Macarca	64%	61%	66%	62%
Maltaca	75%	73%	77%	74%
Portekizce	85%	83%	86%	83%
Türkçe	93%	91%	94%	92%
Vietnamca	99%	98%	98%	99%

4.4. Test 2

Bu grup testler web sitelerine ait içerik dilinin zeki yazılım tarafından algılanma başarısını ölçmeyi hedeflemektedir. Bu testler için bahsi geçen dillere ait içeriğe sahip üniversite ve medya siteleri temel alınarak yapılmıştır. Bu sitelere ait içerikler günün herhangi bir saatinde değişebilir nitelikte olduğu için bu bölümde başarı oranları her dil için o dile ait en çok bilinen web sitesi üzerinden sağlanan başarı oranıyla ölçülmüştür.

Geliştirilen yazılım sayısal dokümanlar için Şekil 4.5’de verilen akış diyagramını çalışma zamanında Şekil 4.10’da PDF ve DOC dosyaları için Şekil 4.11’de HTML tabanlı dosyalar için özelleştirmektedir. Şekil 4.11’de verilen Web Sayfası giriş modülü ile kullanıcılar içerik dillerini bilmedikleri web sayfalarını sisteme giriş

olarak uygulayabilmektedirler. Geliştirilen yazılım giriş olarak uygulanan web sayfaları için ilk etapta internet protokolü analizi yaparak yüklendikleri ülkeyi, sonrasında ise kod analizi yaparak karakter kodlama kümesine erişmektedir. Bu bilgiler kullanıcılara bildirilerek dil tanıma işleminin iyileştirilmesi ve farklı açılardan desteklenmesi sağlanmaktadır. Bir web sitesinde, içerik bilgisinin sayfanın her metin dizisinde aynı olmaması ve HTML kodlama dili İngilizce olduğu için sayfa üzerinden erişilen ve HTML filtreleme ile temizlenen içerik bilgisinin sağlıklı yapılamaması karşılaşılan en büyük sorunlardandır. Bu nedenle, bu aşamada HTML kodu budama ve bileşen tabanlı HTML'den metin ayrıştırma ile iki farklı yaklaşım kullanılmıştır. Bu yaklaşımlar sayesinde farklı şekillerde çıkartılan metin bilgisi ile dil tanıma gerçekleştirildikten sonra tespit edilen dilin, istenilen dile dönüşümünü yapılarak yazılım içine gömülmüş olan internet göstericisinde kullanıcı tarafından giriş olarak uygulanan web sitesinin hiç bir içerik bozulmasına uğramamış çevirisi kullanıcılara sunulmaktadır. Kullanıcılar çevirilerin yazıcı çıktılarına erişebilmektedir.



Şekil 4.11. Web sayfaları için sistem modeli blok diyagramı

Bu testlere konu olan Almanca, Arnavutça, Fransızca, Galce, Hırvatça, İngilizce, İrlandaca, İspanyolca, İtalyanca, Letonca, Macarca, Maltaca, Portekizce, Türkçe, Vietnamca dillerine ait detaylı sonuçlar aşağıda başlıklar halinde verilmiştir.

Almanca

Almanca olan web siteleriyle ilgili olarak yapılan örnek test sonucu Çizelge 4.18’de verilmiştir. Elde edilen sonuç 1. Yöntem (#1) ile %93 ve 2. Yöntem (#2) ile %92 oranında bir başarı elde edildiğini göstermekle birlikte aynı tablodan görülebileceği gibi sonuçların Almanca olması beklenirken elde edilen sonuçlarda Almanca’dır. Bu sebeple Almanca için web sitelerinde dil tanıma başarısının %100 olduğu görülmüştür. Almanca örnek web sitesi ile diğer dillere ait olan sonuçların detayları Çizelge 4.18’de verilmiştir.

Çizelge 4.18. Almanca için web sitesi bazında başarı oranları

Dil	Almanca	
Web Sitesi	www.bild.de	
Yöntem	#1	#2
Almanca	93%	92%
Arnavutça	53%	46%
Fransızca	35%	30%
Galce	26%	23%
Hırvatça	21%	18%
İngilizce	17%	15%
İrlandaca	15%	13%
İspanyolca	13%	11%
İtalyanca	11%	10%
Letonca	10%	9%
Macarca	9%	8%
Maltaca	8%	7%
Portekizce	8%	7%
Türkçe	7%	6%
Vietnamca	7%	6%

Arnavutça

Arnavutça olan web siteleriyle ilgili olarak yapılan örnek test sonucu Çizelge 4.19’da verilmiştir. Elde edilen sonuç 1. Yöntem (#1) ile %99 ve 2. Yöntem (#2) ile %99 oranında bir başarı elde edildiğini göstermekle birlikte aynı tablodan görülebileceği gibi sonuçların Arnavutça olması beklenirken elde edilen sonuçlarda Arnavutça’dır. Bu sebeple Arnavutça için web sitelerinde dil tanıma başarısının %100 olduğu görülmüştür. Arnavutça örnek web sitesi ile diğer dillere ait olan sonuçların detayları Çizelge 4.19’da verilmiştir.

Çizelge 4.19. Arnavutça için web sitesi bazında başarı oranları

Dil	Arnavutça	
Web Sitesi	www.mash.gov.al	
Yöntem	#1	#2
Almanca	0%	0%
Arnavutça	99%	99%
Fransızca	66%	67%
Galce	49%	50%
Hırvatça	39%	40%
İngilizce	33%	33%
İrlandaca	28%	28%
İspanyolca	24%	25%
İtalyanca	22%	22%
Letonca	19%	20%
Macarca	18%	18%
Maltaca	16%	16%
Portekizce	15%	15%
Türkçe	14%	14%
Vietnamca	13%	13%

Fransızca

Fransızca olan web siteleriyle ilgili olarak yapılan örnek test sonucu Çizelge 4.20’de verilmiştir. Elde edilen sonuç 1. Yöntem (#1) ile %99 ve 2. Yöntem (#2) ile %75 oranında bir başarı elde edildiğini göstermekle birlikte aynı tablodan görülebileceği

gibi sonuçların Fransızca olması beklenirken elde edilen sonuçlarda Fransızca'dır. Bu sebeple önerilen birinci yöntemin başarısı %100 olmasına rağmen Fransızca için web sitelerinde dil tanıma başarısının ortalama %94 olduğu görülmüştür. Fransızca örnek web sitesi ile diğer dillere ait olan sonuçların detayları Çizelge 4.20'de verilmiştir.

Çizelge 4.20. Fransızca için web sitesi bazında başarı oranları

Dil	Fransızca	
Web Sitesi	www.univ-tlse1.fr	
Yöntem	#1	#2
Almanca	0%	0%
Arnavutça	62%	13%
Fransızca	91%	75%
Galce	68%	93%
Hırvatça	55%	74%
İngilizce	45%	62%
İrlandaca	39%	53%
İspanyolca	34%	46%
İtalyanca	30%	41%
Letonca	27%	37%
Macarca	25%	33%
Maltaca	22%	31%
Portekizce	21%	28%
Türkçe	19%	26%
Vietnamca	18%	24%

Galce

Galce olan web siteleriyle ilgili olarak yapılan örnek test sonucu Çizelge 4.21'de verilmiştir. Elde edilen sonuç 1. Yöntem (#1) ile %95 ve 2. Yöntem (#2) ile %93 oranında bir başarı elde edildiğini göstermekle birlikte aynı tablodan görülebileceği gibi sonuçların Galce olması beklenirken elde edilen sonuçlarda Galce'dir. Bu sebeple Galce için web sitelerinde dil tanıma başarısının %100 olduğu görülmüştür.

Galce örnek web sitesi ile diğer dillere ait olan sonuçların detayları Çizelge 4.21’de verilmiştir.

Çizelge 4.21. Galce için web sitesi bazında başarı oranları

Dil	Galce	
Web Sitesi	www.caerdydd.ac.uk	
Yöntem	#1	#2
Almanca	0%	0%
Arnavutça	9%	0%
Fransızca	72%	57%
Galce	95%	93%
Hırvatça	76%	85%
İngilizce	63%	71%
İrlandaca	54%	61%
İspanyolca	47%	53%
İtalyanca	42%	47%
Letonca	38%	42%
Macarca	34%	38%
Maltaca	31%	35%
Portekizce	29%	32%
Türkçe	27%	30%
Vietnamca	25%	28%

Hırvatça

Hırvatça olan web siteleriyle ilgili olarak yapılan örnek test sonucu Çizelge 4.22’de verilmiştir. Elde edilen sonuç 1. Yöntem (#1) ile %55 ve 2. Yöntem (#2) ile %95 oranında bir başarı elde edildiğini göstermekle birlikte aynı tablodan görülebileceği gibi sonuçların Hırvatça olması beklenirken elde edilen sonuçlarda Hırvatça’dır. Bu sebeple önerilen ikinci yöntemin başarısı %100 olmasına rağmen Hırvatça için web sitelerinde dil tanıma başarısının ortalama %77 olduğu görülmüştür. Hırvatça örnek web sitesi ile diğer dillere ait olan sonuçların detayları Çizelge 4.22’de verilmiştir.

Çizelge 4.22. Hırvatça için web sitesi bazında başarı oranları

Dil	Hırvatça	
Web Sitesi	www.unizg.hr	
Yöntem	#1	#2
Almanca	0%	0%
Arnavutça	0%	0%
Fransızca	0%	25%
Galce	19%	69%
Hırvatça	55%	95%
İngilizce	79%	87%
İrlandaca	96%	74%
İspanyolca	90%	65%
İtalyanca	80%	58%
Letonca	72%	52%
Macarca	65%	47%
Maltaca	60%	43%
Portekizce	55%	40%
Türkçe	51%	37%
Vietnamca	48%	34%

İngilizce

İngilizce olan web siteleriyle ilgili olarak yapılan örnek test sonucu Çizelge 4.23’de verilmiştir. Elde edilen sonuç 1. Yöntem (#1) ile %99 ve 2. Yöntem (#2) ile %86 oranında bir başarı elde edildiğini göstermekle birlikte aynı tablodan görülebileceği gibi sonuçların İngilizce olması beklenirken elde edilen sonuçlarda İngilizce’dir. Bu sebeple önerilen birinci yöntemin başarısı %100 olmasına rağmen İngilizce için web sitelerinde dil tanıma başarısının ortalama %94 olduğu görülmüştür. İngilizce örnek web sitesi ile diğer dillere ait olan sonuçların detayları Çizelge 4.23’de verilmiştir.

Çizelge 4.23. İngilizce için web sitesi bazında başarı oranları

Dil	İngilizce	
Web Sitesi	www.bbc.co.uk	
Yöntem	#1	#2
Almanca	0%	0%
Arnavutça	0%	0%
Fransızca	1%	27%
Galce	51%	70%
Hırvatça	80%	96%
İngilizce	99%	86%
İrlandaca	85%	74%
İspanyolca	74%	64%
İtalyanca	66%	57%
Letonca	59%	51%
Macarca	54%	47%
Maltaca	49%	43%
Portekizce	45%	39%
Türkçe	42%	37%
Vietnamca	39%	34%

İrlandaca

İrlandaca olan web siteleriyle ilgili olarak yapılan örnek test sonucu Çizelge 4.24’de verilmiştir. Elde edilen sonuç 1. Yöntem (#1) ile %99 ve 2. Yöntem (#2) ile %96 oranında bir başarı elde edildiğini göstermekle birlikte aynı tablodan görülebileceği gibi sonuçların İrlandaca olması beklenirken elde edilen sonuçlarda İrlandaca’dır. Bu sebeple İrlandaca için web sitelerinde dil tanıma başarısının %100 olduğu görülmüştür. İrlandaca örnek web sitesi ile diğer dillere ait olan sonuçların detayları Çizelge 4.24’de verilmiştir.

Çizelge 4.24. İrlandaca için web sitesi bazında başarı oranları

Dil	İrlandaca	
Web Sitesi	www.foinse.ie	
Yöntem	#1	#2
Almanca	0%	0%
Arnavutça	0%	0%
Fransızca	0%	0%
Galce	24%	30%
Hırvatça	59%	64%
İngilizce	83%	87%
İrlandaca	99%	96%
İspanyolca	87%	84%
İtalyanca	77%	75%
Letonca	70%	67%
Macarca	63%	61%
Maltaca	58%	56%
Portekizce	53%	52%
Türkçe	50%	48%
Vietnamca	46%	45%

İspanyolca

İspanyolca olan web siteleriyle ilgili olarak yapılan örnek test sonucu Çizelge 4.25’de verilmiştir. Elde edilen sonuç 1. Yöntem (#1) ile %97 ve 2. Yöntem (#2) ile %98 oranında bir başarı elde edildiğini göstermekle birlikte aynı tablodan görülebileceği gibi sonuçların İspanyolca olması beklenirken elde edilen sonuçlarda İspanyolca’dır. Bu sebeple İspanyolca için web sitelerinde dil tanıma başarısının %100 olduğu görülmüştür. İspanyolca örnek web sitesi ile diğer dillere ait olan sonuçların detayları Çizelge 4.25’de verilmiştir.

Çizelge 4.25. İspanyolca için web sitesi bazında başarı oranları

Dil	İspanyolca	
Web Sitesi	www.uniovi.es	
Yöntem	#1	#2
Almanca	0%	0%
Arnavutça	0%	0%
Fransızca	0%	0%
Galce	4%	3%
Hırvatça	43%	42%
İngilizce	69%	68%
İrlandaca	88%	87%
İspanyolca	97%	98%
İtalyanca	87%	87%
Letonca	78%	78%
Macarca	71%	71%
Maltaca	65%	65%
Portekizce	60%	60%
Türkçe	55%	56%
Vietnamca	52%	52%

İtalyanca

İtalyanca olan web siteleriyle ilgili olarak yapılan örnek test sonucu Çizelge 4.26’da verilmiştir. Elde edilen sonuç 1. Yöntem (#1) ile %98 ve 2. Yöntem (#2) ile %97 oranında bir başarı elde edildiğini göstermekle birlikte aynı tablodan görülebileceği gibi sonuçların İtalyanca olması beklenirken elde edilen sonuçlarda İtalyanca’dır. Bu sebeple İtalyanca için web sitelerinde dil tanıma başarısının %100 olduğu görülmüştür. İtalyanca örnek web sitesi ile diğer dillere ait olan sonuçların detayları Çizelge 4.26’da verilmiştir.

