

Uygulamalı Türkçe Doğal Dil İşleme Evreleri

GÜLŞEN ERYİĞİT

İstanbul Teknik Üniversitesi, Bilgisayar ve Bilişim Fakültesi



Akademisyen kısa özgeçmiş

İstanbul Teknik Üniversitesi (İTÜ) Bilgisayar ve Bilişim Fakültesi öğretim üyesi olan Dr. Gülşen Eryiğit İTÜ Doğal Dil İşleme Grubunun kurucu üyesi ve yöneticisidir. Yüksek lisans ve doktora derecelerini 2002 ve 2007 yıllarında İTÜ Bilgisayar Mühendisliği Bölümünden almıştır. 2006 yılında İsveç Vaxjö Üniversitesinde ziyaretçi araştırmacı olarak bulunmuş ve çalıştığı grup ile CoNLL (doğal dil öğrenme konferans serisi) 2006 ve 2007 çok-dilli bağıllık ayrıştırması çalıştaylarında birinci olarak seçilen çalışmada yer almıştır. 2007 yılında doktora çalışmaları nedeni ile Siemens Mükemmeliyet Ödülüne laik görülmüştür.

Gülşen Eryiğit'in aktif araştırma alanı doğal dil işleme konusundadır. Bu alanda, pek çok prestijli dergi ve konferansta hakem ve yazar olarak görev almıştır. Avrupa Birliği CLARIN (EU 7th Framework Programme, CLARIN - Common language resources and technology infrastructure) projesinde Türkiye temsilcisi olarak yer almıştır. Bunun yanı sıra yakın geçmişte pek çok AB (Cost), Tubitak (1001 ve 1003) ve Sanayi Bakanlığı (Santez) destekli projede yürütücülük ve araştırmacılık yapmış, 6 Tubitak TEYDEB destekli projede danışman olarak görev almıştır. Aynı zamanda bu fonlama kuruluşlarında (AB H2020, Cost Aksiyonu programları, Tubitak, Sanayi Bakanlığı) hem araştırma hem endüstriyel proje önerilerinde hakemlik ve izleyicilik yapmaktadır. Araştırmacıların hizmetine sunduğu Türkçe Doğal Dil İşleme Web Servisleri (tools.nlp.itu.edu.tr) 2018 yılı itibari ile 190'nın üzerinde araştırmacıya hizmet vermektedir.

Uygulamalı Türkçe Doğal Dil İşleme Evreleri

GÜLŞEN ERYİĞİT

İstanbul Teknik Üniversitesi, Bilgisayar ve Bilişim Fakültesi





ITU Turkish NLP Web Service



Gülşen Eryiğit [gulsen.cebiroglu@itu.edu.tr]

Department of Computer Engineering, Istanbul Technical University, Istanbul, Turkey

Project Website : <http://tools.nlp.itu.edu.tr/>

ITU NLP Web Service in a nutshell

The platform provides current state of the art Turkish NLP modules which are interoperable.

Motivation:

comes from the following real word problems:

- 1- Need to provide assistance for the installation and the usage of different tools,
- 2- Difficulty to share the updates and the new modules introduced into the pipeline,
- 3- Difficulty of using the tools for educational purposes within the classrooms and term projects.

ITU NLP Web Service:

- Operates as a SaaS (Software as a Service) ,
- Provides the researchers and the students the state of the art NLP tools in many layers:
- Preprocessing, morphology, syntax and entity recognition.

The users may communicate with the platform via three channels:

- 1- Via a user friendly web interface,
- 2- By file uploads,
- 3- By using the provided Web APIs within their own codes for constructing higher level applications.

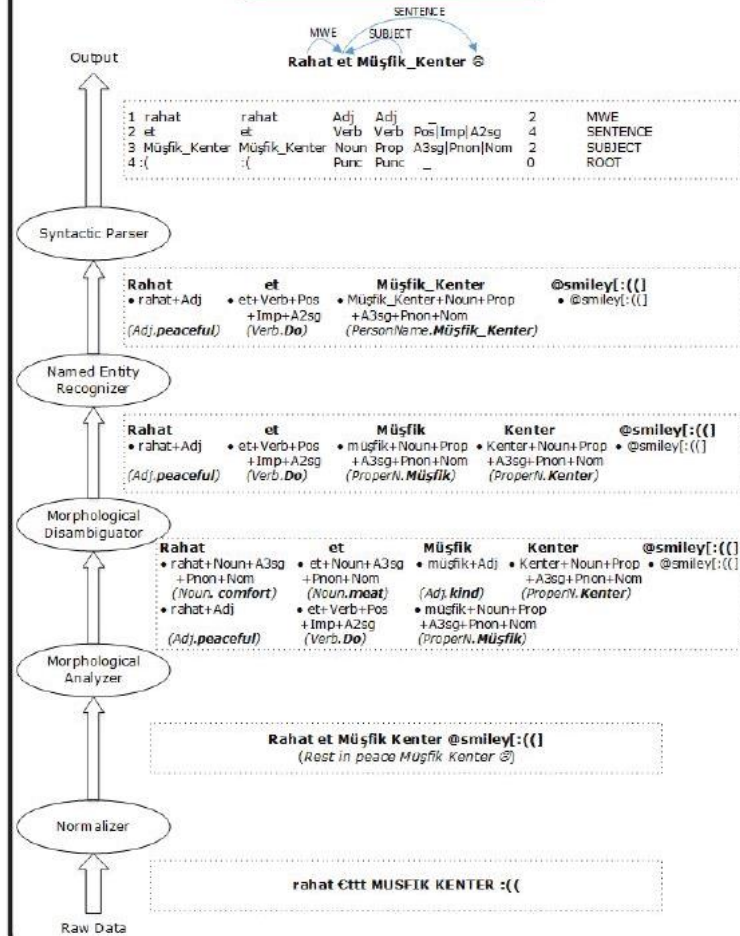
NLP Web Service Components

- Tokenizer
- Deasciifier
- Vowelizer
- Spelling Corrector
- Normalizer
- isTurkish
- Morphological Analyzer
- Morphological Disambiguator
- Named Entity Recognizer
- Dependency Parser
- Whole Pipeline

NLP Web Service Layers

- **Preprocessing**
 - Tokenizer
 - Deasciifier
 - Vowelizer
 - Spelling Corrector
 - Normalizer
- **Morphological Processing**
 - Morphological Analyzer
 - Morphological Disambiguator
- **Multiword Expression Handling**
 - Named Entity Recognizer
- **Syntactic Processing**
 - Dependency Parsing

ITU Turkish NLP Pipeline



- 2014'den bu yana
- >190 kayıtlı aktif araştırmacı (y.lisans, doktora, araştırma projesi çalışanı)



I would like to thank to my students; Thomas Joole, Dilara Torunoğlu, Umur Sulubacak and Hasan Kaya without whose it would be impossible to produce the ITU Turkish NLP pipeline. This work is part of a research project supported by TUBITAK 1001 (Grant number: 112E276) as an ICT cost action (ICT207) project.



Sunum İçeriği

- *Türkçe Doğal Dil İşleme Yazılım Zinciri Evreleri*
 - *Normalizasyon*
 - *Sözcük Analizi*
 - *Varlık İsmi Tanıma*
 - *Cümle Analizi*
 - *Veri Kümeleri*
- *Örnek Projeler*
- *Aktif Araştırma Evreleri*
 - *Anlamsal Rol Etiketleme*
 - *Eşgönderge Çözümlemesi*
 - *Çok sözcüklü terimlerin Bulunması*

Sunum İçeriği

- *Türkçe Doğal Dil İşleme*
 - Normalizasyon
 - Sözcük Analizi
 - Varlık İsmi Tanıma
 - Cümle Analizi
 - Veri Kümeleri
- Örnek Projeler
- Aktif Araştırma Evreleri
 - Anlamsal Rol Etiketleme
 - Eşgönderge Çözümlemesi
 - Çok sözcüklü terimlerin Bulunması

Kullanılan Yapay Zeka Yöntemleri

- Sınıflandırıcılar
- Dizilim Etiketleyicileri
- Kural Kodlayıcıları

- ✓ **CRF** (conditional random fields, koşullu rasgele alanlar)
- ✓ **FST** (Finite State Transducers – Sonlu Durumlu Otomatlar)
- ✓ **SVM** (Support Vektor Machines – Karar Destek Makineleri)
- ✓ **Yapay Sinir Ağları**
- ✓ **Derin Öğrenme - LSTM** (Long-Short Term Memory)

Sunum İçeriği

- *Türkçe Doğal Dil İşleme*
 - ***Normalizasyon***
 - *Sözcük Analizi*
 - *Varlık İsmi Tanıma*
 - *Cümle Analizi*
 - *Veri Kümeleri*
- *Örnek Projeler*
- *Aktif Araştırma Evreleri*
 - *Anlamsal Rol Etiketleme*
 - *Eşgönderge Çözümlemesi*
 - *Çok sözcüklü terimlerin Bulunması*



rahat €ttt MUSFIK KENTER :((



Rahat et Müşfik Kenter @smiley[:((



Normalizasyon

Gülşen Eryiğit and Dilara Torunoğlu-Selamet. 2017.

Social media text normalization for Turkish.

Natural Language Engineering Journal, 23(6): 835–875.

Sorunlu Sözcük Tespiti

Öneri Üretimi

Harf Yazımı
Dönüşümü

Dönüştürme
Kuralları

Özel İsim
Tespiti

Türkçe
Karakter
Düzeltilici

Sesli
Harf
Üreteci

Şive
Düzeltilici

Yazım Hatası
Düzeltilici

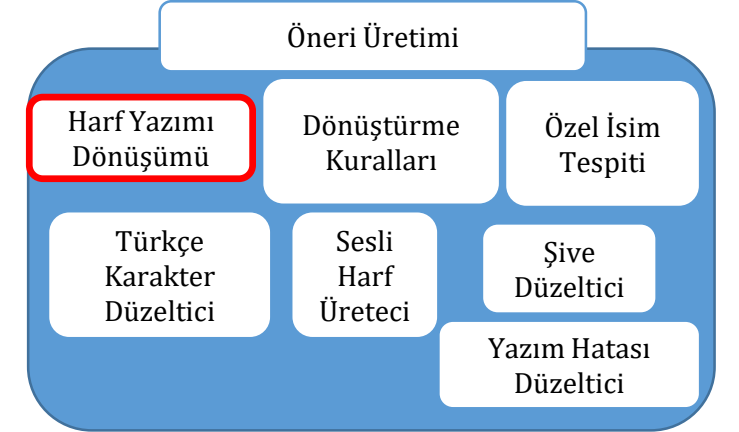
Normalizasyon

tümüküçük
TÜMÜBÜYÜK
Özel İsim Yazımı
karIŞIK YaZıM

Gülşen Eryiğit and Dilara Torunoğlu-Selamet. 2017.

Social media text normalization for Turkish.

Natural Language Engineering Journal, 23(6): 835–875.

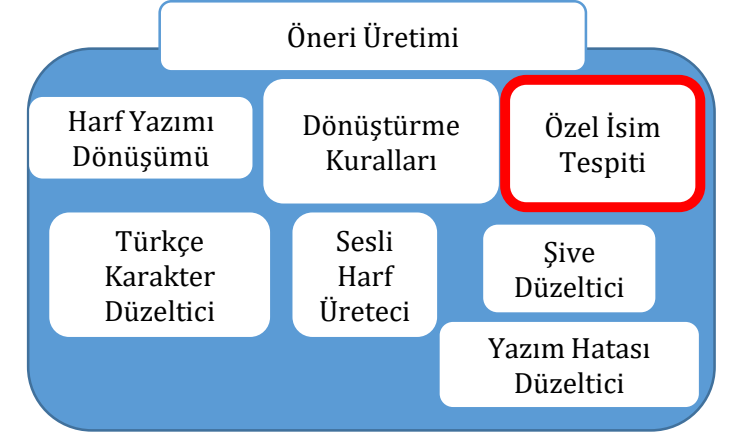


Normalizasyon

Gülşen Eryiğit and Dilara Torunoğlu-Selamet. 2017.

Social media text normalization for Turkish.

Natural Language Engineering Journal, 23(6): 835–875.



umuttan → Umut'tan ?

meltem yanık konuşmacılarımızdan biri.

ayşenden → Ayşe'nden , Ayşen'den ?

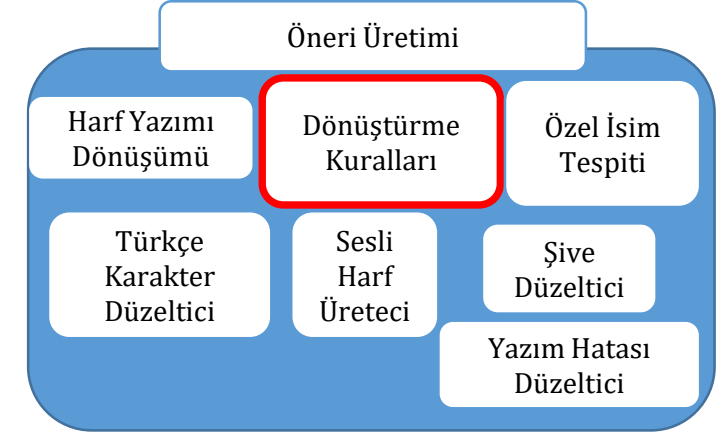
Normalizasyon

Karakter tekrarları,
Logogramlar (\$ → ,s, € → e, 3 → e, @ → a, ! → i, and ß → b),
Web 2.0'a özgü sözcükler,
Eposta, hashtag, mention ve URL'leri yakalamaya yönelik

Gülşen Eryiğit and Dilara Torunoğlu-Selamet. 2017.