Çizelge 4.26. İtalyanca için web sitesi bazında başarı oranları

Dil	İtalyanca	
Web Sitesi	www.unito.it	
Yöntem	#1	#2
Almanca	0%	0%
Arnavutça	0%	0%
Fransızca	0%	0%
Galce	0%	0%
Hırvatça	22%	15%
İngilizce	52%	46%
İrlandaca	73%	68%
İspanyolca	89%	84%
İtalyanca	98%	97%
Letonca	88%	92%
Macarca	80%	83%
Maltaca	73%	76%
Portekizce	68%	70%
Türkçe	63%	65%
Vietnamca	59%	61%

Letonca

Letonca olan web siteleriyle ilgili olarak yapılan örnek test sonucu Çizelge 4.27’de verilmiştir. Elde edilen sonuç 1. Yöntem (#1) ile %95 ve 2. Yöntem (#2) ile %99 oranında bir başarı elde edildiğini göstermekle birlikte aynı tablodan görülebileceği gibi sonuçların Letonca olması beklenirken elde edilen sonuçlarda Letonca’dır. Bu sebeple Letonca için web sitelerinde dil tanıma başarısının %100 olduğu görülmüştür. Letonca örnek web sitesi ile diğer dillere ait olan sonuçların detayları Çizelge 4.27’de verilmiştir.

Çizelge 4.27. Letonca için web sitesi bazında başarı oranları

Dil	Letonca	
Web Sitesi	www.llu.lv	
Yöntem	#1	#2
Almanca	0%	0%
Arnavutça	0%	0%
Fransızca	0%	0%
Galce	0%	0%
Hırvatça	0%	0%
İngilizce	25%	33%
İrlandaca	50%	57%
İspanyolca	69%	74%
İtalyanca	83%	88%
Letonca	95%	99%
Macarca	94%	90%
Maltaca	87%	83%
Portekizce	80%	76%
Türkçe	74%	71%
Vietnamca	69%	66%

Macarca

Macarca olan web siteleriyle ilgili olarak yapılan örnek test sonucu Çizelge 4.28’de verilmiştir. Elde edilen sonuç 1. Yöntem (#1) ile %95 ve 2. Yöntem (#2) ile %97 oranında bir başarı elde edildiğini göstermekle birlikte aynı tablodan görülebileceği gibi sonuçların Macarca olması beklenirken elde edilen sonuçlarda Macarca’dır. Bu sebeple Macarca için web sitelerinde dil tanıma başarısının %100 olduğu görülmüştür. Macarca örnek web sitesi ile diğer dillere ait olan sonuçların detayları Çizelge 4.28’de verilmiştir.

Çizelge 4.28. Macarca için web sitesi bazında başarı oranları

Dil	Macarca	
Web Sitesi	www.ises.hu	
Yöntem	#1	#2
Almanca	0%	0%
Arnavutça	0%	0%
Fransızca	0%	0%
Galce	0%	0%
Hırvatça	0%	0%
İngilizce	25%	21%
İrlandaca	50%	46%
İspanyolca	69%	66%
İtalyanca	83%	80%
Letonca	95%	92%
Macarca	95%	97%
Maltaca	87%	89%
Portekizce	80%	82%
Türkçe	74%	76%
Vietnamca	69%	71%

Maltaca

Maltaca olan web siteleriyle ilgili olarak yapılan örnek test sonucu Çizelge 4.29’da verilmiştir. Elde edilen sonuç 1. Yöntem (#1) ile %99 ve 2. Yöntem (#2) ile %99 oranında bir başarı elde edildiğini göstermekle birlikte aynı tablodan görülebileceği gibi sonuçların Maltaca olması beklenirken elde edilen sonuçlarda Maltaca’dır. Bu sebeple Maltaca için web sitelerinde dil tanıma başarısının %100 olduğu görülmüştür. Maltaca örnek web sitesi ile diğer dillere ait olan sonuçların detayları Çizelge 4.29’da verilmiştir.

Çizelge 4.29. Maltaca için web sitesi bazında başarı oranları

Dil	Maltaca	
Web Sitesi	www.illum.com.mt	
Yöntem	#1	#2
Almanca	0%	0%
Arnavutça	0%	0%
Fransızca	0%	0%
Galce	0%	0%
Hırvatça	0%	0%
İngilizce	0%	0%
İrlandaca	29%	28%
İspanyolca	50%	49%
İtalyanca	67%	66%
Letonca	80%	79%
Macarca	91%	90%
Maltaca	99%	99%
Portekizce	91%	92%
Türkçe	85%	85%
Vietnamca	79%	80%

Portekizce

Portekizce olan web siteleriyle ilgili olarak yapılan örnek test sonucu Çizelge 4.30’da verilmiştir. Elde edilen sonuç 1. Yöntem (#1) ile %99 ve 2. Yöntem (#2) ile %96 oranında bir başarı elde edildiğini göstermekle birlikte aynı tablodan görülebileceği gibi sonuçların Portekizce olması beklenirken elde edilen sonuçlarda Portekizce’dir. Bu sebeple Portekizce için web sitelerinde dil tanıma başarısının %100 olduğu görülmüştür. Portekizce örnek web sitesi ile diğer dillere ait olan sonuçların detayları Çizelge 4.30’da verilmiştir.

Çizelge 4.30. Portekizce için web sitesi bazında başarı oranları

Dil	Portekizce	
Web Sitesi	www.uminho.pt	
Yöntem	#1	#2
Almanca	0%	0%
Arnavutça	0%	0%
Fransızca	0%	0%
Galce	0%	0%
Hırvatça	0%	0%
İngilizce	0%	0%
İrlandaca	15%	21%
İspanyolca	38%	43%
İtalyanca	56%	61%
Letonca	71%	75%
Macarca	82%	86%
Maltaca	92%	95%
Portekizce	99%	96%
Türkçe	92%	89%
Vietnamca	85%	83%

Türkçe

Türkçe olan web siteleriyle ilgili olarak yapılan örnek test sonucu Çizelge 4.31’de verilmiştir. Elde edilen sonuç 1. Yöntem (#1) ile %98 ve 2. Yöntem (#2) ile %99 oranında bir başarı elde edildiğini göstermekle birlikte aynı tablodan görülebileceği gibi sonuçların Türkçe olması beklenirken elde edilen sonuçlarda Türkçe’dir. Bu sebeple Türkçe için web sitelerinde dil tanıma başarısının %100 olduğu görülmüştür. Türkçe örnek web sitesi ile diğer dillere ait olan sonuçların detayları Çizelge 4.31’de verilmiştir.

Çizelge 4.31. Türkçe için web sitesi bazında başarı oranları

Dil	Türkçe	
Web Sitesi	www.hurriyet.com.tr	
Yöntem	#1	#2
Almanca	0%	0%
Arnavutça	0%	0%
Fransızca	0%	0%
Galce	0%	0%
Hırvatça	0%	0%
İngilizce	0%	0%
İrlandaca	3%	0%
İspanyolca	28%	25%
İtalyanca	47%	44%
Letonca	62%	60%
Macarca	75%	73%
Maltaca	85%	83%
Portekizce	94%	92%
Türkçe	98%	99%
Vietnamca	91%	93%

Vietnamca

Vietnamca olan web siteleriyle ilgili olarak yapılan örnek test sonucu Çizelge 4.32’de verilmiştir. Elde edilen sonuç 1. Yöntem (#1) ile %98 ve 2. Yöntem (#2) ile %99 oranında bir başarı elde edildiğini göstermekle birlikte aynı tablodan görülebileceği gibi sonuçların Vietnamca olması beklenirken elde edilen sonuçlarda Vietnamca’dır. Bu sebeple Vietnamca için web sitelerinde dil tanıma başarısının %100 olduğu görülmüştür. Vietnamca örnek web sitesi ile diğer dillere ait olan sonuçların detayları Çizelge 4.32’de verilmiştir.

Çizelge 4.32. Vietnamca için web sitesi bazında başarı oranları

Dil	Vietnamca	
Web Sitesi	www.hanu.vn	
Yöntem	#1	#2
Almanca	0%	0%
Arnavutça	0%	0%
Fransızca	0%	0%
Galce	0%	0%
Hırvatça	0%	0%
İngilizce	0%	0%
İrlandaca	0%	0%
İspanyolca	15%	13%
İtalyanca	36%	33%
Letonca	52%	50%
Macarca	65%	64%
Maltaca	77%	75%
Portekizce	86%	84%
Türkçe	94%	93%
Vietnamca	98%	99%

5. SONUÇ VE ÖNERİLER

Bu tez çalışmasında web tabanlı otomatik dil tanıma ve çeviri sistemi geliştirilmiştir. Geliştirilen sistem iki farklı yöntemden oluşmaktadır. İlk yöntemde geliştirilen sistem başarılı olsada bazı dillerde karşılaşılan başarısızlık sebebiyle gözardı edilmiştir. YSA'nın öğrenme sürecinde başarı düşüklüğü, dil tanıma başarısının yetersizliği sadece web siteleri ile çalışması, çeviri yeteneklerinin sınırlı oluşu, sunucu bilgisayar ve bant genişliği bağımlı olması ve tanınması istenilen dil sayısı arttıkça bu artışa paralel olarak artan eğitim veri kümesi ve YSA giriş sayısındaki büyük artış sistem karmaşıklığını yükseltip eğitim performansını düşürdüğü için farklı bir yöntem arayışına girilmiştir. Karşılaşılan en büyük güçlük web sayfalarının aynı standart ve formatta bulunmamlarından kaynaklanan içeriğin doğru analiz edilememesi sorunudur.

Bu sebeple tanınması istenilen dil sayısı ile sistem karmaşıklığı orantısını kabul edilebilir düzeyde tutan hatta bazı durumlarda bu dil sayısı ve karmaşıklığı ters orantılı olarak değiştiren yöntemler ile kullanıcıların dillerini bilmedikleri sayısal dokümanları istediklere dile otomatik olarak dönüştürebilecek ikinci bir yöntem önerilmiştir. Daha güzel bir sistem tasarımı yapılarak platform tüm dünya dillerini destekleyebilecek şu anda 41 dil desteği veren bir platforma dönüştürülmüştür. Geliştirilen sistemde kullanılan iki yöntem ile farklı teknolojiler kullanılarak geliştirilen sistemin içerik tanımlama başarısı oldukça arttırılmıştır. Bu teknolojiler bir bütün olarak entegre olduklarında önerilen alfabetik kesişim ve birleşim yöntemleri ışığında yazılım tasarım ve test süreçleri gerçekleştirilmiştir. Bu yazılım sayesinde internet ortamında sıklıkla karşımıza çıkan ve literatürde geniş kabul görmüş olan Microsoft Word ve Adobe Acrobat dosya formatları için dil tanımlama işlemleri başarıyla yapılabilmektedir. Bu sistem kendi başına çalışır (standalone) olduğu için herhangi bir hizmet sağlayıcı sunucu veya bant genişliği kapasitesi gibi kısıtlar ortadan kaldırılmıştır. Bu çalışmada kullanılan özel bileşenler sayesinde web

içeriklerinin analiz edilmesinde yaşanan betik dil analizi gibi güçlüklerde ortadan kaldırılmıştır.

Bu çalışmada yapılan web sitesi ve doküman tanıma testlerinde sistemin %95-%100 arasında başarı oranlarında kararlı ve kesin sonuçlar verdiği görülmüştür. Bazı yanlış sonuç veren testlerde coğrafi olarak yakın olmayan bölgelerin dillerinin birbirlerine yakınsadıkları gösterilmiştir. Bu dillerin kökeni ve dağılımları için araştırmalara kaynak olabilecek bir çalışmadır. Sistemin henüz kendi kendine öğrenme sürecini bulunmadığından ölçeklenebilir veritabanı ve ağ yapısının sonraki çalışmalarda kullanıcılardan gelen geri beslemeler doğrultusunda güncellemesi gerekebileceği değerlendirilmektedir.

Bu tez çalışmasında önerilen sistemin en büyük avantajı dil tanıma sistemi için kullanılan yöntemin dil grubuna özgü olmayışı ve diğer diller içinde sistemin dönüştürülme sürecinin esnekliğidir. Dil tanıma veritabanı için önerilen kesişim ve birleşim yöntemlerinin tanınması istenilen dillere ait alfabelerde tekrarlanan harf sayıları hedef alınmaktadır. Bu sebeple tanınması istenilen dil sayısındaki artış, giriş olarak uygulanan karakter frekans analizinin yapıldığı ortak alfabe kümesinde Latin alfabesindeki dillerin birbirlerine yakınsadığı coğrafyalarda tekrarlanan harf sayıları artacağından minimal düzeyde artışlara, birbirlerinden uzaklaştığı yerlerde ise tekrarlanma azalacağı için ortak alfabe kümesinde azalmalara neden olacaktır. Bu sayede sistemin hızlı çalışması, küçük veritabanı gereksinimi ve ölçeklenebilirliği ile yüksek kararlılık gerektiren endüstriyel uygulamalarda kullanılabileceği düşünülmektedir.

Bu çalışmadan elde edilen bilgiler doğrultusunda, YSA tabanlı zeki bir Dil Tanıyıcı ve Dönüştürücü tasarımı yapılarak elde edilen sonuçlar aşağıda başlıklar halinde sunulmuştur.

- YSA'nın Dil Tanıyıcı tasarımında başarılı olduğu,
- YSA'ların ara katman sayısı, nöron sayısı ve seçilecek olan fonksiyon tipine bağlı olarak başarıların arttığı,

- Eğitim ve Test için gerekli olan ortak alfabe kümelerinin birbirlerine yakınsayan dillerde karmaşıklığın artmasına neden olduğu,
- YSA'ların farklı dillerde başarılarının ölçülmesinin şart olduğu,
- Tasarlanan YSA yapısının kolaylıkla gerçek-zamanlı uygulamalarda kullanılabileceği,
- Ülkemizde ve dünyada zeki dil tanıma sistemlerine yönelik yeterli çalışma bulunmadığı,
- Literatürde dil tanıma üzerine birçok çalışma bulunmasına rağmen veri kümeleri elde etmek için kullanılan yöntemlerin yetersiz olduğu görülmüştür.