Social media text normalization for Turkish.

Natural Language Engineering Journal, 23(6): 835–875.



@dida → *@mention[@dida]*
nbr → *ne* *haber*
neden → *neden*
aramion → *aramion*
#kirildim → *@hashtag[#kirildim]*
: (→ *@smiley[: (]*

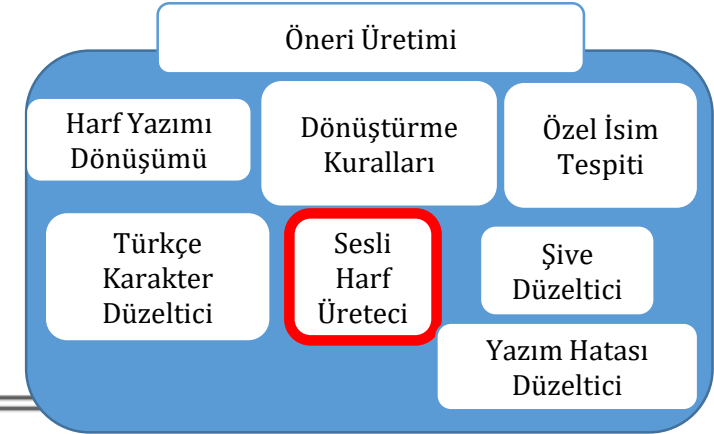
Normalizasyon

Kübra Adalı and Gülşen Eryiğit. 2014.

Vowel and diacritic restoration for social media texts.

In 5th Workshop on Language Analysis for Social Media (LASM) at EACL.

_k_l_d_n → okuldan



Curr. Word	Neigh. Ch(-3)	Neigh. Ch(-2)	Neigh. Ch(-1)	Neigh. Ch(+1)	Neigh. Ch(+2)	Neigh. Ch(+3)	Class Label
---------------	------------------	------------------	------------------	------------------	------------------	------------------	----------------

kldn	-	-	-	k	l	d	o
kldn	-	-	k	l	d	-	u
kldn	-	k	l	d	n	-	null
kldn	k	l	d	n	-	-	a
kldn	l	d	n	-	-	-	null

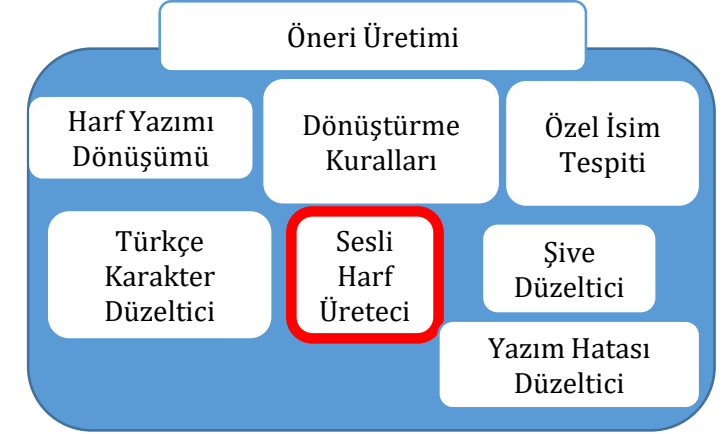
Normalizasyon

- *Dizilim etiketleyicileri*
- *CRF – Koşullu Rasgele Alanlar*
- *Kod çözme aşamasında*
Kısıtlanmış Viterbi algoritması

Gülşen Eryiğit and Dilara Torunoğlu-Selamet. 2017.

Social media text normalization for Turkish.

Natural Language Engineering Journal, 23(6): 835–875.

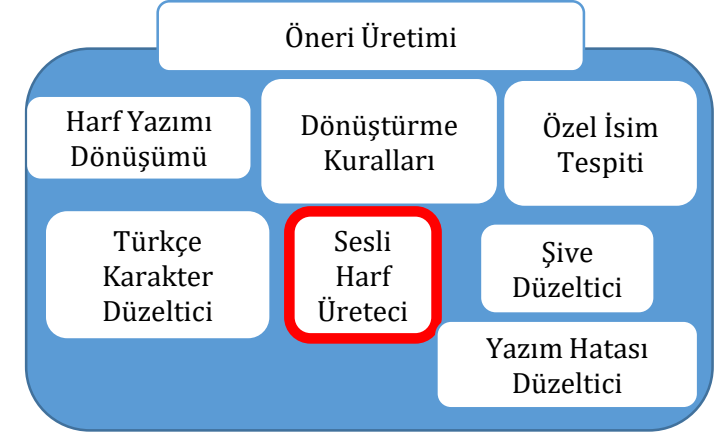
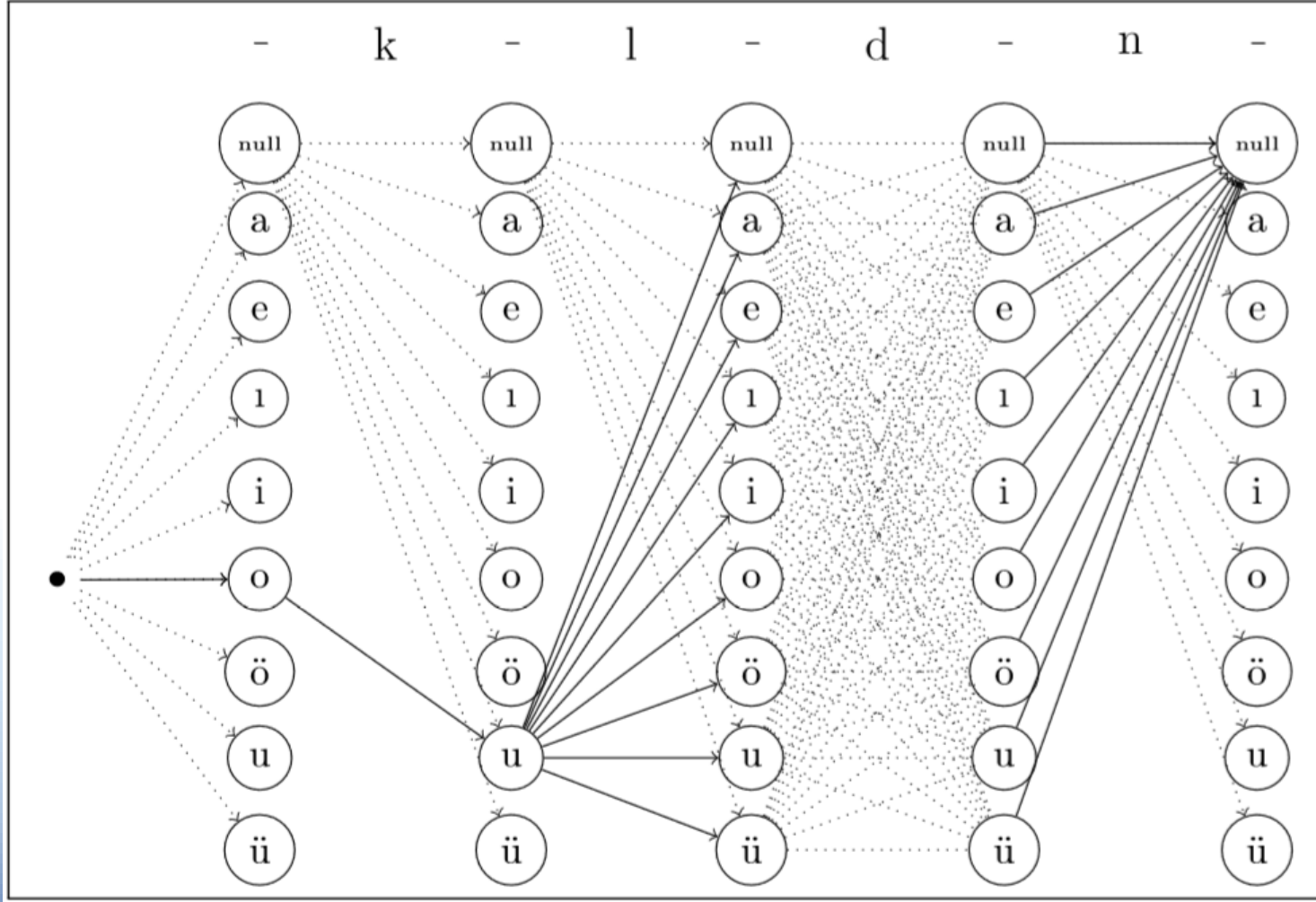


Normalizasyon

Gülşen Eryiğit and Dilara Torunoğlu-Selamet. 2017.

Social media text normalization for Turkish.

Natural Language Engineering Journal, 23(6): 835–875.

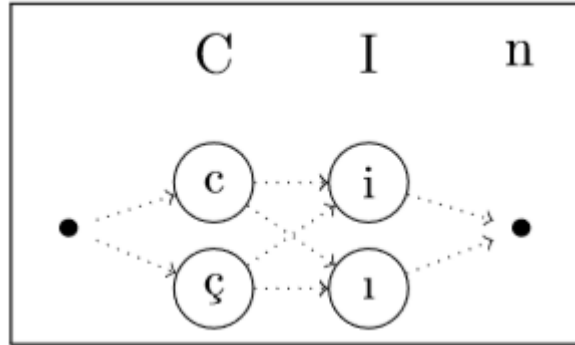


Normalizasyon

Kübra Adalı and Gülşen Eryiğit. 2014.

Vowel and diacritic restoration for social media texts.

In 5th Workshop on Language Analysis for Social Media (LASM) at EACL.



Input	Output 1	Output 2	Output 3	Output 4
<i>cin</i>	<i>cin</i> ('genie')	<i>çin</i> ('China')	<i>çın</i> ('ding')	<i>cın</i> (OOV)
<i>kus</i>	<i>kus</i> ('vomit')	<i>kuş</i> ('bird')	<i>küs</i> ('sulk at')	<i>küş</i> (OOV)
<i>sok</i>	<i>sok</i> ('insert')	<i>şok</i> ('shock')	<i>sök</i> ('disassemble')	<i>şök</i> (OOV)

Normalizasyon

Gülşen Eryiğit and Dilara Torunoğlu-Selamet. 2017.

Social media text normalization for Turkish.

Natural Language Engineering Journal, 23(6): 835–875.

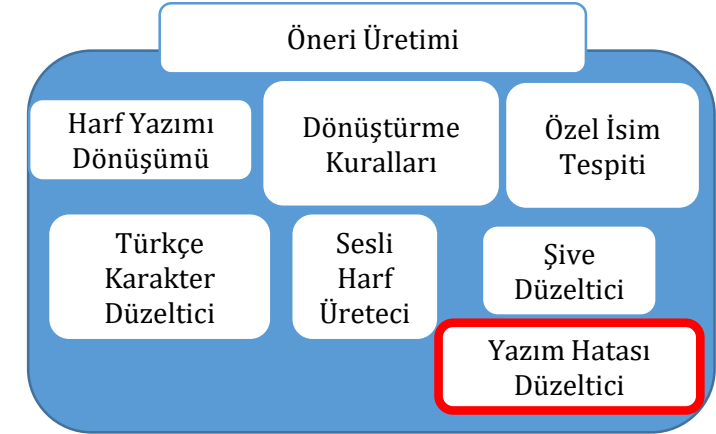
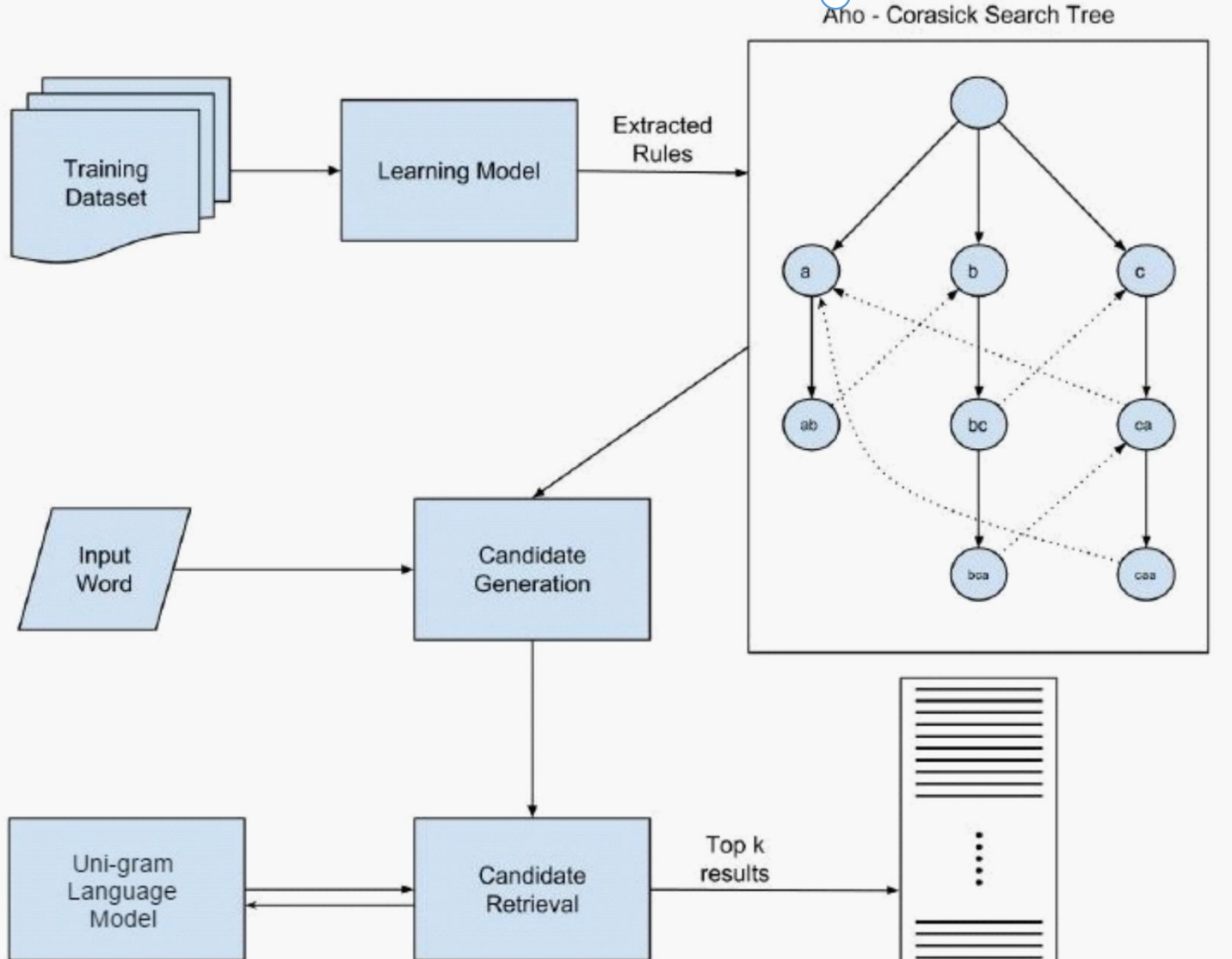
gidmiycem → gitmeyeceğim



yapacağım-yapacaksın-yapacak-yapacağız-yapacaksınız-yapacaklar
(1stsg) (2ndsg) (3rdsg) (1stpl) (2ndpl) (3rdpl)

yapcam - yapcan - yapcak - yapcaz - yapcanız - yapcaklar
yapçam - yapçan - yapçak - yapçaz - yapçaksınız - yapçaklar
yapıcam - yapıcan - yapıcak - yapıcaz - yapıcaksınız - yapıcaklar
yapacam - yapacan - yapacak - yapacaz - yapacanız - yapacaklar

Normalizasyon



Normalizasyon

Gülşen Eryiğit and Dilara Torunoğlu-Selamet. 2017.

Social media text normalization for Turkish.

Natural Language Engineering Journal, 23(6): 835–875.

Sorunlu Sözcük Tespiti

Öneri Üretimi

Harf Yazımı
Dönüşümü

Dönüştürme
Kuralları

Özel İsim
Tespiti

Türkçe
Karakter
Düzeltilici

Sesli
Harf
Üreteci

Şive
Düzeltilici

Yazım Hatası
Düzeltilici

ITU Turkish Natural Language Processing Pipeline

ITU Türkçe Doğal Dil İşleme Yazılım Zinciri

Login

Normalization

- Tokenizer
- Normalization
- Deasciifier
- Vowelizer
- Spelling Corrector
- isTurkish
- Morphological Analyzer
- Morphological Disambiguator
- Named Entity Recognizer
- Dependency Parsing
- Whole pipeline
- UD Mapper (Beta)
- Datasets
- Simple API
- Changelog

@dida bi yere gitcem
"büttttttüüüünnnn" insanlar hür,
hysyt ve haklari bkmndn eşit
doğarlaaaar.
aaaaaakıl ve vicdana sahiptirler ve
birbirlerine karşı kardeslik zhnyt
ile hrkt etmelidirler ." diye
bağırcaam :)) #insanhaklari
www.insanhaklari.com

Normalize

@mention[@dida] bir yere
gideceğim "bütün" insanlar hür ,
haysiyet ve hakları bakımından
eşit doğarlar . Akıl ve vicdana
sahiptirler ve birbirlerine karşı
kardeşlik zihniyet ile hareket
etmelidirler ." diye bağıracağım
@smiley[:))]
@hashtag[#insanhaklari]
@url[www.insanhaklari.com]

The details of this work are given in the following publications please refer to them if you w

Report a
bug

Sunum İçeriği

- *Türkçe Doğal Dil İşleme Yazılım Zinciri Evreleri*
 - *Normalizasyon*
 - ***Sözcük Analizi***
 - *Varlık İsmi Tanıma*
 - *Cümle Analizi*
 - *Veri Kümeleri*
- *Örnek Projeler*
- *Aktif Araştırma Evreleri*
 - *Anlamsal Rol Etiketleme*
 - *Eşgönderge Çözümlemesi*
 - *Çok sözcüklü terimlerin Bulunması*

Rahat et Müşfik Kenter

Sözcük Analizi

Sıfat

rahat+Adj

2. tekil kişi emir kipinde
çekilmiş olumlu eylem
et+Verb+Pos+Imp+A2sg

Özel İsim

Müşfik+Noun+Prop+A3sg
+Pnon+Nom

Özel İsim

Kenter+Noun+Prop+A3sg
+Pnon+Nom

Rahat et Müşfik Kenter

Gülşen Eryiğit, 2017.

ITU Turkish NLP Web Service.

In Proc. of the Demonstrations at EACL 2014.

Göteborg, Sweden,

Sözcük Analizi

İyelik eki almamış yalın halde
3. tekil kişi isim

rahat+Noun+A3sg+Pnon+Nom

Sıfat

rahat+Adj

İyelik eki almamış yalın halde
3. tekil kişi isim

et+Noun+A3sg+Pnon+Nom

Sıfat

müşfik+Adj

Özel İsim

Kenter+Noun+Prop+A3sg
+Pnon+Nom

2. tekil kişi emir kipinde
çekilmiş olumlu eylem

et+Verb+Pos+Imp+A2sg

Özel İsim

Müşfik+Noun+Prop+A3sg
+Pnon+Nom

Biçimbilimsel Çözümleme

Türkçe Doğal Dil İşleme Evreleri © 2018 by Gülşen Eryiğit

Rahat et Müşfik Kenter

Gülşen Eryiğit, 2017.

ITU Turkish NLP Web Service.

In Proc. of the Demonstrations at EACL 2014.

Gothenburg, Sweden,

Sözcük Analizi

İyelik eki almamış yalın halde
3. tekil kişi isim

rahat+Noun+A3sg+Pnon+Nom

Sıfat

rahat+Adj

İyelik eki almamış yalın halde
3. tekil kişi isim

et+Noun+A3sg+Pnon+Nom

2. tekil kişi emir kipinde
çekilmiş olumlu eylem
et+Verb+Pos+Imp+A2sg

Sıfat

müşfik+Adj

Özel İsim

Kenter+Noun+Prop+A3sg
+Pnon+Nom

Özel İsim

Müşfik+Noun+Prop+A3sg
+Pnon+Nom

Biçimbilimsel Çözümleme

Rahat et Müşfik Kenter

Gülşen Eryiğit, 2017.

ITU Turkish NLP Web Service.

In Proc. of the Demonstrations at EACL 2014.

Gothenburg, Sweden,

Sözcük Analizi

İyelik eki almamış yalın halde
3. tekil kişi isim

rahat+Noun+A3sg+Pnon+Nom

Sıfat

rahat+Adj

İyelik eki almamış yalın halde
3. tekil kişi isim

et+Noun+A3sg+Pnon+Nom

2. tekil kişi emir kipinde
çekilmiş olumlu eylem

et+Verb+Pos+Imp+A2sg

Sıfat

müşfik+Adj

Özel İsim

Kenter+Noun+Prop+A3sg
+Pnon+Nom

Özel İsim

Müşfik+Noun+Prop+A3sg
+Pnon+Nom

Biçimbilimsel Çözümleme

Türkçe Doğal Dil İşleme Evreleri © 2018 by Gülşen Eryiğit

Rahat et Müşfik Kenter

Gülşen Eryiğit, 2017.

ITU Turkish NLP Web Service.

In Proc. of the Demonstrations at EACL 2014.

Gothenburg, Sweden,

Sözcük Analizi

İyelik eki almamış yalın halde
3. tekil kişi isim

rahat+Noun+A3sg+Pnon+Nom

Sıfat

rahat+Adj

İyelik eki almamış yalın halde
3. tekil kişi isim

et+Noun+A3sg+Pnon+Nom

Sıfat

müşfik+Adj

Özel İsim

Kenter+Noun+Prop+A3sg
+Pnon+Nom

2. tekil kişi emir kipinde
çekilmiş olumlu eylem

et+Verb+Pos+Imp+A2sg

Özel İsim

Müşfik+Noun+Prop+A3sg
+Pnon+Nom

Biçimbilimsel Çözümleme

Türkçe Doğal Dil İşleme Evreleri © 2018 by Gülşen Eryiğit

Rahat et Müşfik Kenter

Gülşen Eryiğit, 2017.

ITU Turkish NLP Web Service.

In Proc. of the Demonstrations at EACL 2014.
Gothenburg, Sweden,

HFST & XFST uyumlu
1.5M sözcük
22saniye

Sözcük
Analizi

İyelik eki almamış yalın halde
3. tekil kişi isim

rahat+Noun+A3sg+Pnon+Nom

Sıfat

rahat+Adj

İyelik eki almamış yalın halde
3. tekil kişi isim

et+Noun+A3sg+Pnon+Nom

Sıfat

müşfik+Adj

Özel İsim

Kenter+Noun+Prop+A3sg
+Pnon+Nom

2. tekil kişi emir kipinde
çekilmiş olumlu eylem

et+Verb+Pos+Imp+A2sg

Özel İsim

Müşfik+Noun+Prop+A3sg
+Pnon+Nom

Biçimbilimsel Çözümleme

Türkçe Doğal Dil İşleme Evreleri © 2018 by Gülşen Eryiğit



ITU Turkish Natural Language Processing Pipeline

ITU Türkçe Doğal Dil İşleme Yazılım Zinciri

[Login](#)

Morphological Analyzer

Tokenizer
Normalization
Deasciifier
Vowelizer
Spelling Corrector
isTurkish
Morphological Analyzer
Morphological Disambiguator
Named Entity Recognizer
Dependency Parsing
Whole pipeline
UD Mapper (Beta)
Datasets
Simple API
Changelog

Obama 'dan

serverlarda

#Paris'den

:)

[Analyse](#)

Obama 'dan
Obama+Guess+Noun+Prop+A3sg+Pnon+Ab1

serverlarda
server+Guess+Noun+A3pl+Pnon+Loc
serverlar+Guess+Noun+A3sg+Pnon+Loc
serverlard+Guess+Noun+A3sg+Pnon+Dat
serverlarda+Guess+Noun+A3sg+Pnon+Nom

#Paris'den
#Paris+Guess+Noun+Prop+A3sg+Pnon+Ab1

:) :) +Punc

[Report a bug](#)

Biçimbilimsel Çözümleyici ne amaçlarla kullanılıyor?

- *Sözcük Çözümleme*

GET or POST parameters: {tool}, {input} and {token}

token: you can find your token on your login space

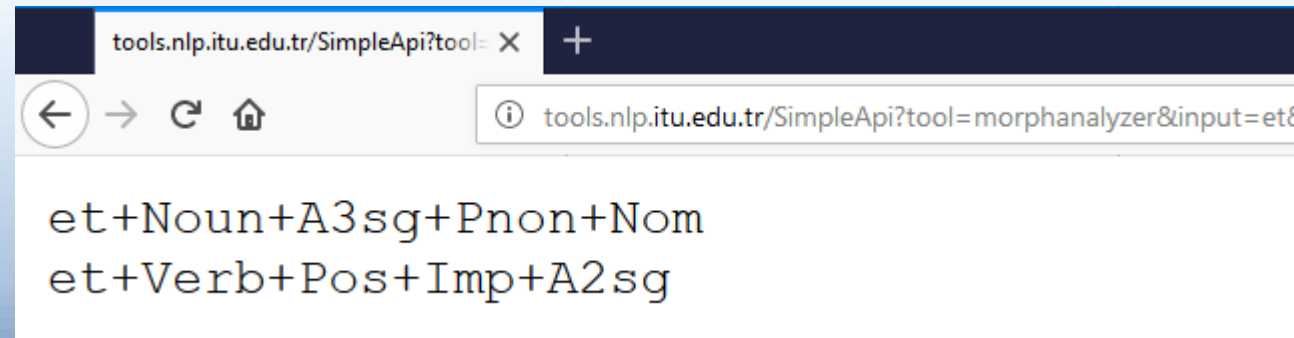
tool: ner, **morphalyzer**, isturkish, morphgeerator, tokenizer, normalize, deasciifier, Vowelizer,

DepParserFormal, DepParserNoisy, spellcheck, disambiguator, pipelineFormal, pipelineNoisy,

input: utf-8 string The response is a text/plain encoded in UTF-8

Example: [http://tools.nlp.itu.edu.tr/SimpleApi?tool=](http://tools.nlp.itu.edu.tr/SimpleApi?tool=morphalyzer&input=et&token=XXXXXXXXXXXXXXXXXXXX)

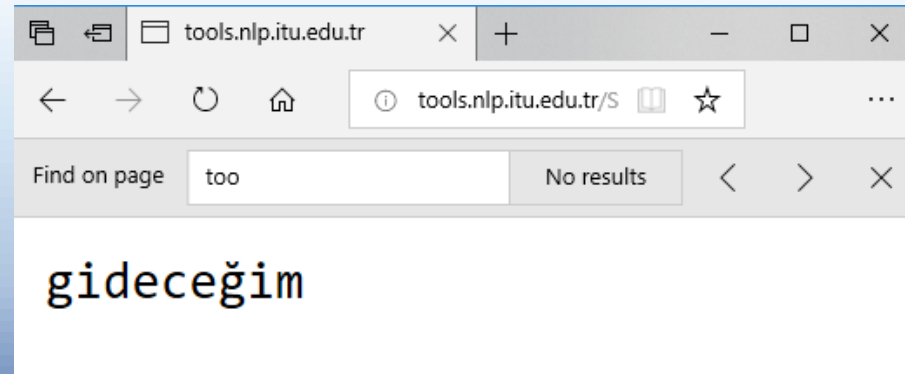
morphalyzer&input=**et**&token=XXXXXXXXXXXXXXXXXXXX
XXXXXXXXXXXX



Biçimbilimsel Çözümleyici ne amaçlarla kullanılıyor?