Gerçekleştirilen sistemin test sonuçları, diğer çalışmalarla karşılaştırıldığında oldukça başarılı sonuçlar ürettiği görülmüştür. EK-41'de çalışmanın Microsoft Word dokümanları için ortalama başarı oranları, EK-42'de çalışmanın Adobe PDF dokümanları için ortalama başarı oranları, EK-43'de çalışmanın Web Sitesi HTML dokümanları için ortalama başarı oranları verilmiştir. Elde edilen bu başarının sonucunda LM algoritmasıyla eğitilmiş sistemin, zeki dil tanıyıcı tasarımında uygulanabilir olduğunu göstermiştir.

Gerçekleştirilen çalışma literatüre, istatistiksel dil tanıma metotlarının yanında farklı bakış açısı kazandırmaktadır. Literatürde de belirtildiği gibi zeki yaklaşımların DT tasarımında kullanılmasının doğru bir tercih olduğu, farklı yapılar seçerken MLP yapısının LM gibi güçlü algoritmalarla eğitilmesiyle YSA'ların DT tasarımında başarıyı arttırdığı tespit edilmiştir.

Bu çalışma sonucunda geliştirilen sistem ile kullanıcılardan Word ve PDF dokümanları ile web sayfalarını giriş olarak aldıktan sonra dil tespiti sonucunda kullanıcının dönüştürmek istediği dile yapılan tercüme yine yazılım üzerinde bulunan dahili web göstericisi arabiriminde sunulmaktadır. Bu nedenle kullanıcılar farklı dillerde yazılmış dokümanları istedikleri dile başka hiçbir araç kullanmadan zeki yöntemler ile dönüştürebilmektedirler.

Bu çalışmanın internetin daha verimli kullanılmasına, farklı dil ve kültürlerde yapılan çalışmaların kolaylıkla okunup öğrenilmesine ve internette karşılaşılan pek çok problemin çözümüne büyük katkılar sağlayacağı değerlendirilmektedir. Bu sistem sayesinde, internetin geniş kitlelere ve halk gruplarına yayıldığı bir dönemde küreselleşme ile ortaya çıkan ana dil yayınlarının başka milletler tarafından takip edilmesinin önünü açacağı değerlendirilmektedir.

Bu çalışma kapsamında;

1. Bu çalışmada her dil için anahtar kelimeleri barındıran 20 doküman kullanılmıştır. Platform başarımını yükseltmek için farklı konularda ve içeriklerde sayısal dokümanlar eğitime eklenmesi,
2. Hatalı sonuçların eğitime dahil edilerek sistemin iyileştirilmesi,
3. DOC, PDF ve HTML Ayırıştırma ile metin çıkartma mekanizmaları iyileştirilmesi,
4. Resim formatında olan dokümanlar için OCR (Optik Karakter Tanıma) parçasının geliştirilen platforma entegre edilmesi,
5. Önerilen tespit yöntemlerinin doküman tipine ve dil grubuna olan uygunluklarının incelenmesi,
6. Tasarlanan XML (Extended Markup Language) ortak arabirim dil süzgeçlerinin her dil için yapılandırılması,

Sunulan sistemin başarısını arttıracakları önerilmektedir.

Bu çalışmanın;

- Literatüre, istatistiksel dil tanıma metotlarının yanında kolayca ölçeklenebilir farklı bir bakış açısı kazandırdığı,
- Geliştirilen sistem ile kullanıcılardan Word ve PDF dokümanları ile web sayfalarını giriş olarak aldıktan sonra dil tespiti sonucunda kullanıcının dönüştürmek istediği dile yapılan tercümenin yine yazılım üzerinde bulunan

dahili web göstericisi arabiriminde sunularak kullanıcı dostu doküman tanımlama ve çeviri sisteminin oluşturulduğu,

- İnternetin daha verimli kullanılmasına, farklı dil ve kültürlerde yapılan çalışmaların kolaylıkla okunup öğrenilmesine ve internette karşılaşılan pek çok problemin çözümüne büyük katkılar sağlayacağı değerlendirilmektedir.

KAYNAKLAR

1. Padro M., Padro L., “Comparing Methods for Language Identification” *Procesamiento del Lenguaje Natural*, Barcelona, 33-35 (2004).
2. Botha G.R., Zimu V.Z., Barnard E., “Text-based language identification for the South African languages”, *SAIEE Africa Research Journal*, Cape Town, 141-146 (2007).
3. El-Shishiny H., Troussov A., McCloskey DJ., Takeuchi M., Nevidomsky A., Volkov P., “Word Fragments Based Arabic Language Identification”, *NEMLAR Conference on Arabic Language Resources and Tools*, Mısır, 23-26 (2004).
4. Kruengkrai C., Srichaivattana P., Sornlertlamvanich V., Isahara H., "Language Identification Based on String Kernels" *Communications and Information Technology*, Pekin, 896-899 (2005).
5. Zavorsky P., Wada S., Mikami Y., "Language and Encoding Scheme Identification of Extremely Large Sets of Multilingual Text Documents", *The 10th Machine Translation Summit*, Puket, 354-355 (2005).
6. Peng F., Schuurmans D., Wang S., "Language and Task Independent Text Categorization with Simple Language Models", *North American Chapter of the Association for Computational Linguistics - Human Language Technologies*, Edmonton, 110-117 (2003).
7. Nair A.S., Nair V. V., Chandra V. S. S., "Hidden Markov Model Based Identification of Transliterated Regional Language Words in Text Documents", *Twentieth International Joint Conference on Artificial Intelligence*, Haydarabad, 87-91 (2007).
8. Ahmed B., Cha S-H., Tappert C., "Language Identification from Text Using N-gram Based Cumulative Frequency Addition", *Student/Faculty Research Day*, New York, 121-128 (2004).
9. Constable P.G., "Toward a Model for Language Identification", *Summer Institute of Linguistics International Working Papers*, Dublin (2002).

10. Adams G., Resnik P., "A Language Identification Application Built on the Java Client/Server Platform", *The European Chapter of the Association of Computational Linguistics Workshop*, İspanya (1997).
11. Ölveck T., "N-Gram based Statistics Aimed. at Language Identification", *Student Research Conference in Informatics and Information Technologies*, Brastilava, 1-7 (2005).
12. Bilcu, E.B., Astola J., "A Hybrid Neural Network for Language Identification from Text", *Machine. Learning for Signal Processing Conference*, Maynooth, 253-258 (2006).
13. Liu Y-H., Chang F., Lin C-C., "Language Identification of Character Images Using Machine Learning Techniques", *International Conference on Document Analysis and Recognition*, Seul, 630-634 (2005).
14. Zhu G., Yu X., Li Y., Doermann D., "Unconstrained Language Identification Using A Shape Codebook", *The 11th International Conference on Frontiers in Handwriting Recognition*, Montreal, 13-18 (2008).
15. Baykan E., Henzinger M., Weber I., "Web Page Language Identification Based on URLs", *International Conference on Very Large Data Bases*, Auckland, 176-187 (2008).
16. Sağiroğlu, Ş., Beşdok, E., Erler, M., "Mühendislikte Yapay Zeka Uygulamaları-1:Yapay Sinir Ağları", *Ufuk Kitabevi*, Kayseri, 10-100 (2003).
17. Sağiroğlu Ş., Yavanoğlu U., Güven E.N., "Web Based Machine Learning for Language Identification and Translation" *International Conference on Machine Learning and Applications*, Ohio, 280-285 (2007).
18. İnternet : Wikipedia Web Tabanlı Ansiklopedi "Bayes Teoremi"
http://en.wikipedia.org/wiki/Bayes'_theorem/ (2008)
19. İnternet : Google Yazılım "Web Tabanlı Dil Çeviri Aracı Web Sayfası"
<http://translate.google.com/> (2008)
20. İnternet : Microsoft Yazılım "Visual Studio 2005 C# Windows Form Uygulaması Yazılım Geliştirme Aracı", <http://msdn.microsoft.com/en-us/vstudio/default.aspx> (2005).
21. İnternet : Mathworks Yazılım "Matlab R2007B Deployment Tool, Dinamik bağlantı Kütüphanesi Geliştirme Aracı",
http://www.mathworks.com/products/new_products/release2007b.html (2007).

22. İnternet : Cellbi Yazılım “Microsoft Word OLE Otomasyon Bileşeni”
<http://www.cellbi.com/products/docframework.aspx> (2008).
23. İnternet : Apache Yazılım “Adobe Reader PDF Otomasyon Bileşeni”
<http://incubator.apache.org/pdfbox/> (2008).

EKLER

EK-1 Almanca dili kesişim veri kümesi

[illegible]

[illegible]

EK-2 Arnavutça dili kesişim veri kümesi

AL	AL	AL	AL	AL	AL	AL	AL	AL	AL	AL	AL	AL	AL	AL	AL	AL	AL	AL	AL
8	6	7	8	7	7	7	8	7	7	7	8	8	8	8	8	8	8	8	7
1	3	2	2	1	2	2	2	1	2	2	2	2	2	1	1	1	2	2	1
1	2	1	1	1	1	1	1	1	1	1	1	2	1	1	1	2	1	1	1
5	6	4	4	3	5	4	4	4	3	4	4	4	4	4	4	4	4	3	3
11	8	11	10	20	12	12	11	12	12	20	11	10	11	13	12	12	11	13	19
1	0	2	2	1	1	1	1	2	1	2	1	2	1	1	2	1	2	1	1
2	0	2	2	2	2	3	2	1	2	2	2	2	2	2	1	2	2	2	2
5	6	5	4	6	5	4	5	4	4	4	6	4	6	4	4	5	5	4	4
10	18	9	10	8	10	11	10	12	13	9	10	11	10	10	10	9	10	11	10
3	2	4	4	4	4	3	4	3	3	4	5	3	4	4	3	4	4	3	3
5	4	4	4	4	4	4	4	4	4	4	4	4	5	4	4	4	4	4	3
4	2	3	3	3	3	3	3	3	3	3	3	3	3	3	5	3	3	4	4
5	8	5	5	4	4	5	5	5	5	5	5	4	4	4	4	4	5	5	4
8	5	8	7	7	7	7	7	7	8	7	8	7	8	8	7	7	7	5	7
5	5	5	4	5	5	5	4	5	5	5	4	5	5	4	6	5	5	4	5
4	4	4	4	3	4	5	4	3	5	3	3	4	4	4	5	4	4	4	3
2	0	2	2	2	1	1	1	1	1	1	1	1	2	1	1	2	2	1	1
8	12	8	10	6	9	9	9	8	9	7	8	9	9	9	10	9	8	9	8
6	10	6	6	6	6	7	7	6	7	6	7	6	7	7	6	7	6	8	6
11	11	11	10	9	11	10	11	11	11	10	10	11	10	10	9	12	10	10	10
3	3	5	4	4	4	3	4	4	3	3	4	4	4	4	3	4	3	4	3
2	2	2	3	2	3	2	2	3	2	2	2	3	2	2	2	2	2	3	2
0	0	1	0	1	0	0	1	1	0	0	1	1	1	1	1	0	1	0	0
1	0	1	1	1	1	1	1	1	1	1	1	1	1	1	0	1	1	0	1
1	0	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
1	0	1	1	1	1	1	2	1	1	1	1	1	1	1	1	1	1	2	2
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
1	2	1	1	0	1	1	1	1	1	0	1	1	1	1	1	1	1	0	1
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0

EK-4 Galce dili kesişim veri kümesi

[illegible]

EK-6 İngilizce dili kesişim veri kümesi

EN	EN	EN	EN	EN	EN	EN	EN	EN	EN	EN	EN	EN	EN	EN	EN	EN	EN	EN	EN
8	8	9	8	10	8	9	8	9	9	8	8	9	9	9	8	8	8	8	7
2	1	1	2	2	2	2	2	2	2	2	2	2	2	3	2	2	2	2	2
4	5	5	4	4	5	4	6	4	4	4	6	5	4	7	5	6	4	4	8
4	5	4	4	5	4	4	5	4	4	3	4	4	5	3	5	6	4	4	2
12	13	13	13	14	12	13	12	12	14	10	11	13	10	8	13	12	13	14	13
3	3	3	2	3	4	3	3	2	3	2	3	3	3	1	3	3	3	3	2
3	3	2	2	2	5	2	2	3	2	3	3	2	2	4	2	3	2	2	3
5	4	5	4	5	4	6	3	5	6	4	5	4	4	4	4	3	4	5	3
9	8	9	9	8	9	9	10	7	7	10	9	9	10	7	8	10	8	8	10
1	1	1	1	1	1	1	1	1	1	1	1	1	1	0	1	1	1	1	1
1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
5	4	5	5	5	4	5	4	5	5	4	5	5	7	6	5	4	5	4	5
3	3	3	3	4	3	3	3	3	3	4	3	3	5	4	4	3	3	3	3
8	8	7	8	7	9	8	10	7	7	9	9	8	7	6	9	9	7	8	8
8	8	8	7	8	8	8	7	9	8	7	8	7	8	8	9	9	9	8	9
3	3	3	3	2	2	3	3	3	3	4	3	3	3	3	3	3	3	3	2
1	1	1	1	1	1	1	1	1	1	1	1	1	1	6	1	1	1	1	0
7	7	7	7	7	7	7	6	7	7	7	6	7	7	6	6	7	8	7	9
7	9	8	6	7	7	7	8	8	7	8	6	7	7	8	8	7	7	7	9
10	10	9	10	10	9	11	9	10	10	9	9	9	8	7	9	8	10	11	9
3	3	3	4	3	3	2	3	3	3	4	3	3	3	3	3	3	3	3	2
2	1	1	2	2	2	2	2	1	1	1	2	2	2	1	2	2	2	2	1
2	2	2	2	1	2	1	1	2	2	2	3	2	2	1	2	1	2	2	1
1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
2	2	2	3	3	2	2	2	2	2	2	1	2	2	5	2	2	2	2	2
1	1	1	1	1	1	1	1	1	1	0	1	1	1	1	1	1	1	1	1
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0
0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
1	0	0	1	0	0	0	0	0	0	0	1	0	0	0	0	0	1	0	0
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0
1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
0	0	0	1	0	0	0	0	0	0	0	1	0	1	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
0	0	0	1	0	0	0	0	0	0	0	1	0	1	0	0	0	0	0	0

EK-7 İrlandaca dili kesişim veri kümesi

[illegible]