- *Sözcük Çözümleme*
- *Sözcük Üretimi*

GET or POST parameters: {tool}, {input} and {token}
token: you can find your token on your login space
tool: ner, morphanalyzer, isturkish, morphgenerator, tokenizer, normalize, deasciifier, Vowelizer, DepParserFormal, DepParserNoisy, spellcheck, disambiguator, pipelineFormal, pipelineNoisy,
input: utf-8 string The response is a text/plain encoded in UTF-8
Example: [http://tools.nlp.itu.edu.tr/SimpleApi?tool=morphgenerator](http://tools.nlp.itu.edu.tr/SimpleApi?tool=morphgenerator&input=git+Verb+Pos+Fut+A1sg&token=XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX) &input=**git+Verb+Pos+Fut+A1sg** &token=XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX

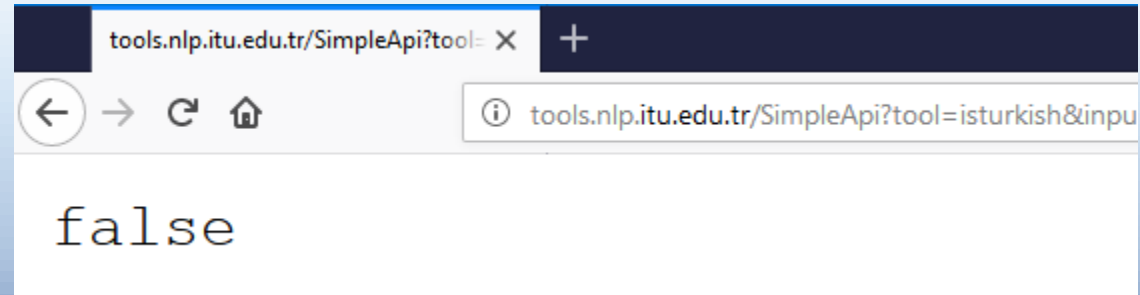


Biçimbilimsel Çözümleyici ne amaçlarla kullanılıyor?

- *Sözcük Çözümleme*
- *Sözcük Üretimi*
- *Sözcük Doğrulama*

GET or POST parameters: {tool}, {input} and {token}
token: you can find your token on your login space
tool: ner, morphanalyzer, isturkish, morphgenerator, tokenizer, normalize, deasciifier, Vowelizer, DepParserFormal, DepParserNoisy, spellcheck, disambiguator, pipelineFormal, pipelineNoisy,
input: utf-8 string The response is a text/plain encoded in UTF-8

Example: [http://tools.nlp.itu.edu.tr/SimpleApi?tool=isturkish](http://tools.nlp.itu.edu.tr/SimpleApi?tool=isturkish&input=giteceğim&token=XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX) &input= **giteceğim** &token=X
XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX



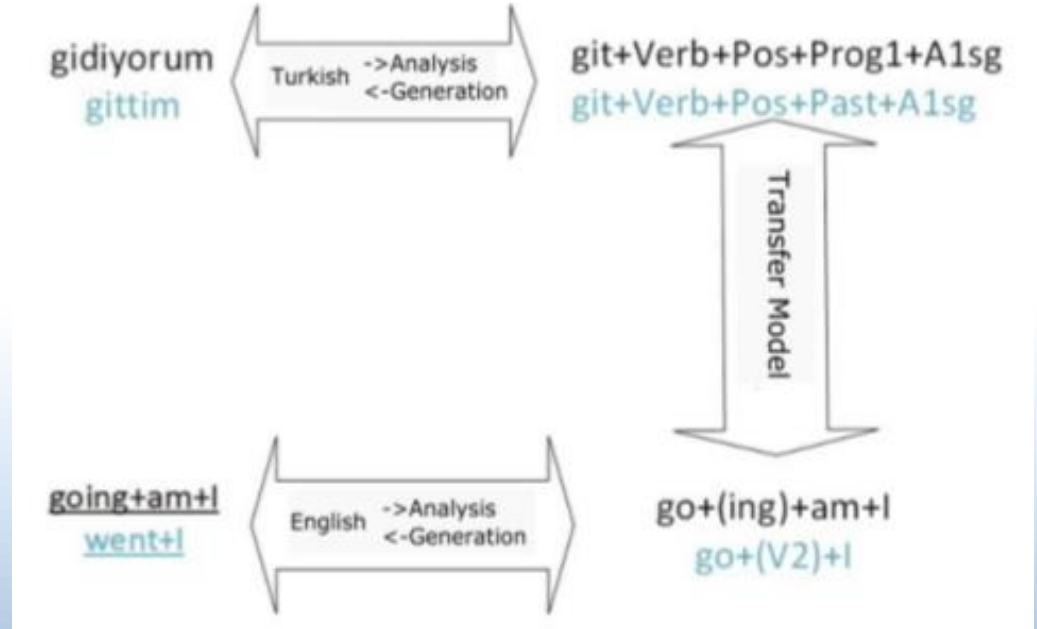
Biçimbilimsel Çözümleyici ne amaçlarla kullanılıyor?

- *Sözcük Çözümleme*
- *Sözcük Üretimi*
- *Sözcük Doğrulama*
- *Sözcük bazında transfere dayalı çeviri*

Hasan Kaya and Gülşen Eryiğit. 2015.

Using finite state transducers for helping foreign language learning.

2nd Workshop on Natural Language Processing Techniques for Educational Applications at ACL, Beijing, China



Biçimbilimsel Çözümleyici ne amaçlarla kullanılıyor?

- Sözcük Çözümleme
- Sözcük Üretimi
- Sözcük Doğrulama
- Sözcük bazında

transfere dayalı çeviri

Turkish	ITUMorphTrans4ET	Google	Bing	Tureng	Yandex
gidebilirse	if he can go ✓	can go	if he can go ✓	go	he could leave
gidecek	he will go ✓	will go	will go	be destined for	go
gitmelilerdi	they should have gone ✓	they should go	they need to get it	go	they have to go
gitmişlerse	if they had gone✓	they have gone	if they went	go	if they're gone
gidebilecekse	if he will be able to go✓	go abilecekse	can go	go	if you can go
yapmalılarsa	if they should do ✓	sign mAllArsA	if they're making	go	do they
masalarımızla	with our tables ✓	our table	our table	table	our desks
English	ITUMorphTrans4ET	Google	Bing	Yandex	
if he can go	gidebilirse✓	O gidebilirsiniz	Eğer gidebilir	eğer gidip o	
he will go	gidecek ✓	O gidecek ✓	o-ecek gitmek	gidecek ✓	
they should have gone	gitmelilerdi✓ gitmelilermiş✓	Onlar gitmiş olmalı	Onlar gitmiş olmalı	gitmelilerdi ✓	
if they had gone	gitmişlerse✓ gitmişseler✓	onlar gitmişti eğer	Onlar ne gitseydin	eğer gitmiş olsalardı	
if he will be able to go	gidebilecekse✓	O gitmek mümkün olacak eğer	Eğer o-ecek var olmak güçlü-e doğru gitmek için	eğer gitmek mümkün olacak	
if they should do	yapmalılarsa✓ etmelilerse✓	Onlar yapmalıyım	onlar yoksa	eğer yapmalıyım eğer	
with our tables	listelerimizle✓ tablolarımızla✓	Bizim tablolarla ✓	Bizim tablolarla ✓	bizim tablolar ile ✓	

Rahat et Müşfik Kenter

Sözcük Analizi

İyelik eki almamış yalın halde
3. tekil kişi isim
rahat+Noun+A3sg+Pnon+Nom

Sıfat
rahat+Adj

İyelik eki almamış yalın halde
3. tekil kişi isim
et+Noun+A3sg+Pnon+Nom

2. tekil kişi emir kipinde
çekilmiş olumlu eylem
et+Verb+Pos+Imp+A2sg

Sıfat
müşfik+Adj

Özel İsim
Müşfik+Noun+Prop+A3sg+Pnon+Nom

Özel İsim
Kenter+Noun+Prop+A3sg+Pnon+Nom

Biçimbilimsel Belirsizlik Giderme



Sıfat
rahat+Adj

2. tekil kişi emir kipinde
çekilmiş olumlu eylem
et+Verb+Pos+Imp+A2sg

Özel İsim
*Müşfik+Noun+Prop+A3sg
+Pnon+Nom*

Özel İsim
*Kenter+Noun+Prop+A3sg
+Pnon+Nom*

Biçimbilimsel Belirsizlik Giderme

Rahat et Müşfik Kenter

Sözcük Analizi

Sıfat

rahat+Adj

2. tekil kişi emir kipinde
çekilmiş olumlu eylem
et+Verb+Pos+Imp+A2sg

Özel İsim

Müşfik+Noun+Prop+A3sg
+Pnon+Nom

Özel İsim

Kenter+Noun+Prop+A3sg
+Pnon+Nom

Gülşen Eryiğit. 2012.

The Impact of Automatic Morphological Analysis &
Disambiguation on Dependency Parsing of Turkish.

LREC

Sunum İçeriği

- *Türkçe Doğal Dil İşleme Yazılım Zinciri Evreleri*
 - *Normalizasyon*
 - *Sözcük Analizi*
 - ***Varlık İsmi Tanıma***
 - *Cümle Analizi*
 - *Veri Kümeleri*
- *Örnek Projeler*
- *Aktif Araştırma Evreleri*
 - *Anlamsal Rol Etiketleme*
 - *Eşgönderge Çözümlemesi*
 - *Çok sözcüklü terimlerin Bulunması*

Varlık İsmi Tanıma

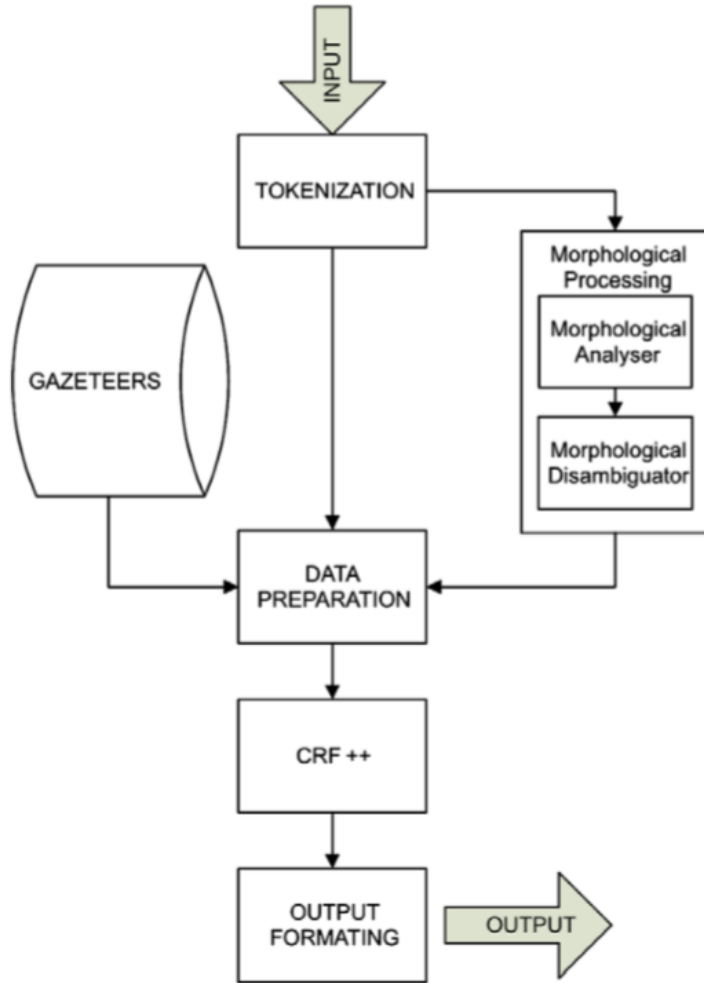


Fig. 1. Proposed Framework.

Gökhan Akın Şeker and Gülşen Eryiğit. 2017.

Extending a CRF-based named entity recognition model for Turkish well formed text and user generated content.

Semantic Web Journal, 8(5):625–642.

Token	IOB2 Tags
Mustafa	B-PERSON
Kemal	I-PERSON
Atatürk	I-PERSON
1919	O
yılında	O
Samsun	B-LOCATION
'a	O
çıktı	O
.	O

Varlık İsmi Tanıma

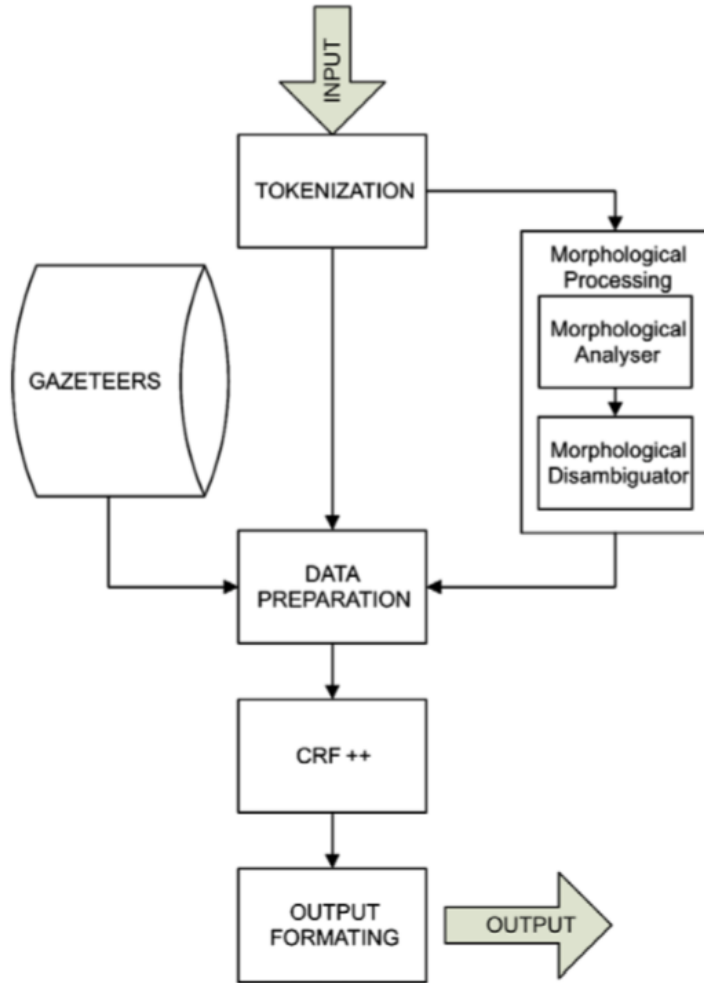


Fig. 1. Proposed Framework.

Gökhan Akın Şeker and Gülşen Eryiğit. 2017.

Extending a CRF-based named entity recognition model for Turkish well formed text and user generated content.

Semantic Web Journal, 8(5):625–642.

Group	Type	News articles [52]	Tweets [2]	IWT [33]
		492K	55K	43K
ENAMEX	Person	15,352	681	380
ENAMEX	Location	10,404	240	260
ENAMEX	Organization	9,571	428	401
TIMEX	Date	1,486	57	59
TIMEX	Time	169	20	9
NUMEX	Money	638	24	45
NUMEX	Percentage	710	5	8
TOTAL		38,330	1,455	1,162

Varlık İsmi Tanıma

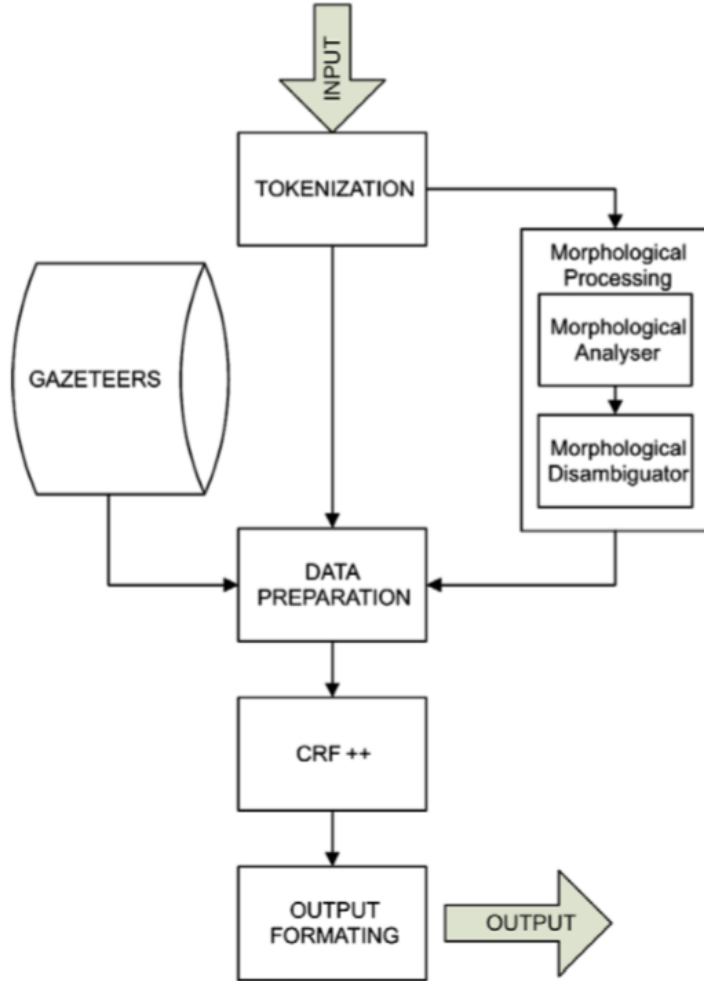


Fig. 1. Proposed Framework.

Gökhan Akın Şeker and Gülşen Eryiğit. 2017.

Extending a CRF-based named entity recognition model for Turkish well formed text and user generated content.

Semantic Web Journal, 8(5):625–642.

Token	IOB2 Tags
Mustafa	B-PERSON
Kemal	I-PERSON
Atatürk	I-PERSON
1919	O
yılında	O
Samsun	B-LOCATION
'a	O
çıktı	O
.	O

- Biçimbilimsel Özellikler
- Sözcüksel Özellikler
- Sözlüklerden Elde Edilen Özellikler

Sunum İçeriği

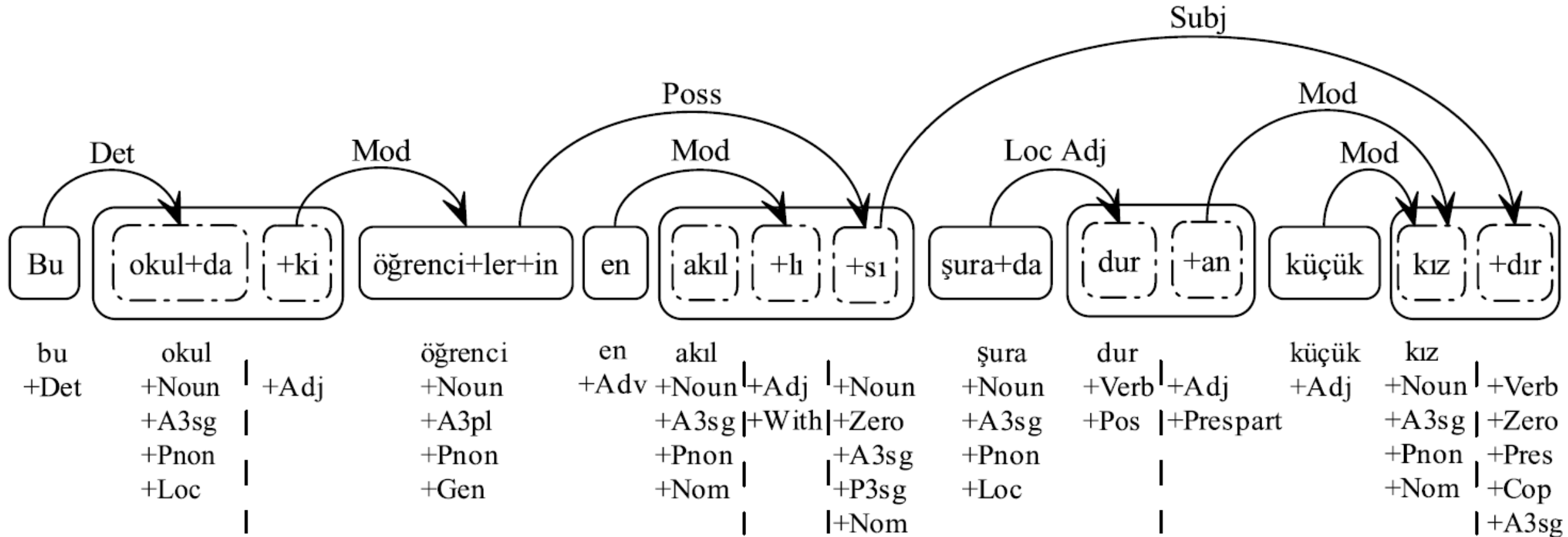
- *Türkçe Doğal Dil İşleme Yazılım Zinciri Evreleri*
 - *Normalizasyon*
 - *Sözcük Analizi*
 - *Varlık İsmi Tanıma*
 - ***Cümle Analizi***
 - *Veri Kümeleri*
- *Örnek Projeler*
- *Aktif Araştırma Evreleri*
 - *Anlamsal Rol Etiketleme*
 - *Eşgönderge Çözümlemesi*
 - *Çok sözcüklü terimlerin Bulunması*

Cümle Analizi

Gülşen Eryiğit, Joakim Nivre and Kemal Oflazer. 2008.

Dependency parsing of Turkish.

Computational Linguistics, 34(3):357–389.



This school-at+that-is student-s-' most intelligence+with+of there stand+ing little girl+is
The most intelligent of the students in this school is the little girl standing there.

Cümle Analizi

Gülşen Eryiğit, Joakim Nivre and Kemal Oflazer. 2008.
Dependency parsing of Turkish.
Computational Linguistics, 34(3):357–389.

ID	LEX	LEMMA	CPOS	POS	INF	HEAD	DEP
1	Bu	bu	Det	Det	–	2	DETERMINER
2	–	okul	Noun	Noun	A3sg Pnon Loc	3	DERIV
3	okuldaki	–	Adj	Adj	Rel	4	MODIFIER
4	öğrencilerin	öğrenci	Noun	Noun	A3pl Pnon Gen	8	POSSESSOR
5	en	en	Adv	Adv	–	7	MODIFIER
6	–	akıl	Noun	Noun	A3sg Pnon Nom	7	DERIV
7	–	–	Adj	Adj	With	8	DERIV
8	akıllısı	–	Noun	Zero	A3sg P3sg Nom	14	SUBJECT
9	şurada	şura	Noun	Noun	A3sg Pnon Loc	10	LOCATIVE.ADJUNCT
10	–	dur	Verb	Verb	Pos	11	DERIV
11	duran	–	Adj	PresPart	–	13	MODIFIER
12	küçük	küçük	Adj	Adj	–	13	MODIFIER
13	–	kız	Noun	Noun	A3sg Pnon Nom	14	DERIV
14	kızdır	–	Verb	Zero	Pres Cop A3sg	15	SENTENCE
15	.	.	Punc	Punc	–	0	ROOT

Geçiş Tabanlı Ayırıştırma



**ötele
indirge
ilişkilendir**



Dependency Parser

Gülşen Eryiğit. 2013-2015.

Türkçe Web 2.0 Cümlelerinin Ayrıştırılması

Tubitak 1001 (EU ICT Cost Aksiyonu) Projesi

Tokenizer

Normalization

Deasciifier

Vowelizer

Spelling Corrector

isTurkish

Morphological Analyzer

Morphological Disambiguator

Named Entity Recognizer

Dependency Parsing

Whole pipeline

UD Mapper (Beta)

Datasets

Simple API

Changelog

☒ Formal language ☐ Noisy language

1	Bütün	bütün	Adj	Adj	—		
2	insanlar		insan	Noun	Noun	A3pl Pnon Nom	
3	hür	hür	Adj	Adj	—		
4	,	,	Punc	Punc	—		
5	haysiyet		haysiyet		Noun	Noun	A3sg Pnon Nom
6	ve	ve	Conj	Conj	—		
7	haklar	hak	Noun	Noun	A3pl Pnon Nom		

Parse

1	Bütün	bütün	Adj	Adj	—			2	MODIFIER		
2	insanlar		insan	Noun	Noun			A3pl Pnon Nom	10	SUBJECT	
3	hür	hür	Adj	Adj	—			10	MODIFIER		
4	,	,	Punc	Punc	—			3	PUNCTUATION		
5	haysiyet		haysiyet		Noun			Noun	A3sg Pnon Nom	7	COORDINATION
6	ve	ve	Conj	Conj	—			5	CONJUNCTION		
7	haklar	hak	Noun	Noun	A3pl Pnon Nom			8	POSSESSOR		
8	bakımından		bakım	Noun	Noun			A3sg P3sg Abl	10	MODIFIER	
9	eşit	eşit	Adj	Adj	—			10	MODIFIER		
10	doğarlar		doğ	Verb	Verb			Pos Aor A3pl	0	PREDICATE	
11	.	.	Punc	Punc	—			10	PUNCTUATION		

Veri Kümeleri

Tuğba Pamay, Umut Sulubacak, Dilara Torunoğlu-Selamet and Gülşen Eryiğit. 2015.
The annotation process of the ITU Web treebank.
LAW at NAACL

ITU Turkish Natural Language Processing Pipeline ITU Türkçe Doğal Dil İşleme Yazılım Zinciri

Umut Sulubacak, Tuğba Pamay Gülşen Eryiğit. 2016.
IMST: A revisited Turkish dependency treebank.
TURCLING

Tokenizer
Normalization
Deasciifier
Vowelizer
Spelling Corrector
isTurkish
Morphological Analyzer
Morphological Disambiguator
Named Entity Recognizer
Dependency Parsing
Whole pipeline
UD Mapper (Beta)
Datasets
Simple API
Changelog

Datasets

IMST Treebank v1.4.1

ITU Web Treebank v1.5.1

IMST-UD Treebank (UD v2.1)

IWT-UD Treebank (TBD)

Turkish PropBank

Turkish PropBank-UD

Veri Kümeleri

Umut Sulubacak and Gülşen Eryiğit. 2018.