EK-8 İspanyolca dili kesişim veri kümesi

ES	ES	ES	ES	ES	ES	ES	ES	ES	ES	ES	ES	ES	ES	ES	ES	ES	ES	ES	ES
12	13	11	13	13	11	12	12	12	11	12	12	12	11	11	13	12	12	11	12
2	2	3	2	2	2	2	2	1	2	2	1	1	2	2	2	2	2	2	1
5	6	7	4	5	5	5	6	7	6	6	6	5	5	6	6	5	8	5	7
6	6	6	6	6	6	6	5	7	6	6	6	5	5	7	5	5	6	6	6
15	13	14	14	13	14	13	13	13	14	13	13	14	14	12	12	14	12	15	12
2	1	2	1	1	1	1	1	2	1	1	1	1	1	1	1	1	1	1	2
1	3	1	2	2	1	2	2	1	1	2	1	2	1	2	2	2	2	2	2
1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
6	9	9	7	7	7	8	8	9	8	9	9	8	7	8	9	7	9	7	9
1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
0	0	1	0	1	1	1	1	1	1	0	1	0	1	1	1	1	1	0	1
5	7	7	6	6	6	6	6	6	6	6	7	7	6	6	5	6	5	5	6
4	3	3	3	4	3	3	3	2	3	4	4	3	4	3	4	3	3	2	2
7	8	8	7	8	8	8	8	6	7	8	7	8	8	7	7	7	9	7	6
8	8	9	8	8	9	9	9	7	8	8	9	9	9	10	8	9	8	10	10
3	3	3	3	3	4	4	4	3	3	3	3	4	3	4	3	3	3	3	3
1	1	1	2	1	1	1	1	1	1	1	1	1	2	1	1	2	1	1	1
7	7	6	7	7	7	7	6	7	7	7	6	6	7	6	8	7	6	8	7
10	7	8	7	9	8	8	9	8	8	8	7	9	7	8	7	8	8	9	8
5	5	6	5	5	5	5	5	7	5	5	6	5	5	6	6	6	5	6	5
5	4	4	5	4	4	4	4	4	3	4	4	3	5	3	3	4	4	5	4
1	2	1	2	2	1	1	1	2	2	2	1	1	2	1	1	2	1	1	2
1	0	1	1	1	1	0	1	1	0	0	1	1	1	1	1	0	1	1	1
1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
2	1	1	1	1	1	2	1	1	1	1	1	1	2	2	1	1	1	1	1
1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	0	1
0	0	1	1	0	1	0	0	1	0	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
0	0	1	0	0	0	0	1	1	0	0	0	0	0	0	0	0	0	0	0
1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
0	0	1	0	1	0	0	1	1	0	0	0	0	1	0	0	0	0	0	0
0	0	0	0	1	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0
1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
1	2	2	1	1	1	1	2	2	2	1	1	1	1	1	2	1	2	1	1
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0
1	1	1	1	1	1	1	1	1	1	0	1	1	1	1	1	1	1	1	1
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
0	0	1	1	1	1	1	1	0	0	0	0	0	1	0	1	1	0	0	0

EK-10 Letonca dili kesişim veri kümesi

LV	LV	LV	LV	LV	LV	LV	LV	LV	LV	LV	LV	LV	LV	LV	LV	LV	LV	LV	LV
15	14	11	14	13	13	12	12	14	12	13	13	11	14	14	11	12	14	14	14
4	2	3	2	2	3	2	2	2	3	3	2	2	2	2	2	2	2	2	2
1	2	1	2	3	2	2	2	2	2	2	2	2	2	2	2	2	2	1	3
6	4	4	4	4	4	4	5	4	4	3	4	4	3	4	3	4	5	3	4
7	6	6	8	7	7	7	8	8	6	7	8	8	6	8	7	8	7	9	7
1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	2	1	1
2	2	1	2	2	2	2	2	2	3	2	3	2	2	2	2	2	2	2	2
1	1	1	1	1	1	1	1	1	1	1	1	1	1	2	1	1	1	1	1
10	9	10	11	11	12	9	11	12	12	11	9	11	13	10	13	12	10	9	9
3	4	2	3	4	4	3	2	2	3	3	3	3	3	3	3	3	4	3	4
5	4	5	6	4	5	5	5	6	4	4	6	5	4	4	6	6	5	8	4
3	5	6	5	4	4	4	4	4	5	4	4	4	5	4	4	3	4	3	5
4	5	6	4	3	5	5	4	5	5	4	4	4	3	4	5	4	4	4	5
5	5	5	5	6	5	6	6	5	6	6	7	7	6	7	5	6	5	4	5
4	4	4	4	5	4	5	4	4	5	5	5	5	4	5	4	4	5	5	4
3	3	4	4	3	4	4	4	4	4	4	4	5	5	5	3	3	4	3	4
0	0	0	0	1	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0
8	6	6	7	7	6	7	7	7	6	6	8	7	7	7	7	6	7	9	7
9	10	12	9	9	9	8	10	10	9	11	10	8	12	9	10	10	10	12	8
8	6	8	7	9	8	7	8	8	7	9	6	8	6	8	8	9	9	8	5
6	6	8	6	6	7	5	5	5	6	6	6	7	4	5	7	7	7	5	5
4	3	4	5	2	4	3	3	5	4	4	3	4	4	4	4	5	3	2	2
1	0	0	0	1	1	1	0	0	1	0	1	1	1	0	0	0	1	0	0
0	0	0	0	0	0	1	0	0	1	0	1	0	0	0	0	0	1	1	0
0	0	0	1	0	1	1	0	0	1	0	0	1	0	0	1	0	0	0	0
2	2	4	2	2	3	2	4	3	2	3	2	2	2	3	2	2	2	2	2
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
0	3	0	0	0	0	5	0	0	0	0	0	0	0	0	0	0	0	0	4
0	3	0	0	1	0	2	0	0	0	0	0	0	0	0	1	0	0	0	2
0	0	0	0	2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
0	1	0	0	1	0	1	0	0	0	0	0	0	0	0	0	0	0	0	1
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
0	1	0	0	1	0	1	0	0	0	0	0	0	0	0	0	0	0	0	1
0	2	0	0	2	0	3	0	0	0	0	0	0	0	0	0	0	0	0	2
0	1	0	0	0	0	1	0	0	0	0	0	0	0	0	1	0	0	0	1
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
0	1	0	0	1	0	1	0	0	0	0	0	0	0	0	0	0	0	0	2
0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0

EK-11 Macarca dili kesişim veri kümesi

HU	HU	HU	HU	HU	HU	HU	HU	HU	HU	HU	HU	HU	HU	HU	HU	HU	HU	HU	HU	
10	9	9	9	10	11	10	11	10	8	10	9	10	10	10	10	10	10	9	9	10
2	2	3	2	2	2	3	3	2	3	3	2	1	3	3	3	2	3	3	2	
1	1	1	1	1	1	1	1	1	1	2	1	1	1	1	1	1	1	1	1	
3	2	2	2	3	3	2	2	3	3	2	3	3	4	2	3	2	2	3	3	
11	12	12	11	10	11	11	10	10	10	10	11	9	10	11	10	10	10	11	10	
2	1	1	1	1	1	1	1	1	2	1	2	3	1	2	2	2	2	2	2	
4	4	4	4	4	5	5	5	4	4	4	4	4	3	4	4	4	4	3	4	
2	2	2	2	2	2	2	2	2	1	2	2	1	2	2	2	3	2	2	2	
6	4	5	5	5	6	5	5	5	5	5	6	5	4	5	6	5	4	5	5	
2	2	1	2	2	2	1	1	2	1	1	1	2	1	1	2	1	2	1	2	
7	6	6	5	5	6	5	6	6	5	6	5	5	5	6	5	5	6	6	6	
8	7	6	6	7	7	7	8	8	7	7	6	7	7	7	6	6	6	7	7	
5	4	5	5	4	4	4	4	4	4	3	3	3	4	4	3	3	4	4	3	
6	7	7	6	6	7	7	7	6	5	7	6	4	7	6	7	5	6	5	7	
5	4	4	4	5	5	5	5	4	6	5	5	5	4	5	4	4	5	5	4	
2	1	2	2	2	2	2	1	1	2	2	1	2	2	1	2	3	2	1	2	
0	0	0	0	1	1	0	1	0	0	1	1	0	1	0	0	0	0	0	0	
5	4	5	5	5	5	5	4	4	6	5	5	5	4	5	5	6	4	4	5	
8	7	7	7	7	8	6	7	8	7	6	7	8	8	7	7	8	7	7	7	
10	9	7	8	8	10	7	7	8	9	8	8	6	8	8	8	9	8	8	7	
2	2	2	2	2	2	1	2	1	1	2	2	2	2	1	2	1	1	1	2	
3	3	3	3	2	2	2	2	2	2	2	3	2	3	2	3	2	3	3	2	
0	0	0	1	1	1	0	1	0	1	1	1	1	1	1	1	0	0	0	1	
1	1	1	0	1	1	1	1	0	1	1	1	1	1	1	1	1	1	1	1	
3	2	3	3	2	3	3	3	3	2	3	2	2	2	2	3	2	3	2	3	
5	5	5	5	5	6	5	5	4	5	4	5	6	5	5	4	6	5	5	5	
0	4	3	4	4	1	4	5	4	5	4	4	4	5	4	4	4	4	4	4	
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
0	6	5	5	4	1	4	4	5	6	3	4	4	4							

EK-14 Türkçe dili kesişim veri kümesi

TR	TR	TR	TR	TR	TR	TR	TR	TR	TR	TR	TR	TR	TR	TR	TR	TR	TR	TR	TR
14	12	12	12	14	13	10	13	12	14	13	14	12	11	12	13	13	12	12	14
3	3	4	3	2	3	3	3	2	2	2	3	3	3	3	2	2	4	2	3
1	2	1	1	1	2	1	2	1	2	2	1	2	2	1	1	1	1	1	1
5	5	5	6	4	5	5	4	4	4	4	5	5	4	5	4	4	5	4	4
10	12	11	11	11	11	10	12	12	10	10	12	12	10	10	12	10	11	10	10
1	1	1	1	1	1	2	1	1	1	1	1	1	1	1	2	1	1	1	2
2	2	2	2	2	2	2	2	2	1	2	2	2	2	2	2	2	2	2	2
1	1	1	2	1	2	2	2	1	1	2	2	2	1	2	2	1	1	1	1
9	9	12	11	9	10	11	9	11	10	11	9	10	9	10	9	10	12	10	8
1	1	1	1	2	1	1	1	1	1	1	1	0	1	1	1	1	1	1	1
5	5	4	5	5	6	5	4	6	6	6	6	5	7	5	5	5	4	6	5
9	9	10	8	8	9	9	9	12	10	9	8	8	10	8	10	9	9	9	9
4	4	6	4	4	5	4	6	4	4	4	5	5	4	5	5	6	6	5	3
9	9	8	9	8	9	8	8	7	8	7	9	8	7	8	8	8	8	9	10
4	4	4	3	4	2	3	3	3	3	4	2	3	5	3	2	3	3	4	3
2	1	1	1	2	2	1	1	1	2	2	1	1	1	2	2	2	1	1	2
0	1	1	1	0	0	1	0	1	0	0	0	0	0	1	0	1	0	0	0
8	9	8	8	9	8	9	9	9	9	8	9	7	7	9	9	8	8	8	8
4	4	4	4	5	4	4	3	4	4	4	4	4	5	3	4	4	4	5	5
4	5	4	5	5	4	5	5	4	5	6	3	4	4	6	4	6	5	5	5
4	3	3	4	3	4	4	5	5	3	3	4	5	6	3	3	4	3	5	3
2	2	2	2	2	2	1	2	2	2	2	2	2	2	2	2	2	1	2	2
1	1	1	1	1	1	1	0	1	1	0	0	0	0	1	1	0	1	1	0
1	1	1	1	0	0	0	0	1	1	0	0	1	0	1	1	0	0	1	0
4	3	3	3	4	3	3	3	3	4	3	4	3	4	4	4	4	4	3	3
1	2	2	2	1	2	2	2	1	1	2	2	3	2	2	2	1	2	2	2
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
0	0	1	0	1	1	1	1	0	0	0	0	1	0	1	0	0	0	0	0
1	1	2	2	1	2	1	1	2	1	1	2	1	2	1	1	1	1	1	2
1	0	0	0	0	0	0	0	1	0	0	0	1	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
0	0	1	0	0	0	1	1	0	0	0	0	1	0	1	1	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
1	1	1	1	1	1	1	1	1	1	1	2	1	1	1	1	1	1	1	1
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0
3	2	2	3	3	2	5	2	2	3	2	3	2	2	3	4	2	2	2	4

EK-16 Birleşim tespit 1

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V	W	X	Y	Z	Á	À	Â	Ä
	a	b	c	d	e	f	g	h	i	j	k	l	m	n	o	p	q	r	s	t	u	v	w	x	y	z	á	à	â	ä
Almanca	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V	W	X	Y	Z				
Arnavutça	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V		X	Y	Z				Ä
Fransızca	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V	W	X	Y	Z		À	Â	
Galce	A	B	C	D	E	F	G	H	I			L	M	N	O	P		R	S	T	U		W		Y		Á	À	Â	Ä
Hırvatça	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P		R	S	T	U	V				Z				
İngilizce	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V	W	X	Y	Z		À	Â	Ä
İrlandaca	A	B	C	D	E	F	G	H	I			L	M	N	O	P		R	S	T	U						Á			
İspanyolca	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V	W	X	Y	Z	Á			
İtalyanca	A	B	C	D	E	F	G	H	I			L	M	N	O	P	Q	R	S	T	U	V				Z		À		
Letonca	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P		R	S	T	U	V				Z				
Macarca	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V	W	X	Y	Z	Á			
Maltaca	A	B		D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V	W	X		Z		À		
Portekizce	A	B	C	D	E	F	G	H	I	J		L	M	N	O	P	Q	R	S	T	U	V		X		Z	Á	À	Â	
Türkçe	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P		R	S	T	U	V			Y	Z			Â	
Vietnamca	A	B	C	D	E		G	H	I		K	L	M	N	O	P	Q	R	S	T	U	V		X	Y		Á	À	Â	
Tekrar	15	15	14	15	15	14	15	15	15	11	11	15	15	15	15	15	10	15	15	15	15	13	7	9	9	12	6	7	6	3