Implementing universal dependency, morphology and multiword expression annotation standards for Turkish language processing.

Turkish Journal of Electrical Engineering & Computer Sciences, 26(3):1662-1672.

Umut Sulubacak, Memduh Gokirmak, Francis Tyers, Çağrı Çöltekin, Joakim Nivre, and Gülşen Eryiğit. 2016.

Universal dependencies for Turkish.

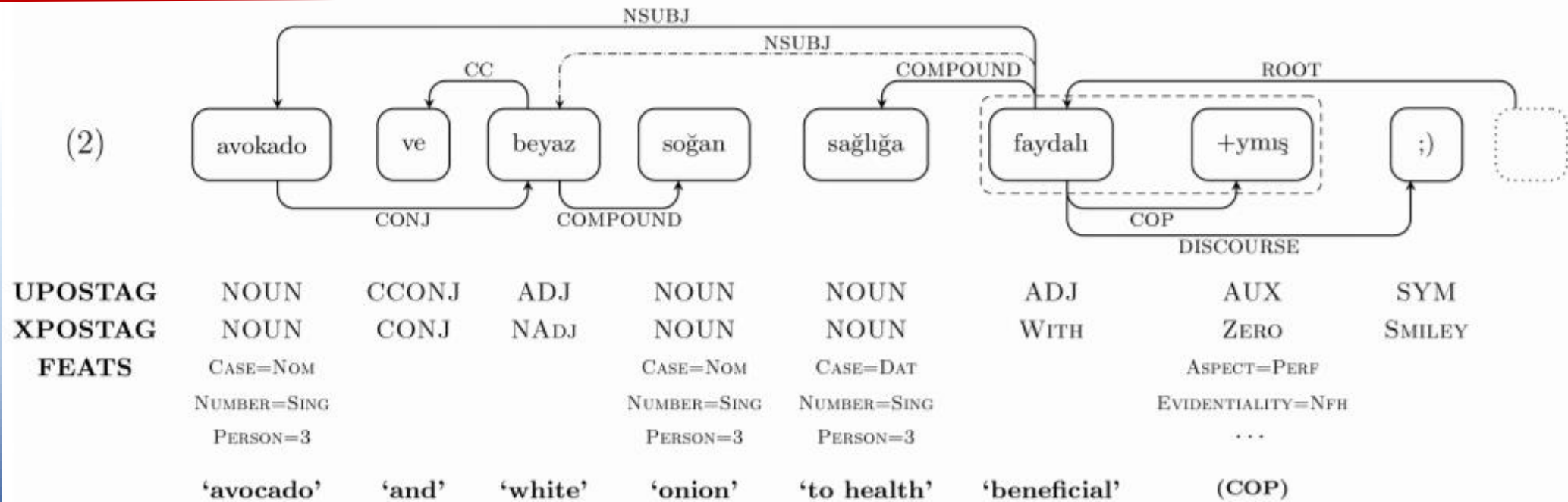
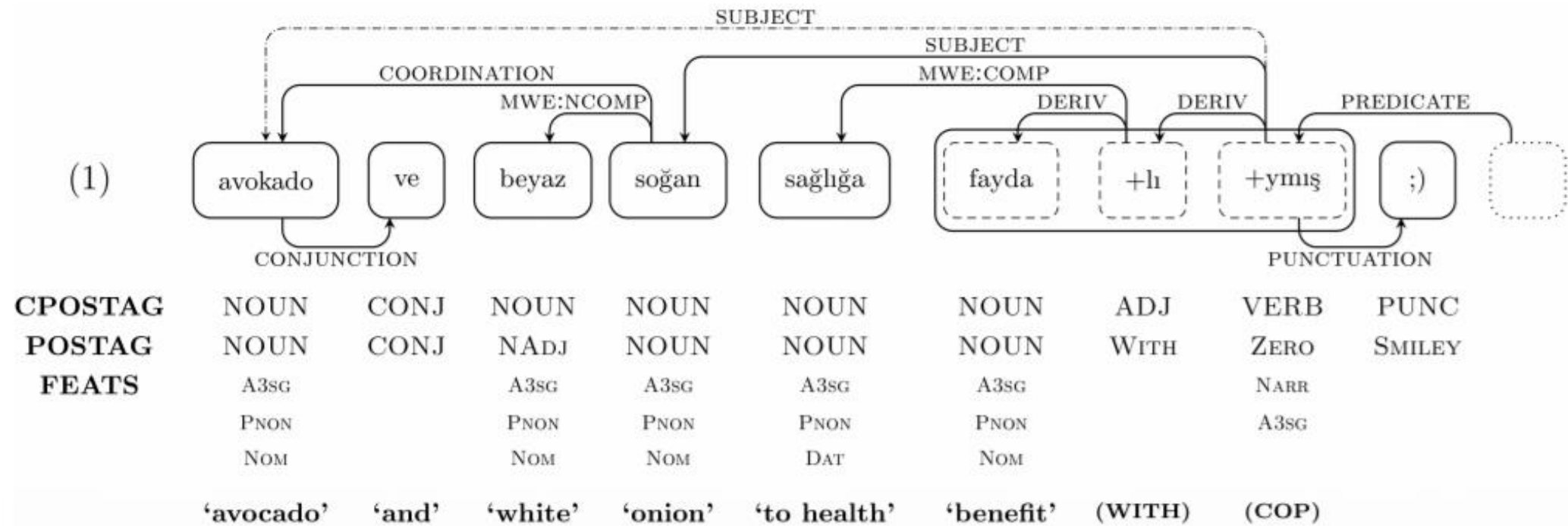
COLING

- *Uluslararası Bağılılık Projesi*
‘Universal Dependencies Project’

UD

<http://universaldependencies.org/>

71 dil , 122 ağaç yapılı derlem



Sunum İçeriği

- *Türkçe Doğal Dil İşleme Yazılım Zinciri Evreleri*
 - *Normalizasyon*
 - *Sözcük Analizi*
 - *Varlık İsmi Tanıma*
 - *Cümle Analizi*
 - *Veri Kümeleri*

➤ ***Örnek Projeler***

- *Aktif Araştırma Evreleri*
 - *Anlamsal Rol Etiketleme*
 - *Eşgönderge Çözümlemesi*
 - *Çok sözcüklü terimlerin Bulunması*

ITU Natural Language Processing Group

Projects and Collaborations

Parsing Web2.0 Sentences

(funding agency: The Scientific and Technological Research Council of Turkey TÜBİTAK and EU COST ACTION IC1207)

Turkish Mobile Personal Assistant

(funding agency: Republic of Turkey Ministry of Science, Industry and Technology, collaboration with HUAWEI)

Machine Translation from Turkish to Turkish Sign Language and the Avatar Animation

(funding agency TÜBİTAK)

Information Retrieval from Fax Messages

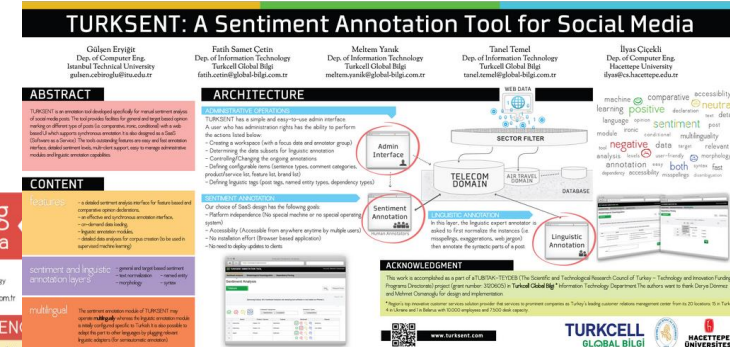
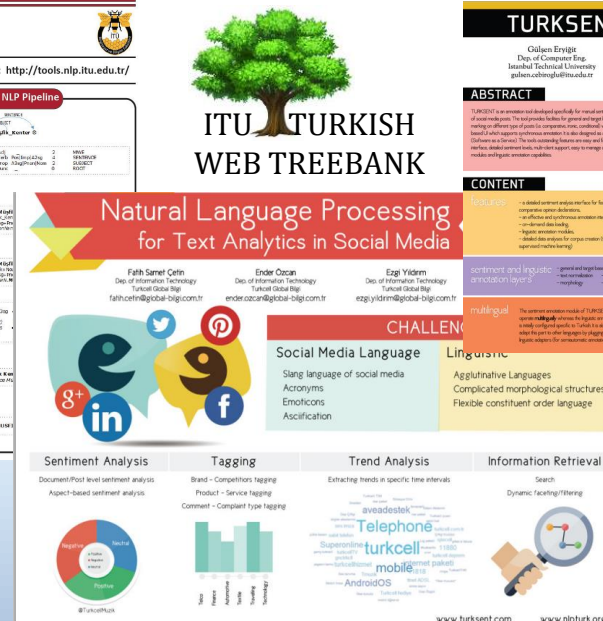
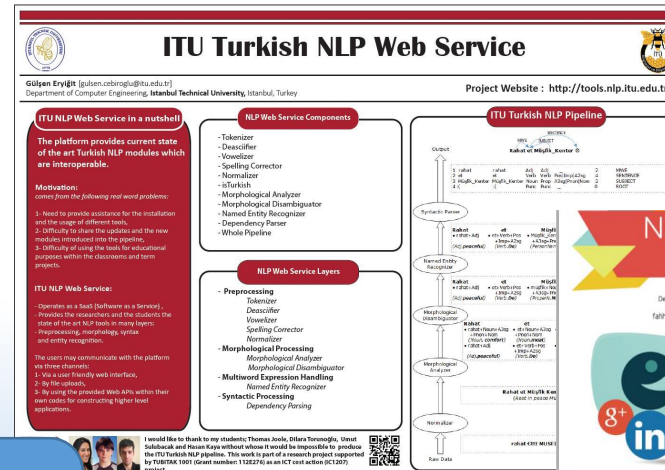
(funding agency: : TÜBİTAK – TEYDEB, collaboration with YapıKredi Teknoloji)

Aspect Based Sentiment Analysis for Turkish

(funding agency: TÜBİTAK – TEYDEB Technology and Innovation Funding Programs Directorate, collaboration with TURKCELL)

Sentiment Analysis and Big Data Platform for Turkish

(funding agency: TÜBİTAK – TEYDEB Technology and Innovation Funding Programs Directorate, collaboration with TURKCELL)



Social CRM

(funding agency: : TÜBİTAK – TEYDEB, collaboration with INTERTECH)

The usage of NLP for Geocoding

(funding agency: : TÜBİTAK – TEYDEB, collaboration with MAPTRIKS)

Sanal Asistanlar

Gökhan Çelikkaya and Gülşen Eryiğit. 2017.

Use of NLP Techniques for an Enhanced Mobile Personal Assistant: The Case of Turkish.

International Journal of Intelligent Systems and Applications in Engineering,
5(3):94–104

- *Soru-Soru & Soru-Cevap benzerlikleri hesaplama*
- *Varlık tespiti*
- *Niyet tespiti*
- *Aksiyon Planlama*

ITU Natural Language Processing Group

Projects and Collaborations

Parsing Web2.0 Sentences

(funding agency: The Scientific and Technological Research Council of Turkey TÜBİTAK and EU COST ACTION IC1207)

Sentiment Analysis and Big Data Platform for Turkish

(funding agency: TÜBİTAK – TEYDEB Technology and Innovation Funding Programs Directorate, collaboration with TURKCELL)

Turkish Mobile Personal Assistant

(funding agency: Republic of Turkey Ministry of Science, Industry and Technology, collaboration with HUAWEI)

Machine Translation from Turkish to Turkish Sign Language and the Avatar Animation

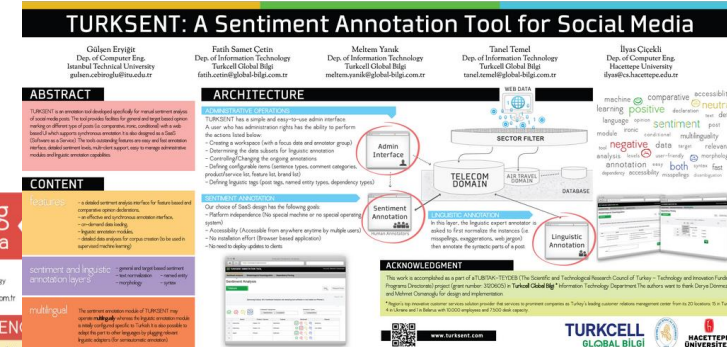
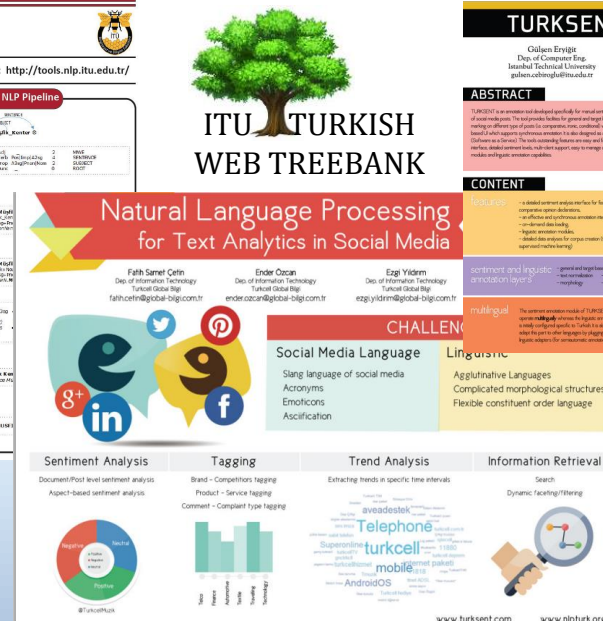
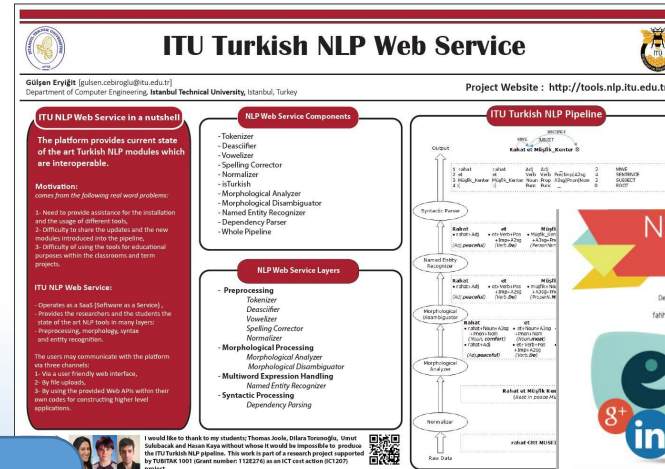
(funding agency TÜBİTAK)

Information Retrieval from Fax Messages

(funding agency: : TÜBİTAK – TEYDEB, collaboration with YapıKredi Teknoloji)

Aspect Based Sentiment Analysis for Turkish

(funding agency: TÜBİTAK – TEYDEB Technology and Innovation Funding Programs Directorate, collaboration with TURKCELL)



Social CRM

(funding agency: : TÜBİTAK – TEYDEB, collaboration with INTERTECH)

The usage of NLP for Geocoding

(funding agency: : TÜBİTAK – TEYDEB, collaboration with MAPTRIKS)

Türkçeden Türk İşaret Diline Çeviri

Written Turkish:

I love you.
Seni seviyorum.
you-acc love-imperfective-1sg
Sen-i sev-iyor-um

TID Glosses:

Ben sevmek sen

Visual TID:



Virtual TID:



Human avatar

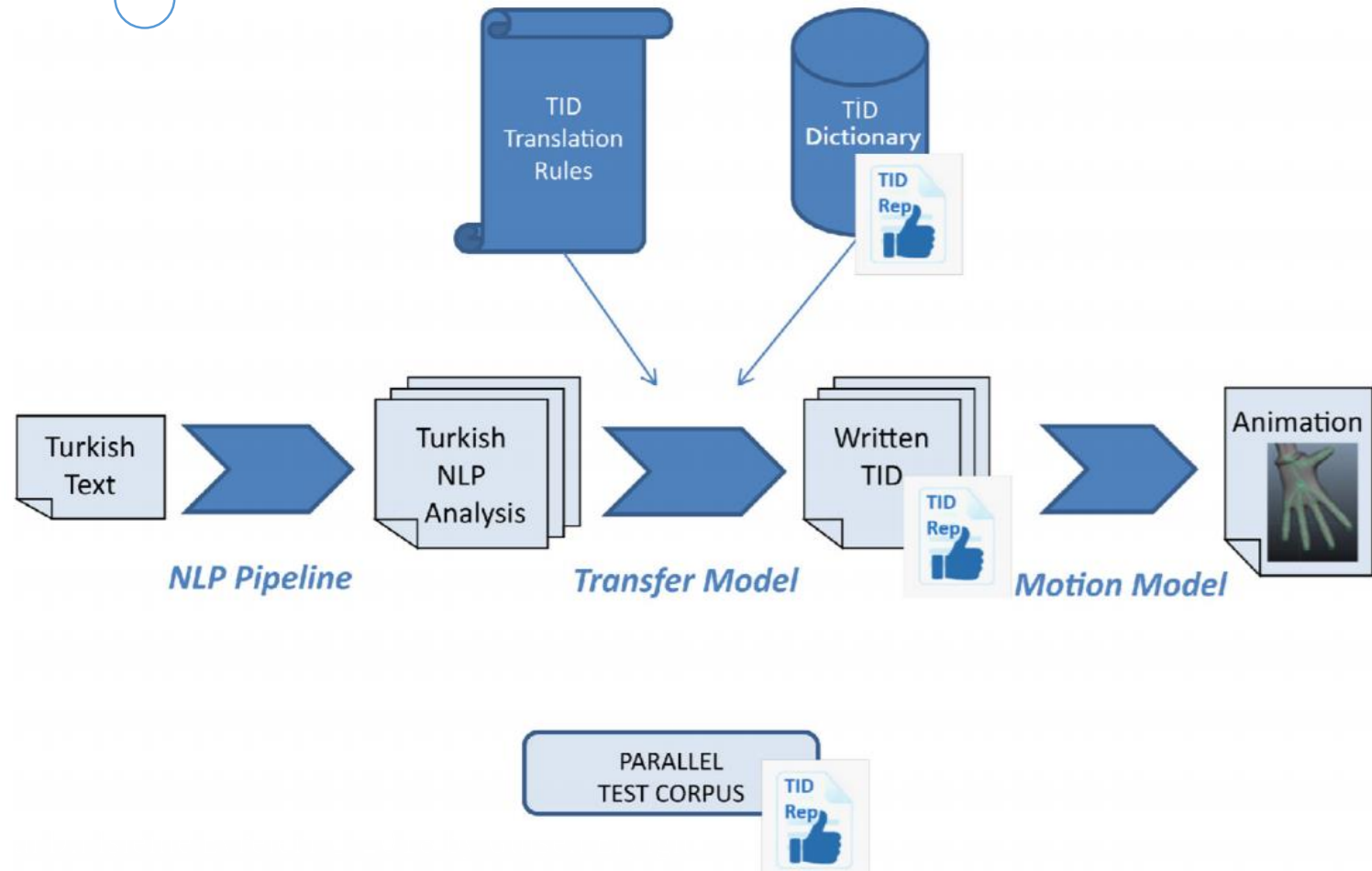


Humanoid avatar

Cihat Eryiğit, Hatice Köse, Meltem Kelepir and Gülşen Eryiğit.
2016.

Building machine-readable knowledge
representations for Turkish sign language
generation.

Knowledge-Based Systems, 108:179–194



ITU Natural Language Processing Group

Projects and Collaborations

Parsing Web2.0 Sentences

(funding agency: The Scientific and Technological Research Council of Turkey TÜBİTAK and EU COST ACTION IC1207)

Turkish Mobile Personal Assistant

(funding agency: Republic of Turkey Ministry of Science, Industry and Technology, collaboration with HUAWEI)

Machine Translation from Turkish to Turkish Sign Language and the Avatar Animation

(funding agency TÜBİTAK)

Information Retrieval from Fax Messages

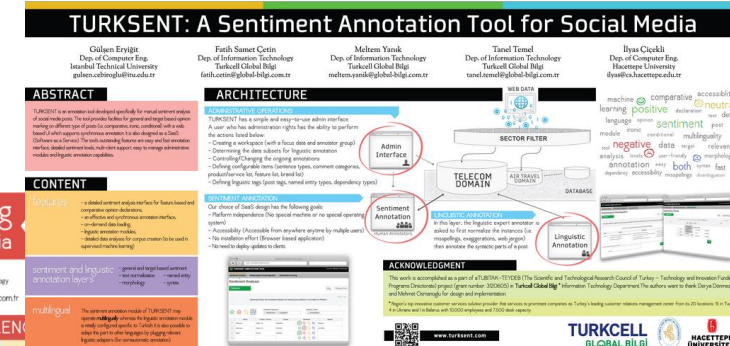
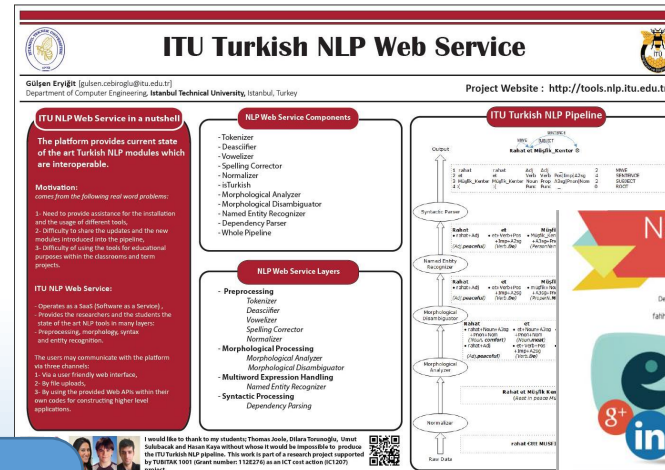
(funding agency: : TÜBİTAK – TEYDEB, collaboration with YapıKredi Teknoloji)

Aspect Based Sentiment Analysis for Turkish

(funding agency: TÜBİTAK – TEYDEB Technology and Innovation Funding Programs Directorate, collaboration with TURKCELL)

Sentiment Analysis and Big Data Platform for Turkish

(funding agency: TÜBİTAK – TEYDEB Technology and Innovation Funding Programs Directorate, collaboration with TURKCELL)



Social CRM

(funding agency: : TÜBİTAK – TEYDEB, collaboration with INTERTECH)

The usage of NLP for Geocoding

(funding agency: : TÜBİTAK – TEYDEB, collaboration with MAPTRIKS)

Turkish Journal
of Electrical Engineering & Computer Sciences, 26(2):830–843.

ITU Natural Language Processing Group

Projects and Collaborations

Parsing Web2.0 Sentences

(funding agency: The Scientific and Technological Research Council of Turkey TÜBİTAK and EU COST ACTION IC1207)

Turkish Mobile Personal Assistant

(funding agency: Republic of Turkey Ministry of Science, Industry and Technology, collaboration with HUAWEI)

Machine Translation from Turkish to Turkish Sign Language and the Avatar Animation

(funding agency TÜBİTAK)

Information Retrieval from Fax Messages

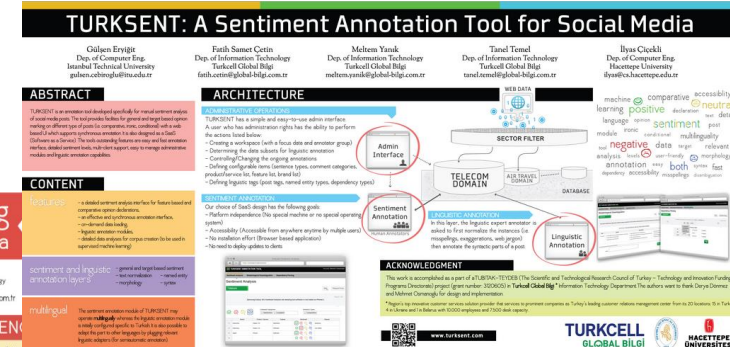
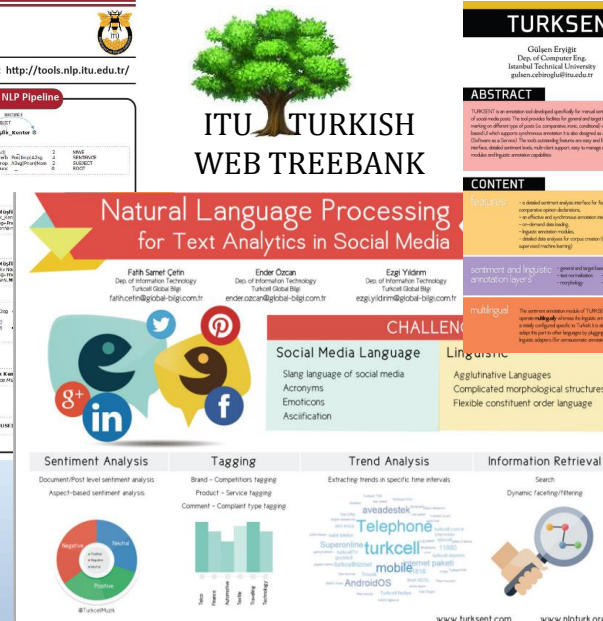
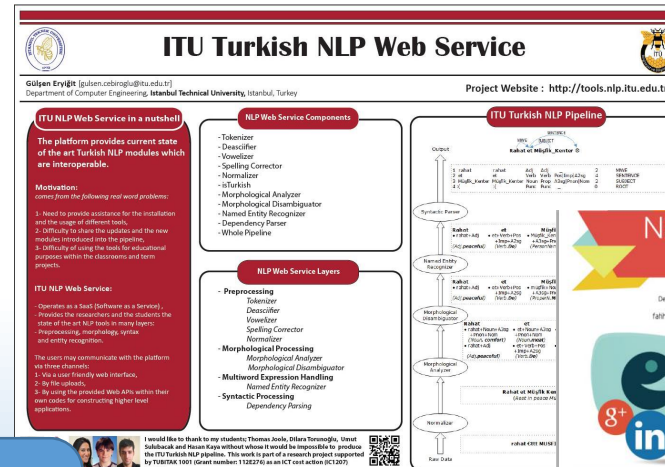
(funding agency: : TÜBİTAK – TEYDEB, collaboration with YapıKredi Teknoloji)