EK-17 Birleşim tespit 2

	Ã	Æ	Č	Ç	Ð	É	È	Ê	Ë	Í	Ì	Î	Ï	Ñ	Ó	Ò	Ô	Ö	Õ	Œ	Š	Ú	Ù	Û	Ü	Ý	Ỳ	Ŷ	Ỹ	Ž
	ã	æ	č	ç	ð	é	è	ê	ë	í	ì	î	ï	ñ	ó	ò	ô	ö	õ	œ	š	ú	ù	û	ü	ý	ỳ	ŷ	ỹ	ž
Almanca				Ç					Ë									Ö						Ü						
Arnavutça												Î					Ô						Û				Ŷ			
Fransızca		Æ		Ç		É	È	Ê	Ë			Î	Ï				Ô			Œ			Ù	Û	Ü				Ỹ	
Galce						É	È	Ê	Ë	Í	Ì	Î	Ï		Ó	Ò	Ô	Ö				Ú	Ù	Û	Ü	Ý	Ỳ	Ŷ	Ỹ	
Hırvatça			Č		Ð																Š									Ž
İngilizce		Æ		Ç		É	È	Ê	Ë			Î	Ï	Ñ	Ó		Ô	Ö		Œ				Û	Ü					
İrlandaca						É				Í					Ó							Ú								
İspanyolca						É				Í				Ñ	Ó							Ú			Ü					
İtalyanca						É	È				Ì	Î				Ò							Ù							
Letonca			Č																		Š									Ž
Macarca						É				Í					Ó			Ö				Ú			Ü					
Maltaca							È				Ì					Ò							Ù							
Portekizce	Ã			Ç		É		Ê		Í					Ó		Ô		Õ			Ú			Ü					
Türkçe				Ç								Î						Ö						Û	Ü					
Vietnamca	Ã				Ð	É	È	Ê		Í	Ì				Ó	Ò	Ô		Õ			Ú	Ù			Ý	Ỳ			
Tekrar	2	2	2	5	2	9	6	5	4	6	4	6	3	2	7	4	6	5	2	2	2	6	5	5	8	2	2	2	2	2

EK-18 Kesişim tespit 1

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V	W	X	Y	Z
	a	b	c	d	e	f	g	h	i	j	k	l	m	n	o	p	q	r	s	t	u	v	w	x	y	z
Almanca	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V	W	X	Y	Z
Arnavutça	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V		X	Y	Z
Fransızca	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V	W	X	Y	Z
Galce	A	B	C	D	E	F	G	H	I			L	M	N	O	P		R	S	T	U		W		Y	
Hırvatça	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P		R	S	T	U	V				Z
İngilizce	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V	W	X	Y	Z
İrlandaca	A	B	C	D	E	F	G	H	I			L	M	N	O	P		R	S	T	U					
İspanyolca	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V	W	X	Y	Z
İtalyanca	A	B	C	D	E	F	G	H	I			L	M	N	O	P	Q	R	S	T	U	V				Z
Letonca	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P		R	S	T	U	V				Z
Macarca	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V	W	X	Y	Z
Maltaca	A	B		D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V	W	X		Z
Portekizce	A	B	C	D	E	F	G	H	I	J		L	M	N	O	P	Q	R	S	T	U	V		X		Z
Türkçe	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P		R	S	T	U	V			Y	Z
Vietnamca	A	B	C	D	E		G	H	I		K	L	M	N	O	P	Q	R	S	T	U	V		X	Y	
Tekrar	15	15	14	15	15	14	15	15	15	11	11	15	15	15	15	15	10	15	15	15	15	13	7	9	9	12

EK-19 Kesişim tespit 2

	Á	À	Â	Ç	É	È	Ê	Í	Î	Ó	Ô	Ö	Ú	Ù	Û	Ü
	á	à	â	ç	é	è	ê	í	î	ó	ô	ö	ú	ù	û	ü
Almanca				Ç								Ö				Ü
Arnavutça									Î		Ô				Û	
Fransızca		À	Â	Ç	É	È	Ê		Î		Ô			Ù	Û	Ü
Galce	Á	À	Â		É	È	Ê	Í	Î	Ó	Ô	Ö	Ú	Ù	Û	Ü
Hırvatça																
İngilizce		À	Â	Ç	É	È	Ê		Î	Ó	Ô	Ö			Û	Ü
İrlandaca	Á				É			Í		Ó			Ú			
İspanyolca	Á				É			Í		Ó			Ú			Ü
İtalyanca		À			É	È			Î					Ù		
Letonca																
Macarca	Á				É			Í		Ó		Ö	Ú			Ü
Maltaca		À				È								Ù		
Portekizce	Á	À	Â	Ç	É		Ê	Í		Ó	Ô		Ú			Ü
Türkçe			Â	Ç					Î			Ö			Û	Ü
Vietnamca	Á	À	Â		É	È	Ê	Í		Ó	Ô		Ú	Ù		
Tekrar	6	7	6	5	9	6	5	6	6	7	6	5	6	5	5	8

EK-20 Genel alfabe kümesi - 1

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V	W	X	Y	Z
	a	b	c	d	e	f	g	h	i	j	k	l	m	n	o	p	q	r	s	t	u	v	w	x	y	z
Almanca	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V	W	X	Y	Z
Arnavutça	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V		X	Y	Z
Fransızca	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V	W	X	Y	Z
Galce	A	B	C	D	E	F	G	H	I			L	M	N	O	P		R	S	T	U		W		Y	
Hırvatça	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P		R	S	T	U	V				Z
İngilizce	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V	W	X	Y	Z
İrlandaca	A	B	C	D	E	F	G	H	I			L	M	N	O	P		R	S	T	U					
İspanyolca	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V	W	X	Y	Z
İtalyanca	A	B	C	D	E	F	G	H	I			L	M	N	O	P	Q	R	S	T	U	V				Z
Letonca	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P		R	S	T	U	V				Z
Macarca	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V	W	X	Y	Z
Maltaca	A	B		D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V	W	X		Z
Portekizce	A	B	C	D	E	F	G	H	I	J		L	M	N	O	P	Q	R	S	T	U	V		X		Z
Türkçe	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P		R	S	T	U	V			Y	Z
Vietnamca	A	B	C	D	E		G	H	I		K	L	M	N	O	P	Q	R	S	T	U	V		X	Y	
Tekrar	15	15	14	15	15	14	15	15	15	11	11	15	15	15	15	15	10	15	15	15	15	13	7	9	9	12

EK-21 Genel alfabe kümesi – 2

	Á	À	Â	Ä	Ǻ	Ǻ	Ā	Ã	Ǻ	Ǻ	Ǻ	Æ	Æ	Æ	Ɓ	Ć	Č	Ĉ	Č	Ç	Ǿ	Ɖ	Ɖ	Ɖ	Ɖ	Ɖ	É	È	Ê	Ê	
	á	à	â	ä	Ǻ	Ǻ	ā	ã	Ǻ	Ǻ	Ǻ	æ	æ	æ	ɓ	ć	č	ĉ	č	ç	d'	d	d	d	d	d	é	è	ê	ê	
Almanca																				Ç										Ë	
Arnavutça				Ä																											
Fransızca		À	Â									Æ								Ç							É	È		Ê	Ë
Galce	Á	À	Â	Ä																							É	È		Ê	Ë
Hırvatça																Ć			Č				Ɖ								
İngilizce		À	Â	Ä					Ǻ			Æ								Ç							É	È		Ê	Ë
İrlandaca	Á																										É				
İspanyolca	Á																										É				
İtalyanca		À																									É	È			
Letonca							Ā												Č												
Macarca	Á																										É				
Maltaca		À																Č										È			
Portekizce	Á	À	Â					Ã												Ç							É			Ê	
Türkçe			Â																	Ç											
Vietnamca	Á	À	Â			Ǻ		Ã															Ɖ				É	È		Ê	
Tekrar	6	7	6	3	0	1	1	2	1	0	0	2	0	0	0	1	1	0	2	5	0	0	2	0	0	0	9	6	0	5	4

EK-22 Genel alfabe kümesi – 3

	Ě	Ě	Ē	Ɛ	Ɛ	Ǝ	Ə	Ɛ	Ġ	Ġ	Ǧ	Ǧ	Ġ	Ƴ	Ĥ	Ĥ	Ĥ	ı	í	ì	ı	î	ĩ	ĩ	ĩ	ĩ	ĩ	ı	ı	ı
	ě	ě	ē	ɛ	ɛ	ə	ə	ɛ	ğ	ğ	ǧ	ǧ	ğ	Ʒ	ĥ	h	h	ı	í	ì	ı	î	ĩ	ĩ	ĩ	ĩ	ĩ	ı	ı	ı
Almanca																														
Arnavutça																						î								
Fransızca																						î	ĩ							
Galce																				í	ì		î	ĩ						
Hırvatça																														
İngilizce																						î	ĩ							
İrlandaca																				í										
İspanyolca																				í										
İtalyanca																					ì		î							
Letonca			Ē										Ģ													Ī				
Macarca																				í										
Maltaca									Ġ							Ħ					ì									
Portekizce																				í										
Türkçe												Ğ						ı			ı	î								
Vietnamca					Ẻ														í	ì							ĩ		ị	
Tekrar	0	0	1	0	1	0	0	0	1	0	0	1	1	0	0	0	1	1	6	4	1	6	3	0	0	1	1	0	1	0

EK-23 Genel alfabe kümesi – 4

												Ñ																		
	Ĵ	Ƙ	Ƙ	(ƙ)	Ĺ	Ł	Ł	Ł	Ł	'N	Ń		Ň	Ñ	Ŋ	ɖ	Ó	Ò	Ô	Ö	Ǿ	Ǿ	Ō	Õ	Ǫ	Ɔ	Ø	Ǿ	Ɔ	Ɔ
												ñ		ñ	ɳ	ɖ	ó	ò	ô	ö	Ǿ	Ǿ	ō	õ	ǫ	ɔ	ø	ǫ	ɔ	œ
	ĵ	ƙ	ƙ	ƙ	ĺ	ł	ł	ł	ł	ł	'n	ń		ň	ñ	ɳ	ɖ	ó	ò	ô	ö	Ǿ	Ǿ	ō	õ	ǫ	ɔ	ø	ǫ	ɔ
Almanca																				Ö										
Arnavutça																			Ô											
Fransızca																			Ô											Ɔ
Galce																	Ó	Ò	Ô	Ö										
Hırvatça																														
İngilizce													Ñ				Ó		Ô	Ö										Ɔ
İrlandaca																	Ó													
İspanyolca													Ñ				Ó													
İtalyanca																		Ò												
Letonca		Ƙ				Ł								Ŋ																
Macarca																	Ó			Ö						Ǫ				
Maltaca																		Ò												
Portekizce																	Ó		Ô					Õ						
Türkçe																				Ö										
Vietnamca																	Ó	Ò	Ô					Õ		Ɔ			Ɔ	
Tekrar	0	1	0	0	0	1	0	0	0	0	0	0	0	2	1	0	7	4	6	5	0	0	0	2	1	1	0	0	1	2

EK-24 Genel alfabe kümesi – 5

	Á	Ř	R̥	Ś	Ŝ	Š	Ş	Ș	Ŧ	(ɒ)	(B)	Ǻ	Ṭ	T̤	F̥	P̣	Ú	Ù	Û	Ü	Ǿ	Ǟ	Ū	Ŭ	Ǘ	ǖ	Ț	Ț	U'	W
	á	ř	r̄	ś	ŝ	š	ş	ș	ț	f	B	tʰ	t̵	t̸	f̴	p̵	ú	ù	û	ü	ǿ	ǟ	ū	ŭ	Ǚ	ǘ	ţ	ţ	u	w
Almanca											ß									Ü										
Arnavutça																			Û											
Fransızca																		Û	Û	Û										
Galce																	Ú	Ù	Û	Û										W
Hırvatça						Š																								
İngilizce																			Û	Û										
İrlandaca																	Ú													
İspanyolca																	Ú			Û										
İtalyanca																		Ù												
Letonca						Š																	Ū							
Macarca																	Ú			Û						Ű				
Maltaca																		Ù												
Portekizce																	Ú			Û										
Türkçe							Ş												Û	Û										
Vietnamca																	Ú	Ù						Ũ				Ư	Ư'	
Tekrar	0	0	0	0	0	2	1	0	0	0	1	0	0	0	0	0	6	5	5	8	0	0	1	1	1	0	0	1	1	1

EK-25 Genel alfabe kümesi – 6

	Ẁ	Ẃ	Ẅ	Ɔ	Ý	Ỳ	Ỳ	Ỳ	Ỳ	Ỳ	Ỳ	Ẑ	Ẓ	Ž	Ƶ
	ẁ	ẃ	ẅ	Ɔ	ý	ỳ	ỳ	ỳ	ỳ	ỳ	ƶ	ẑ	ẓ	ž	ƶ
Almanca															
Arnavutça							Ỳ								
Fransızca								Ỳ							
Galce	Ẁ	Ẃ	Ẅ		Ý	Ỳ	Ỳ	Ỳ							
Hırvatça														Ž	
İngilizce															
İrlandaca															
İspanyolca															
İtalyanca															
Letonca														Ž	
Macarca															
Maltaca													Ž		
Portekizce															
Türkçe															
Vietnamca					Ý	Ỳ				Ỳ					
Tekrar	1	1	1	0	2	2	2	2	0	1	0	0	1	2	0

EK-26 Almanca dili GENIUS test sonuçları

Dil	Almanca																			
Format	PDF Dosya Tipi		DOC Dosya Tipi		PDF Dosya Tipi		DOC Dosya Tipi		PDF Dosya Tipi		DOC Dosya Tipi		PDF Dosya Tipi		DOC Dosya Tipi		PDF Dosya Tipi		DOC Dosya Tipi	
Yöntem	#1	#2	#1	#2	#1	#2	#1	#2	#1	#2	#1	#2	#1	#2	#1	#2	#1	#2	#1	#2
Almanca	77%	91%	92%	83%	95%	93%	98%	83%	98%	95%	69%	98%	96%	97%	94%	93%	97%	96%	93%	92%
Arnavutça	38%	54%	53%	58%	41%	46%	78%	41%	90%	52%	65%	89%	48%	61%	92%	46%	51%	48%	63%	93%
Fransızca	25%	36%	35%	38%	94%	31%	81%	27%	72%	34%	43%	73%	32%	40%	61%	31%	34%	32%	42%	62%
Galce	19%	27%	26%	29%	79%	23%	60%	20%	54%	26%	32%	55%	24%	30%	46%	23%	25%	24%	31%	46%
Hırvatça	15%	21%	21%	23%	63%	18%	48%	16%	43%	20%	26%	44%	19%	24%	37%	18%	20%	19%	25%	37%
İngilizce	12%	18%	17%	19%	52%	15%	40%	13%	36%	17%	21%	36%	16%	20%	30%	15%	17%	16%	21%	31%
İrlandaca	11%	15%	15%	16%	45%	13%	34%	11%	31%	14%	18%	31%	13%	17%	26%	13%	14%	13%	18%	26%
İspanyolca	9%	13%	13%	14%	39%	11%	30%	10%	27%	13%	16%	27%	12%	15%	23%	11%	12%	12%	15%	23%
İtalyanca	8%	12%	11%	12%	35%	10%	27%	9%	24%	11%	14%	24%	10%	13%	20%	10%	11%	10%	14%	20%
Letonca	7%	10%	10%	11%	31%	9%	24%	8%	21%	10%	13%	22%	9%	12%	18%	9%	10%	9%	12%	18%
Macarca	7%	9%	9%	10%	28%	8%	22%	7%	19%	9%	11%	20%	8%	11%	16%	8%	9%	8%	11%	17%
Maltaca	6%	9%	8%	9%	26%	7%	20%	6%	18%	8%	10%	18%	8%	10%	15%	7%	8%	8%	10%	15%
Portekizce	5%	8%	8%	8%	24%	7%	18%	6%	16%	8%	10%	16%	7%	9%	14%	7%	7%	7%	9%	14%
Türkçe	5%	7%	7%	8%	22%	6%	17%	5%	15%	7%	9%	15%	6%	8%	13%	6%	7%	6%	9%	13%
Vietnamca	5%	7%	7%	7%	21%	6%	16%	5%	14%	6%	8%	14%	6%	8%	12%	6%	6%	6%	8%	12%

EK-27 Arnavutça dili GENIUS test sonuçları

Dil	Arnavutça																			
Format	PDF Dosya Tipi		DOC Dosya Tipi		PDF Dosya Tipi		DOC Dosya Tipi		PDF Dosya Tipi		DOC Dosya Tipi		PDF Dosya Tipi		DOC Dosya Tipi		PDF Dosya Tipi		DOC Dosya Tipi	
Yöntem	#1	#2	#1	#2	#1	#2	#1	#2	#1	#2	#1	#2	#1	#2	#1	#2	#1	#2	#1	#2
Almanca	0%	0%	75%	0%	17%	50%	13%	0%	0%	0%	1%	0%	0%	0%	0%	0%	0%	0%	0%	0%
Arnavutça	99%	99%	62%	90%	91%	74%	93%	99%	100%	100%	99%	99%	99%	99%	98%	99%	88%	52%	99%	99%
Fransızca	66%	66%	41%	72%	60%	49%	62%	66%	66%	66%	66%	66%	66%	66%	67%	66%	74%	98%	0%	66%
Galce	49%	50%	31%	54%	45%	37%	46%	50%	50%	50%	49%	49%	49%	50%	50%	50%	55%	73%	22%	49%
Hırvatça	39%	40%	24%	43%	36%	29%	37%	40%	40%	40%	39%	39%	39%	40%	40%	40%	44%	59%	58%	39%
İngilizce	33%	33%	20%	36%	30%	24%	31%	33%	33%	33%	33%	33%	33%	33%	33%	33%	37%	49%	81%	33%
İrlandaca	28%	28%	17%	31%	26%	21%	26%	28%	28%	28%	28%	28%	28%	28%	28%	28%	31%	42%	98%	28%
İspanyolca	24%	25%	15%	27%	22%	18%	23%	25%	25%	25%	24%	24%	24%	25%	25%	25%	27%	36%	88%	24%
İtalyanca	22%	22%	13%	24%	20%	16%	20%	22%	22%	22%	22%	22%	22%	22%	22%	22%	24%	32%	78%	22%
Letonca	19%	20%	12%	21%	18%	14%	18%	20%	20%	20%	19%	19%	19%	20%	20%	20%	22%	29%	70%	19%
Macarca	18%	18%	11%	19%	16%	13%	16%	18%	18%	18%	18%	18%	18%	18%	18%	18%	20%	26%	64%	18%
Maltaca	16%	16%	10%	18%	15%	12%	15%	16%	16%	16%	16%	16%	16%	16%	16%	16%	18%	24%	59%	16%
Portekizce	15%	15%	9%	16%	14%	11%	14%	15%	15%	15%	15%	15%	15%	15%	15%	15%	17%	22%	54%	15%
Türkçe	14%	14%	8%	15%	13%	10%	13%	14%	14%	14%	14%	14%	14%	14%	14%	14%	15%	21%	50%	14%
Vietnamca	13%	13%	8%	14%	12%	9%	12%	13%	13%	13%	13%	13%	13%	13%	13%	13%	14%	19%	47%	13%

EK-28 Fransızca dili GENIUS test sonuçları

Dil	Fransızca																			
Format	PDF Dosya Tipi		DOC Dosya Tipi		PDF Dosya Tipi		DOC Dosya Tipi		PDF Dosya Tipi		DOC Dosya Tipi		PDF Dosya Tipi		DOC Dosya Tipi		PDF Dosya Tipi		DOC Dosya Tipi	
Yöntem	#1	#2	#1	#2	#1	#2	#1	#2	#1	#2	#1	#2	#1	#2	#1	#2	#1	#2	#1	#2
Almanca	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%
Arnavutça	31%	61%	47%	33%	0%	8%	68%	25%	46%	2%	32%	57%	96%	50%	15%	12%	2%	28%	53%	42%
Fransızca	87%	92%	98%	89%	99%	72%	87%	83%	97%	68%	88%	95%	68%	99%	76%	75%	68%	85%	97%	95%
Galce	84%	69%	76%	83%	0%	95%	65%	87%	76%	98%	83%	71%	51%	74%	92%	93%	98%	85%	73%	78%
Hırvatça	67%	55%	60%	66%	11%	76%	52%	69%	61%	78%	66%	57%	41%	59%	73%	74%	79%	68%	58%	62%
İngilizce	56%	46%	50%	55%	43%	63%	43%	58%	51%	65%	55%	47%	34%	49%	61%	62%	65%	57%	48%	52%
İrlandaca	48%	39%	43%	47%	65%	54%	37%	49%	43%	56%	47%	40%	29%	42%	52%	53%	56%	48%	41%	44%
İspanyolca	42%	34%	38%	41%	82%	47%	32%	43%	38%	49%	41%	35%	25%	37%	46%	46%	49%	42%	36%	39%
İtalyanca	37%	30%	33%	36%	95%	42%	29%	38%	34%	43%	37%	31%	22%	33%	41%	41%	43%	38%	32%	34%
Letonca	33%	27%	30%	33%	94%	38%	26%	34%	30%	39%	33%	28%	20%	29%	36%	37%	39%	34%	29%	31%
Macarca	30%	25%	27%	30%	85%	34%	23%	31%	27%	35%	30%	25%	18%	27%	33%	34%	35%	31%	26%	28%
Maltaca	28%	23%	25%	27%	78%	31%	21%	29%	25%	32%	27%	23%	17%	24%	30%	31%	32%	28%	24%	26%
Portekizce	25%	21%	23%	25%	72%	29%	20%	26%	23%	30%	25%	21%	15%	22%	28%	28%	30%	26%	22%	24%
Türkçe	24%	19%	21%	23%	67%	27%	18%	24%	21%	28%	23%	20%	14%	21%	26%	26%	28%	24%	20%	22%
Vietnamca	22%	18%	20%	22%	62%	25%	17%	23%	20%	26%	22%	19%	13%	19%	24%	24%	26%	22%	19%	20%

EK-29 Galce dili GENIUS test sonuçları

Dil	Galce																			
Format	PDF Dosya Tipi		DOC Dosya Tipi		PDF Dosya Tipi		DOC Dosya Tipi		PDF Dosya Tipi		DOC Dosya Tipi		PDF Dosya Tipi		DOC Dosya Tipi		PDF Dosya Tipi		DOC Dosya Tipi	
Yöntem	#1	#2	#1	#2	#1	#2	#1	#2	#1	#2	#1	#2	#1	#2	#1	#2	#1	#2	#1	#2
Almanca	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%
Arnavutça	0%	0%	34%	6%	0%	0%	0%	0%	0%	0%	0%	6%	4%	10%	7%	0%	4%	0%	0%	0%
Fransızca	22%	60%	89%	71%	66%	66%	65%	61%	66%	66%	67%	70%	69%	73%	71%	53%	69%	64%	0%	45%
Galce	66%	95%	82%	96%	100%	100%	99%	96%	100%	100%	99%	96%	97%	94%	96%	89%	97%	98%	22%	84%
Hırvatça	93%	83%	66%	77%	80%	80%	80%	82%	80%	80%	79%	77%	78%	75%	76%	88%	78%	81%	57%	92%
İngilizce	88%	69%	55%	64%	66%	66%	67%	69%	66%	66%	66%	64%	65%	63%	64%	73%	65%	67%	81%	77%
İrlandaca	76%	59%	47%	55%	57%	57%	57%	59%	57%	57%	56%	55%	55%	54%	54%	62%	55%	58%	98%	66%
İspanyolca	66%	52%	41%	48%	50%	50%	50%	51%	50%	50%	49%	48%	48%	47%	48%	55%	48%	50%	88%	57%
İtalyanca	59%	46%	36%	42%	44%	44%	44%	46%	44%	44%	44%	43%	43%	42%	42%	48%	43%	45%	79%	51%
Letonca	53%	41%	33%	38%	40%	40%	40%	41%	40%	40%	39%	38%	39%	37%	38%	44%	39%	40%	71%	46%
Macarca	48%	38%	30%	35%	36%	36%	36%	37%	36%	36%	36%	35%	35%	34%	34%	40%	35%	36%	64%	42%
Maltaca	44%	34%	27%	32%	33%	33%	33%	34%	33%	33%	33%	32%	32%	31%	32%	36%	32%	33%	59%	38%
Portekizce	40%	32%	25%	29%	30%	30%	30%	31%	30%	30%	30%	29%	30%	29%	29%	33%	30%	31%	54%	35%
Türkçe	38%	29%	23%	27%	28%	28%	28%	29%	28%	28%	28%	27%	27%	27%	27%	31%	27%	29%	50%	33%
Vietnamca	35%	27%	22%	25%	26%	26%	26%	27%	26%	26%	26%	25%	26%	25%	25%	29%	26%	27%	47%	30%

EK-30 Hırvatça dili GENIUS test sonuçları

Dil	Hırvatça																			
Format	PDF Dosya Tipi		DOC Dosya Tipi		PDF Dosya Tipi		DOC Dosya Tipi		PDF Dosya Tipi		DOC Dosya Tipi		PDF Dosya Tipi		DOC Dosya Tipi		PDF Dosya Tipi		DOC Dosya Tipi	
Yöntem	#1	#2	#1	#2	#1	#2	#1	#2	#1	#2	#1	#2	#1	#2	#1	#2	#1	#2	#1	#2
Almanca	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%
Arnavutça	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%
Fransızca	33%	33%	37%	42%	28%	33%	37%	21%	30%	23%	35%	33%	0%	31%	8%	33%	33%	33%	40%	50%
Galce	75%	75%	78%	82%	71%	74%	77%	66%	72%	67%	76%	75%	0%	73%	56%	75%	75%	75%	80%	87%
Hırvatça	100%	100%	97%	94%	97%	99%	97%	93%	98%	94%	98%	99%	35%	98%	84%	99%	100%	100%	95%	89%
İngilizce	83%	83%	81%	78%	85%	83%	81%	89%	84%	88%	82%	83%	62%	84%	95%	83%	83%	83%	79%	74%
İrlandaca	71%	71%	69%	67%	73%	71%	69%	76%	72%	75%	70%	71%	82%	72%	82%	71%	71%	71%	68%	64%
İspanyolca	62%	62%	60%	58%	64%	62%	61%	66%	63%	66%	61%	62%	97%	63%	71%	62%	62%	62%	59%	56%
İtalyanca	55%	55%	54%	52%	57%	55%	54%	59%	56%	58%	54%	55%	91%	56%	63%	55%	55%	55%	53%	49%
Letonca	50%	50%	48%	47%	51%	50%	48%	53%	50%	52%	49%	49%	82%	50%	57%	49%	50%	50%	47%	44%
Macarca	45%	45%	44%	42%	46%	45%	44%	48%	46%	48%	44%	45%	74%	45%	52%	45%	45%	45%	43%	40%
Maltaca	41%	41%	40%	39%	42%	41%	40%	44%	42%	44%	41%	41%	68%	42%	47%	41%	41%	41%	39%	37%
Portekizce	38%	38%	37%	36%	39%	38%	37%	41%	39%	40%	37%	38%	63%	38%	44%	38%	38%	38%	36%	34%
Türkçe	35%	35%	34%	33%	36%	35%	34%	38%	36%	37%	35%	35%	58%	36%	41%	35%	35%	35%	34%	32%
Vietnamca	33%	33%	32%	31%	34%	33%	32%	35%	33%	35%	32%	33%	54%	33%	38%	33%	33%	33%	31%	29%