Aspect Based Sentiment Analysis for Turkish

(funding agency: TÜBİTAK – TEYDEB Technology and Innovation Funding Programs Directorate, collaboration with TURKCELL)

Sentiment Analysis and Big Data Platform for Turkish

(funding agency: TÜBİTAK – TEYDEB Technology and Innovation Funding Programs Directorate, collaboration with TURKCELL)



Social CRM

(funding agency: : TÜBİTAK – TEYDEB, collaboration with INTERTECH)

The usage of NLP for Geocoding

(funding agency: : TÜBİTAK – TEYDEB, collaboration with MAPTRIKS)

Hedef Tabanlı Duygu Analizi

Maria Pontiki et al. (including Gülşen Eryiğit). 2016.
**SemEval-2016 task 5: Aspect based
sentiment analysis**
SemEval

NDİ Sosyal Medya Analizinin daha doğru yapılmasını

```
<sentence id="1002:3">
```

```
<text>1 kere denedigim fiyatlarinin yüksek yemeklerinin lezzetsiz olduğu bir mekan önermiyorum.</text>
```

```
<Opinions>
```

```
<Opinion target="fiyatlarinin" category="RESTAURANT#PRICES" polarity="negative" from="17" to="29" />
```

```
<Opinion target="yemeklerinin" category="FOOD#QUALITY" polarity="negative" from="37" to="49" />
```

```
</Opinions>
```

```
</sentence>
```

Sunum İçeriği

- *Türkçe Doğal Dil İşleme Yazılım Zinciri Evreleri*
 - *Normalizasyon*
 - *Sözcük Analizi*
 - *Varlık İsmi Tanıma*
 - *Cümle Analizi*
 - *Veri Kümeleri*
- *Örnek Projeler*
- ***Aktif Araştırma Evreleri***
 - *Anlamsal Rol Etiketleme*
 - *Eşgönderge Çözümlemesi*
 - *Çok sözcüklü terimlerin Bulunması*

Sunum İçeriği

- *Türkçe Doğal Dil İşleme Yazılım Zinciri Evreleri*
 - *Normalizasyon*
 - *Sözcük Analizi*
 - *Varlık İsmi Tanıma*
 - *Cümle Analizi*
 - *Veri Kümeleri*
- *Örnek Projeler*
- *Aktif Araştırma Evreleri*
 - ***Anlamsal Rol Etiketleme***
 - *Eşgönderge Çözümlemesi*
 - *Çok sözcüklü terimlerin Bulunması*

Gözde Gül İşgüder, Mark Steedman, 2018.
**Character-Level Models versus
Morphology in Semantic Role
Labeling.**

ACL

m Zinciri

Anlamsal Rol Etiketleme

ITU Turkish Natural Language

ITU Türkçe Doğal Dil İşleme Yaşam Zinciri

sat.01

sell.01

A0: Satan kişi

A0: Seller

[Ayşe]_{A0} [elbisesini]_{A1}
[Fatma'ya]_{A2} [sattı]_{sat.01}

A1: Satılan şey

A1: Thing Sold

[Ayşe]_{A0} [her dress]_{A1}
[to Fatma]_{A2} [sold]_{sell.01}

A2: Alıcı

A2: Buyer

Mor
Morpholo
Name
C

Whole pipeline
UD Mapper (Beta)
Datasets
Simple API
Changelog

Fig. 8 Analysis of sentence “Man relaxed after making everyone laugh” with derived adverb “güldürünce” (when sb makes sb laugh). “Adam” (man) annotated with ArgA:causer, “herkes” (everyone) with Arg0:laugher defined by “gül.01” (laugh.01). “güldürünce” is labeled with ArgM-TMP: temporal adjunct “Adam” (man) with Arg1: thing relaxing for the verb “rahatla.01” (relax.01)

Report

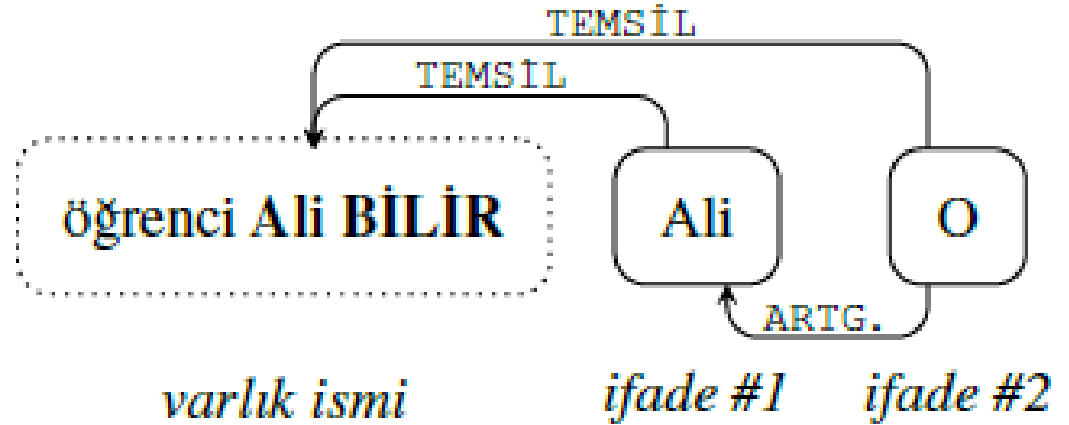
Sunum İçeriği

- *Türkçe Doğal Dil İşleme Yazılım Zinciri Evreleri*
 - *Normalizasyon*
 - *Sözcük Analizi*
 - *Varlık İsmi Tanıma*
 - *Cümle Analizi*
 - *Veri Kümeleri*
- *Örnek Projeler*
- *Aktif Araştırma Evreleri*
 - *Anlamsal Rol Etiketleme*
 - ***Eşgönderge Çözümlemesi***
 - *Çok sözcüklü terimlerin Bulunması*

Eşgönderge Çözümlemesi

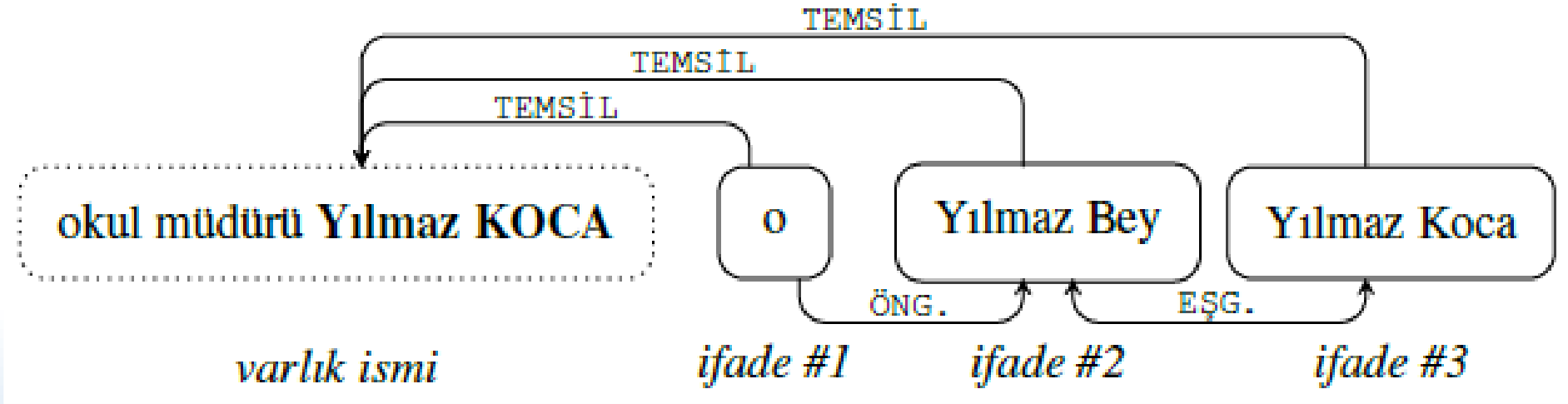
Tuğba Pamay and Gülşen Eryiğit. 2018.
Turkish Coreference
Resolution.
INISTA

- Artgönderim
 - Öngönderim
 - Adıl Çözümlemesi
- vb...*



Eşgönderge Çözümlemesi

Tuğba Pamay and Gülşen Eryiğit. 2018.
Turkish Coreference Resolution.
INISTA



Ali dün hastalanmış. O bugün okula gelemedi.
Bütün bunları o önceden söylemişti. Yine haklı çıktı Yılmaz Bey.
Yılmaz Koca bu konuda her zaman iyidir.

Tuğba Pamay and Gülşen Eryiğit. 2018.
Turkish Coreference Resolution.
INISTA

İfade Saptama

Ali dün hastalanmış. O bugün okula gelemedi.
Bütün bunları o önceden söylemişti. Yine haklı çıktı Yılmaz Bey.
Yılmaz Koca bu konuda her zaman iyidir.

Sınıflandırma

$C = \{(Ali,O),(o,Yılmaz Bey),(o,Yılmaz Koca),(Yılmaz Bey,Yılmaz Koca)\}$
 $C' = \{(Ali,o),(Ali,Yılmaz Bey),(Ali,Yılmaz Koca),$
 $(O,o),(O,Yılmaz Bey),(O,Yılmaz Koca)\}$

İfade Kümesi
Oluşturma

$K = \{(Ali,O), (o, Yılmaz Bey, Yılmaz Koca)\}$

Sunum İçeriği

- *Türkçe Doğal Dil İşleme Yazılım Zinciri Evreleri*
 - *Normalizasyon*
 - *Sözcük Analizi*
 - *Varlık İsmi Tanıma*
 - *Cümle Analizi*
 - *Veri Kümeleri*
- *Örnek Projeler*
- *Aktif Araştırma Evreleri*
 - *Anlamsal Rol Etiketleme*
 - *Eşgönderge Çözümlemesi*
 - ***Çok sözcüklü terimlerin Bulunması***

Çok Sözcüklü Terimler

Mathieu Constant, Gülşen Eryiğit, Johanna Monti, Lonneke van der Plas, Carlos Ramisch, Michael Rosner and Amalia Todirascu. 2017. Multiword Expression Processing: A Survey.

Computational Linguistics, 43(4):837–892

- *MWEs «Pain in the neck» in NLP*
Türkçe’de «boyun ağrısı» değil «karın ağrısı»

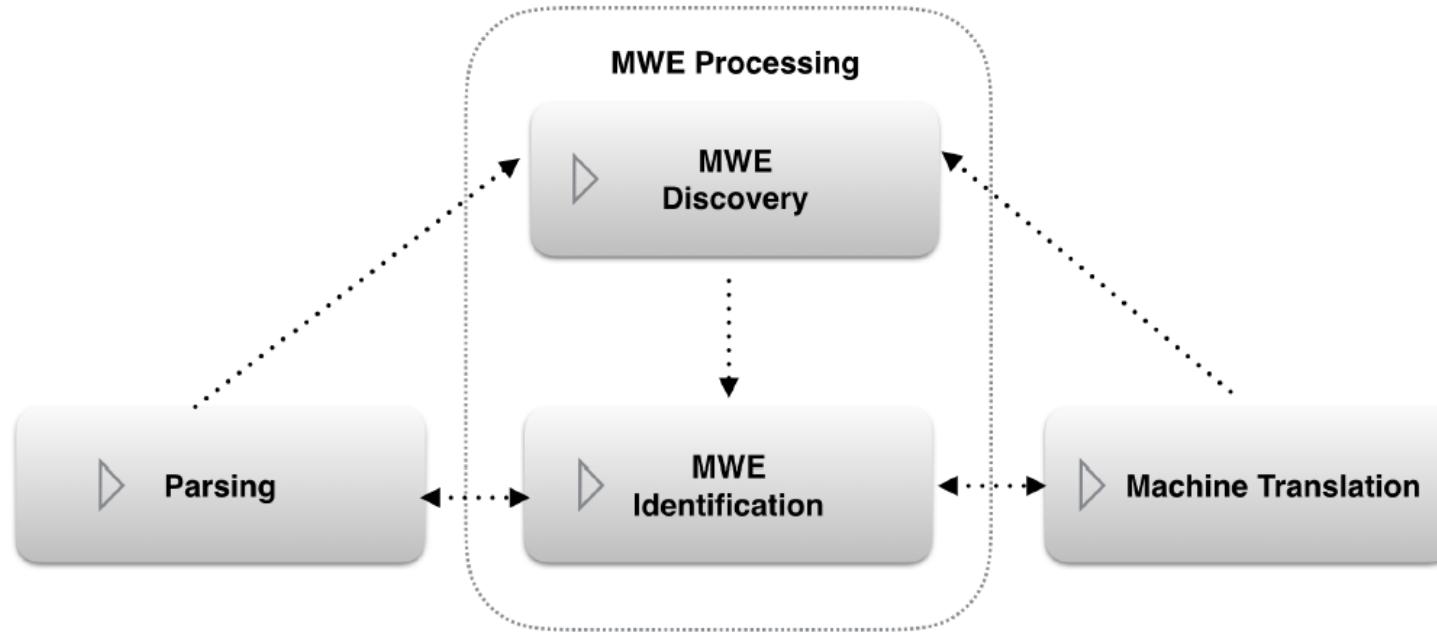
Umut Sulubacak and Gülşen Eryiğit. 2018.

Implementing universal dependency, morphology and multiword expression annotation standards for Turkish language processing.

Turkish Journal of Electrical Engineering & Computer Sciences, 26(3):1662-1672.

Çok Sözcüklü Terimler

«Parseme» Parsing and
Multi Word Expressions
EU Cost Action



Questions