EK-31 İngilizce dili GENIUS test sonuçları

Dil	İngilizce																			
Format	PDF Dosya Tipi		DOC Dosya Tipi		PDF Dosya Tipi		DOC Dosya Tipi		PDF Dosya Tipi		DOC Dosya Tipi		PDF Dosya Tipi		DOC Dosya Tipi		PDF Dosya Tipi		DOC Dosya Tipi	
Yöntem	#1	#2	#1	#2	#1	#2	#1	#2	#1	#2	#1	#2	#1	#2	#1	#2	#1	#2	#1	#2
Almanca	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%
Arnavutça	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%
Fransızca	0%	4%	0%	0%	25%	2%	9%	0%	6%	17%	9%	0%	9%	4%	0%	0%	7%	0%	8%	0%
Galce	50%	53%	47%	50%	69%	52%	56%	46%	55%	62%	56%	49%	56%	53%	30%	41%	55%	6%	56%	48%
Hırvatça	80%	82%	78%	80%	95%	81%	85%	77%	84%	90%	85%	79%	85%	82%	64%	73%	84%	45%	85%	78%
İngilizce	99%	97%	98%	99%	87%	98%	95%	97%	96%	91%	95%	99%	95%	97%	86%	94%	96%	70%	95%	99%
İrlandaca	85%	83%	86%	85%	74%	84%	81%	87%	82%	78%	81%	86%	81%	83%	97%	90%	82%	89%	81%	86%
İspanyolca	74%	73%	76%	74%	65%	73%	71%	76%	72%	68%	71%	75%	71%	73%	84%	79%	72%	96%	71%	75%
İtalyanca	66%	65%	67%	66%	58%	65%	63%	68%	64%	60%	63%	66%	63%	65%	75%	70%	64%	86%	63%	67%
Letonca	59%	58%	60%	59%	52%	59%	57%	61%	57%	54%	57%	60%	57%	58%	67%	63%	57%	77%	57%	60%
Macarca	54%	53%	55%	54%	47%	53%	52%	55%	52%	49%	52%	54%	52%	53%	61%	57%	52%	70%	52%	55%
Maltaca	49%	48%	50%	49%	43%	49%	47%	51%	48%	45%	47%	50%	47%	48%	56%	52%	48%	64%	47%	50%
Portekizce	46%	45%	46%	46%	40%	45%	44%	47%	44%	42%	44%	46%	44%	45%	52%	48%	44%	59%	44%	46%
Türkçe	42%	41%	43%	42%	37%	42%	40%	43%	41%	39%	40%	43%	40%	41%	48%	45%	41%	55%	40%	43%
Vietnamca	39%	39%	40%	39%	34%	39%	38%	40%	38%	36%	38%	40%	38%	39%	45%	42%	38%	51%	38%	40%

EK-32 İrlandaca dili GENIUS test sonuçları

Dil	İrlandaca																			
Format	PDF Dosya Tipi		DOC Dosya Tipi		PDF Dosya Tipi		DOC Dosya Tipi		PDF Dosya Tipi		DOC Dosya Tipi		PDF Dosya Tipi		DOC Dosya Tipi		PDF Dosya Tipi		DOC Dosya Tipi	
Yöntem	#1	#2	#1	#2	#1	#2	#1	#2	#1	#2	#1	#2	#1	#2	#1	#2	#1	#2	#1	#2
Almanca	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%
Arnavutça	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%
Fransızca	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%
Galce	3%	18%	28%	26%	25%	3%	27%	30%	25%	21%	25%	25%	0%	21%	25%	26%	17%	25%	28%	29%
Hırvatça	43%	54%	62%	61%	60%	43%	61%	64%	60%	56%	60%	60%	29%	57%	60%	60%	53%	60%	63%	63%
İngilizce	69%	78%	85%	84%	83%	69%	84%	86%	83%	80%	83%	83%	58%	81%	83%	84%	78%	83%	85%	86%
İrlandaca	87%	96%	98%	99%	99%	87%	98%	96%	99%	97%	99%	99%	78%	98%	99%	99%	95%	99%	97%	97%
İspanyolca	98%	90%	85%	86%	87%	98%	86%	84%	87%	89%	87%	87%	93%	89%	87%	86%	91%	87%	85%	85%
İtalyanca	87%	80%	76%	77%	77%	87%	76%	75%	77%	79%	77%	77%	94%	79%	77%	77%	81%	77%	76%	75%
Letonca	78%	72%	68%	69%	69%	78%	69%	67%	69%	71%	69%	69%	85%	71%	69%	69%	73%	69%	68%	68%
Macarca	71%	66%	62%	63%	63%	71%	62%	61%	63%	65%	63%	63%	77%	64%	63%	63%	66%	63%	62%	62%
Maltaca	65%	60%	57%	57%	58%	65%	57%	56%	58%	59%	58%	58%	70%	59%	58%	57%	60%	58%	57%	56%
Portekizce	60%	55%	52%	53%	53%	60%	53%	52%	53%	55%	53%	53%	65%	54%	53%	53%	56%	53%	52%	52%
Türkçe	56%	51%	49%	49%	49%	56%	49%	48%	49%	51%	49%	49%	60%	50%	49%	49%	52%	49%	48%	48%
Vietnamca	52%	48%	45%	46%	46%	52%	46%	45%	46%	47%	46%	46%	56%	47%	46%	46%	48%	46%	45%	45%

EK-33 İspanyolca dili GENIUS test sonuçları

Dil	İspanyolca																			
Format	PDF Dosya Tipi		DOC Dosya Tipi		PDF Dosya Tipi		DOC Dosya Tipi		PDF Dosya Tipi		DOC Dosya Tipi		PDF Dosya Tipi		DOC Dosya Tipi		PDF Dosya Tipi		DOC Dosya Tipi	
Yöntem	#1	#2	#1	#2	#1	#2	#1	#2	#1	#2	#1	#2	#1	#2	#1	#2	#1	#2	#1	#2
Almanca	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%
Arnavutça	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%
Fransızca	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%
Galce	11%	22%	0%	4%	0%	0%	3%	15%	0%	0%	0%	0%	0%	0%	0%	1%	0%	22%	0%	2%
Hırvatça	49%	58%	39%	43%	32%	40%	42%	52%	6%	33%	38%	38%	0%	37%	31%	40%	19%	57%	34%	42%
İngilizce	74%	81%	65%	69%	60%	66%	69%	77%	38%	61%	65%	65%	32%	64%	59%	67%	49%	81%	62%	68%
İrlandaca	92%	98%	85%	88%	80%	85%	87%	94%	61%	80%	84%	84%	56%	83%	79%	86%	71%	98%	82%	87%
İspanyolca	94%	88%	99%	97%	95%	99%	98%	92%	79%	95%	99%	98%	74%	98%	94%	99%	87%	88%	96%	98%
İtalyanca	83%	78%	89%	87%	92%	88%	87%	81%	92%	92%	89%	89%	88%	90%	93%	88%	99%	78%	91%	87%
Letonca	75%	70%	80%	78%	83%	79%	78%	73%	96%	83%	80%	80%	99%	81%	84%	79%	90%	71%	82%	78%
Macarca	68%	64%	73%	71%	76%	72%	71%	67%	87%	75%	73%	73%	91%	73%	76%	72%	82%	64%	75%	71%
Maltaca	62%	59%	67%	65%	69%	66%	65%	61%	80%	69%	67%	67%	83%	67%	70%	66%	75%	59%	68%	65%
Portekizce	58%	54%	61%	60%	64%	61%	60%	56%	74%	64%	62%	62%	77%	62%	64%	61%	69%	54%	63%	60%
Türkçe	53%	50%	57%	55%	59%	57%	56%	52%	69%	59%	57%	57%	71%	58%	60%	56%	64%	50%	58%	56%
Vietnamca	50%	47%	53%	52%	55%	53%	52%	49%	64%	55%	53%	53%	66%	54%	56%	53%	60%	47%	55%	52%

EK-34 İtalyanca dili GENIUS test sonuçları

Dil	İtalyanca																			
Format	PDF Dosya Tipi		DOC Dosya Tipi		PDF Dosya Tipi		DOC Dosya Tipi		PDF Dosya Tipi		DOC Dosya Tipi		PDF Dosya Tipi		DOC Dosya Tipi		PDF Dosya Tipi		DOC Dosya Tipi	
Yöntem	#1	#2	#1	#2	#1	#2	#1	#2	#1	#2	#1	#2	#1	#2	#1	#2	#1	#2	#1	#2
Almanca	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%
Arnavutça	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%
Fransızca	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%
Galce	0%	0%	0%	0%	0%	0%	0%	0%	28%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%
Hırvatça	23%	19%	26%	38%	0%	21%	0%	20%	63%	14%	10%	19%	8%	21%	8%	17%	0%	18%	29%	20%
İngilizce	52%	49%	55%	65%	19%	51%	10%	50%	85%	45%	42%	49%	40%	51%	40%	48%	27%	49%	57%	50%
İrlandaca	73%	71%	75%	84%	45%	72%	37%	71%	97%	67%	64%	70%	63%	72%	63%	69%	52%	70%	78%	71%
İspanyolca	89%	87%	91%	99%	64%	88%	58%	87%	85%	84%	81%	87%	80%	88%	80%	86%	70%	86%	93%	87%
İtalyanca	98%	99%	96%	89%	79%	99%	73%	99%	76%	97%	94%	99%	93%	98%	93%	98%	85%	99%	94%	99%
Letonca	88%	90%	86%	80%	91%	89%	86%	89%	68%	92%	94%	90%	95%	89%	95%	91%	96%	90%	85%	89%
Macarca	80%	81%	79%	73%	98%	81%	96%	81%	62%	84%	86%	82%	87%	80%	87%	82%	93%	82%	77%	81%
Maltaca	73%	75%	72%	67%	90%	74%	94%	74%	57%	77%	78%	75%	79%	74%	79%	75%	86%	75%	71%	74%
Portekizce	67%	69%	66%	62%	83%	68%	87%	68%	52%	71%	72%	69%	73%	68%	73%	70%	79%	69%	65%	68%
Türkçe	63%	64%	62%	57%	77%	63%	81%	64%	48%	66%	67%	64%	68%	63%	68%	65%	73%	64%	60%	64%
Vietnamca	58%	60%	57%	53%	72%	59%	75%	59%	45%	61%	63%	60%	63%	59%	63%	60%	68%	60%	56%	59%

EK-35 Letonca dili GENIUS test sonuçları

Dil	Letonca																			
Format	PDF Dosya Tipi		DOC Dosya Tipi		PDF Dosya Tipi		DOC Dosya Tipi		PDF Dosya Tipi		DOC Dosya Tipi		PDF Dosya Tipi		DOC Dosya Tipi		PDF Dosya Tipi		DOC Dosya Tipi	
Yöntem	#1	#2	#1	#2	#1	#2	#1	#2	#1	#2	#1	#2	#1	#2	#1	#2	#1	#2	#1	#2
Almanca	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%
Arnavutça	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%
Fransızca	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%
Galce	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%
Hırvatça	8%	0%	0%	0%	0%	0%	4%	0%	11%	0%	30%	0%	32%	0%	0%	0%	0%	0%	7%	0%
İngilizce	40%	16%	26%	33%	18%	3%	36%	33%	42%	33%	58%	33%	60%	32%	24%	33%	28%	33%	39%	33%
İrlandaca	63%	42%	51%	57%	44%	31%	60%	56%	65%	56%	78%	57%	80%	56%	49%	57%	52%	57%	62%	56%
İspanyolca	80%	62%	69%	74%	63%	52%	77%	74%	81%	74%	93%	74%	95%	74%	68%	74%	71%	74%	79%	74%
İtalyanca	93%	77%	84%	88%	78%	68%	91%	88%	95%	88%	94%	88%	93%	88%	82%	88%	85%	88%	93%	88%
Letonca	95%	89%	95%	99%	90%	81%	97%	99%	94%	99%	84%	99%	83%	99%	94%	99%	96%	99%	96%	99%
Macarca	87%	99%	94%	90%	99%	92%	89%	91%	85%	91%	77%	90%	76%	91%	95%	90%	93%	90%	87%	91%
Maltaca	79%	91%	86%	83%	90%	98%	81%	83%	78%	83%	70%	83%	69%	83%	87%	83%	85%	83%	80%	83%
Portekizce	73%	84%	80%	76%	83%	90%	75%	77%	72%	77%	65%	76%	64%	77%	81%	76%	79%	76%	73%	77%
Türkçe	68%	78%	74%	71%	77%	84%	69%	71%	67%	71%	60%	71%	59%	71%	75%	71%	73%	71%	68%	71%
Vietnamca	63%	73%	69%	66%	72%	78%	65%	66%	62%	66%	56%	66%	55%	67%	70%	66%	68%	66%	64%	66%

EK-36 Macarca dili GENIUS test sonuçları

Dil	Macarca																			
Format	PDF Dosya Tipi		DOC Dosya Tipi		PDF Dosya Tipi		DOC Dosya Tipi		PDF Dosya Tipi		DOC Dosya Tipi		PDF Dosya Tipi		DOC Dosya Tipi		PDF Dosya Tipi		DOC Dosya Tipi	
Yöntem	#1	#2	#1	#2	#1	#2	#1	#2	#1	#2	#1	#2	#1	#2	#1	#2	#1	#2	#1	#2
Almanca	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%
Arnavutça	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%
Fransızca	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%
Galce	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%
Hırvatça	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%
İngilizce	16%	21%	22%	12%	16%	18%	12%	23%	6%	16%	3%	26%	6%	18%	17%	20%	12%	15%	13%	19%
İrlandaca	42%	46%	47%	39%	43%	44%	39%	49%	33%	42%	31%	51%	34%	44%	43%	45%	39%	41%	40%	45%
İspanyolca	62%	65%	66%	59%	62%	64%	59%	67%	54%	62%	52%	69%	54%	64%	63%	65%	59%	61%	60%	64%
İtalyanca	77%	80%	81%	74%	77%	79%	74%	82%	70%	77%	69%	84%	71%	79%	78%	80%	74%	76%	75%	79%
Letonca	90%	92%	93%	87%	90%	91%	87%	94%	83%	89%	82%	95%	83%	91%	90%	92%	87%	89%	88%	91%
Macarca	99%	97%	96%	97%	99%	98%	97%	96%	94%	99%	92%	94%	94%	98%	99%	98%	97%	99%	98%	98%
Maltaca	91%	89%	88%	93%	91%	90%	93%	88%	96%	91%	98%	86%	96%	90%	91%	89%	93%	92%	93%	90%
Portekizce	84%	82%	82%	86%	84%	83%	86%	81%	89%	84%	90%	80%	89%	83%	84%	83%	86%	85%	85%	83%
Türkçe	78%	76%	76%	80%	78%	77%	80%	75%	83%	78%	84%	74%	82%	77%	78%	77%	80%	79%	79%	77%
Vietnamca	73%	71%	71%	75%	73%	72%	75%	70%	77%	73%	78%	69%	77%	72%	72%	71%	75%	73%	74%	72%

EK-37 Maltaca dili GENIUS test sonuçları

Dil	Maltaca																			
Format	PDF Dosya Tipi		DOC Dosya Tipi		PDF Dosya Tipi		DOC Dosya Tipi		PDF Dosya Tipi		DOC Dosya Tipi		PDF Dosya Tipi		DOC Dosya Tipi		PDF Dosya Tipi		DOC Dosya Tipi	
Yöntem	#1	#2	#1	#2	#1	#2	#1	#2	#1	#2	#1	#2	#1	#2	#1	#2	#1	#2	#1	#2
Almanca	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%
Arnavutça	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%
Fransızca	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%
Galce	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%
Hırvatça	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%
İngilizce	0%	0%	16%	0%	0%	0%	32%	0%	0%	3%	0%	0%	22%	2%	0%	0%	14%	0%	0%	4%
İrlandaca	28%	28%	42%	28%	12%	26%	56%	28%	20%	31%	28%	28%	47%	30%	26%	28%	41%	29%	27%	32%
İspanyolca	50%	50%	62%	49%	35%	48%	74%	49%	43%	52%	49%	49%	66%	51%	48%	50%	61%	50%	49%	53%
İtalyanca	66%	66%	77%	66%	54%	65%	88%	66%	60%	69%	66%	66%	81%	68%	65%	66%	76%	67%	65%	69%
Letonca	80%	80%	89%	79%	68%	78%	99%	79%	74%	82%	79%	79%	93%	81%	78%	80%	88%	80%	79%	82%
Macarca	90%	90%	99%	90%	80%	89%	91%	90%	85%	93%	90%	90%	96%	92%	89%	90%	99%	91%	90%	93%
Maltaca	100%	100%	91%	99%	90%	98%	83%	99%	95%	98%	99%	99%	88%	98%	98%	99%	92%	99%	99%	97%
Portekizce	92%	92%	84%	92%	99%	93%	77%	92%	96%	90%	92%	92%	82%	91%	93%	92%	85%	92%	92%	90%
Türkçe	85%	85%	78%	85%	93%	86%	71%	85%	89%	84%	85%	85%	76%	84%	86%	85%	79%	85%	86%	83%
Vietnamca	80%	80%	73%	80%	87%	80%	66%	80%	83%	78%	80%	80%	71%	78%	80%	79%	74%	79%	80%	78%

EK-38 Portekizce dili GENIUS test sonuçları

Dil	Portekizce																			
Format	PDF Dosya Tipi		DOC Dosya Tipi		PDF Dosya Tipi		DOC Dosya Tipi		PDF Dosya Tipi		DOC Dosya Tipi		PDF Dosya Tipi		DOC Dosya Tipi		PDF Dosya Tipi		DOC Dosya Tipi	
Yöntem	#1	#2	#1	#2	#1	#2	#1	#2	#1	#2	#1	#2	#1	#2	#1	#2	#1	#2	#1	#2
Almanca	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%
Arnavutça	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%
Fransızca	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%
Galce	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%
Hırvatça	0%	0%	0%	0%	0%	0%	38%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%
İngilizce	2%	0%	0%	0%	0%	0%	65%	0%	0%	0%	0%	0%	6%	0%	3%	0%	0%	0%	0%	0%
İrlandaca	30%	13%	6%	16%	5%	12%	84%	15%	4%	14%	10%	16%	33%	17%	31%	16%	23%	13%	27%	17%
İspanyolca	51%	37%	30%	39%	29%	35%	99%	38%	28%	37%	33%	39%	54%	40%	52%	39%	45%	36%	49%	40%
İtalyanca	68%	55%	49%	56%	48%	53%	89%	56%	47%	55%	52%	57%	70%	58%	68%	57%	62%	54%	65%	58%
Letonca	81%	69%	64%	71%	63%	68%	80%	70%	63%	69%	67%	71%	83%	72%	82%	71%	76%	69%	79%	72%
Macarca	92%	81%	76%	82%	75%	80%	73%	82%	75%	81%	79%	83%	94%	83%	92%	83%	87%	81%	90%	84%
Maltaca	98%	91%	86%	92%	86%	90%	67%	92%	85%	91%	89%	93%	96%	93%	98%	92%	96%	91%	99%	93%
Portekizce	91%	99%	95%	99%	95%	98%	61%	99%	94%	99%	97%	98%	89%	98%	90%	98%	95%	99%	92%	98%
Türkçe	84%	93%	96%	91%	97%	93%	57%	92%	97%	92%	94%	91%	83%	91%	84%	91%	88%	93%	86%	91%
Vietnamca	79%	86%	90%	85%	90%	87%	53%	86%	91%	86%	88%	85%	77%	85%	78%	85%	82%	87%	80%	85%

EK-39 Türkçe dili GENIUS test sonuçları

Dil	Türkçe																			
Format	PDF Dosya Tipi		DOC Dosya Tipi		PDF Dosya Tipi		DOC Dosya Tipi		PDF Dosya Tipi		DOC Dosya Tipi		PDF Dosya Tipi		DOC Dosya Tipi		PDF Dosya Tipi		DOC Dosya Tipi	
Yöntem	#1	#2	#1	#2	#1	#2	#1	#2	#1	#2	#1	#2	#1	#2	#1	#2	#1	#2	#1	#2
Almanca	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%
Arnavutça	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%
Fransızca	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%
Galce	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%
Hırvatça	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%
İngilizce	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%
İrlandaca	1%	1%	25%	1%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	1%	0%	0%	0%	0%
İspanyolca	26%	25%	47%	26%	25%	23%	24%	25%	24%	25%	25%	23%	25%	24%	25%	26%	24%	25%	24%	23%
İtalyanca	45%	45%	64%	45%	44%	43%	44%	44%	44%	44%	44%	42%	44%	44%	44%	45%	44%	44%	44%	43%
Letonca	60%	60%	77%	60%	60%	59%	59%	60%	59%	60%	60%	58%	60%	59%	60%	60%	59%	60%	59%	58%
Macarca	73%	73%	88%	73%	72%	71%	72%	73%	72%	72%	72%	71%	72%	72%	72%	73%	72%	72%	72%	71%
Maltaca	84%	83%	98%	84%	83%	82%	83%	83%	83%	83%	83%	82%	83%	83%	83%	84%	83%	83%	83%	82%
Portekizce	93%	92%	94%	92%	92%	91%	92%	92%	92%	92%	92%	91%	92%	92%	92%	93%	92%	92%	92%	91%
Türkçe	99%	99%	87%	99%	99%	99%	99%	99%	99%	99%	99%	99%	99%	99%	99%	99%	99%	99%	99%	99%
Vietnamca	92%	92%	81%	92%	93%	93%	93%	93%	93%	93%	93%	94%	93%	93%	93%	92%	93%	93%	93%	94%

EK-40 Vietnamca dili GENIUS test sonuçları

Dil	Vietnamca																			
Format	PDF Dosya Tipi		DOC Dosya Tipi		PDF Dosya Tipi		DOC Dosya Tipi		PDF Dosya Tipi		DOC Dosya Tipi		PDF Dosya Tipi		DOC Dosya Tipi		PDF Dosya Tipi		DOC Dosya Tipi	
Yöntem	#1	#2	#1	#2	#1	#2	#1	#2	#1	#2	#1	#2	#1	#2	#1	#2	#1	#2	#1	#2
Almanca	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%
Arnavutça	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%
Fransızca	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%
Galce	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%
Hırvatça	0%	0%	0%	0%	0%	0%	0%	21%	0%	0%	0%	0%	0%	0%	2%	0%	0%	0%	0%	0%
İngilizce	0%	0%	0%	0%	0%	0%	0%	51%	0%	0%	0%	0%	0%	0%	35%	0%	0%	0%	0%	0%
İrlandaca	0%	0%	0%	0%	0%	0%	0%	72%	0%	0%	0%	0%	0%	0%	59%	18%	0%	0%	0%	9%
İspanyolca	12%	17%	13%	10%	16%	11%	14%	88%	16%	17%	13%	12%	12%	21%	76%	41%	12%	12%	16%	33%
İtalyanca	33%	38%	34%	31%	36%	32%	35%	98%	36%	37%	34%	33%	33%	41%	90%	58%	33%	33%	37%	51%
Letonca	49%	54%	50%	48%	52%	49%	51%	89%	52%	54%	51%	50%	50%	56%	98%	72%	50%	49%	53%	66%
Macarca	63%	67%	64%	61%	66%	62%	65%	80%	66%	67%	64%	63%	63%	69%	89%	84%	63%	63%	66%	78%
Maltaca	74%	78%	75%	73%	77%	74%	76%	74%	77%	78%	75%	75%	75%	80%	82%	94%	75%	74%	77%	88%
Portekizce	84%	87%	85%	83%	86%	83%	86%	68%	86%	87%	85%	84%	84%	89%	75%	97%	84%	84%	87%	97%
Türkçe	92%	95%	93%	91%	94%	92%	94%	63%	94%	95%	93%	92%	92%	97%	70%	90%	93%	92%	95%	95%
Vietnamca	99%	97%	99%	98%	98%	99%	98%	59%	98%	97%	99%	99%	99%	95%	65%	84%	99%	99%	97%	88%

EK-41 GENIUS platformu Microsoft Word ortalama başarı sonuçları

Format	DOC Dosya Tipi											
Örnek	#1		#2		#3		#4		#5		Başarı Ortalaması	
Yöntem	#1	#2	#1	#2	#1	#2	#1	#2	#1	#2	Tanıma	Sonuç
Almanca	92%	83%	98%	83%	69%	98%	14%	93%	73%	12%	52%	100%
Arnavutça	62%	90%	93%	99%	99%	99%	98%	99%	0%	99%	84%	100%
Fransızca	98%	89%	87%	83%	88%	95%	76%	75%	97%	95%	88%	100%
Galce	82%	96%	99%	96%	99%	96%	96%	89%	22%	84%	86%	100%
Hırvatça	97%	94%	97%	93%	98%	99%	84%	99%	95%	89%	95%	100%
İngilizce	98%	99%	95%	97%	95%	99%	86%	94%	95%	99%	96%	100%
İrlandaca	98%	99%	98%	96%	99%	99%	99%	99%	97%	97%	98%	100%
İspanyolca	99%	97%	98%	92%	99%	98%	94%	99%	96%	98%	97%	100%
İtalyanca	96%	89%	73%	99%	94%	99%	93%	98%	94%	99%	93%	100%
Letonca	95%	99%	97%	99%	84%	99%	94%	99%	96%	99%	96%	100%
Macarca	96%	97%	97%	96%	92%	94%	99%	98%	98%	98%	97%	100%
Maltaca	91%	99%	83%	99%	99%	99%	98%	99%	99%	97%	96%	100%
Portekizce	95%	99%	61%	99%	97%	98%	90%	98%	92%	98%	93%	100%
Türkçe	87%	99%	99%	99%	99%	99%	99%	99%	99%	99%	98%	100%
Vietnamca	99%	98%	98%	59%	99%	99%	65%	84%	97%	88%	89%	100%

EK-42 GENIUS platformu Adobe PDF ortalama başarı sonuçları

Format	PDF Dosya Tipi											
Örnek	#1		#2		#3		#4		#5		Başarı Ortalaması	
Yöntem	#1	#2	#1	#2	#1	#2	#1	#2	#1	#2	Tanıma	Sonuç
Almanca	77%	91%	96%	93%	96%	95%	96%	77%	97%	96%	72%	100%
Arnavutça	99%	99%	91%	74%	100%	100%	99%	99%	88%	52%	90%	100%
Fransızca	87%	92%	0%	72%	97%	68%	68%	99%	68%	85%	74%	100%
Galce	66%	95%	100%	100%	100%	100%	97%	94%	97%	98%	95%	100%
Hırvatça	100%	100%	97%	99%	98%	94%	35%	98%	100%	100%	92%	100%
İngilizce	99%	97%	87%	98%	96%	91%	95%	97%	96%	70%	93%	100%
İrlandaca	87%	96%	99%	87%	99%	97%	78%	98%	95%	99%	94%	100%
İspanyolca	94%	88%	95%	99%	79%	95%	74%	98%	87%	88%	90%	100%
İtalyanca	98%	99%	79%	99%	76%	97%	93%	98%	85%	99%	92%	100%
Letonca	95%	89%	90%	81%	94%	99%	83%	99%	96%	99%	93%	100%
Macarca	99%	97%	99%	98%	94%	99%	94%	98%	97%	99%	97%	100%
Maltaca	100%	100%	90%	98%	95%	98%	88%	98%	92%	99%	96%	100%
Portekizce	91%	99%	95%	98%	94%	99%	89%	98%	95%	99%	96%	100%
Türkçe	99%	99%	99%	99%	99%	99%	99%	99%	99%	99%	99%	100%
Vietnamca	99%	97%	98%	99%	98%	97%	99%	95%	99%	99%	98%	100%

EK-43 GENIUS platformu web sitesi HTML ortalama başarı sonuçları

Format	Web Sitesi		Başarı Ortalaması	
Yöntem	#1	#2	Tanıma	Sonuç
Almanca	93%	92%	93%	100%
Arnavutça	99%	99%	99%	100%
Fransızca	91%	75%	83%	100%
Galce	95%	93%	94%	100%
Hırvatça	55%	95%	75%	100%
İngilizce	99%	86%	93%	100%
İrlandaca	99%	96%	98%	100%
İspanyolca	97%	98%	98%	100%
İtalyanca	98%	97%	98%	100%
Letonca	95%	99%	97%	100%
Macarca	95%	97%	96%	100%
Maltaca	99%	99%	99%	100%
Portekizce	99%	96%	98%	100%
Türkçe	98%	99%	99%	100%
Vietnamca	98%	99%	99%	100%

ÖZGEÇMİŞ

Kişisel Bilgiler

Soyadı, Adı : YAVANOĞLU, Uraz
 Uyuğu : T.C.
 Doğum tarihi ve yeri : 09.06.1982, Ankara
 Medeni hali : Bekar
 Telefon : 0 (533) 420 94 69
 E-mail : urazyavanoglu@gmail.com

Eğitim

Derece	Eğitim Birimi	Mezuniyet tarihi
Yüksek lisans	Gazi Üniversitesi / Bilgisayar Mühendisliği	2009
Lisans	Atılım Üniversitesi / Bilgisayar Mühendisliği	2005
Lise	Ankara Ayrancı Lisesi	1999

İş Deneyimi

Yıl	Yer	Görev
2007-2009	Gazi Üniversitesi	Araştırma Görevlisi

Yabancı Dil

İngilizce

Yayınlar

1. Sağiroğlu Ş., Yavanoğlu U., Güven E.N., “Web Based Machine Learning for Language Identification and Translation” *International Conference on Machine Learning and Applications*, Ohio, 280-285 (2007).
2. Sağiroğlu Ş., Yavanoğlu U., “Predicting Languages of Web documents with Computers”, *International Scientific Conference on High Technology in Educational System Advancement and Current trends in Teaching Methodology*, Asgabad, (2008).