How to Parse MWEs ?
&
How to use MWEs in Parsing ?

Asst.Prof.Dr.Gülşen Eryiğit
Istanbul Technical University
Faculty of Computer and Informatics

Motivation

- Increasing the dependency parsing performance by using MWEs.
- Understanding «What is the best way of representing MWEs in our treebank?».

Outcomes so far

- A detailed analysis on the impact of different MWE CATEGORIES on dependency parsing.
- An evaluation methodology of parsing performance before and after MWE unification.
- «Can we use the dependency relations in order to decide if a word sequence is a real MWE?>>

STILL A BIG OPEN QUESTION!

Content of the Talk

1. **Multiword Expressions in Statistical Dependency Parsing**
   (*SPRML 2011* at IWPT Dublin.)
2. **Named Entity Recognizer**

- "Initial explorations on using CRFs for Turkish Named Entity Recognition." ***COLING 2012***, Mumbai, India
- "Named Entity Recognition on Real Data: A Preliminary Investigation for Turkish." ***AICT2013***, Baku, Azarbeijan.

3. **Problems in Detecting MWEs in a free constituent order language**

Multiword Expressions in Statistical Dependency Parsing

Gülşen Eryiğit, Tugay İlbay, Ozan Arkan Can
Istanbul Technical University,
Department of Computer Engineering

Abstract

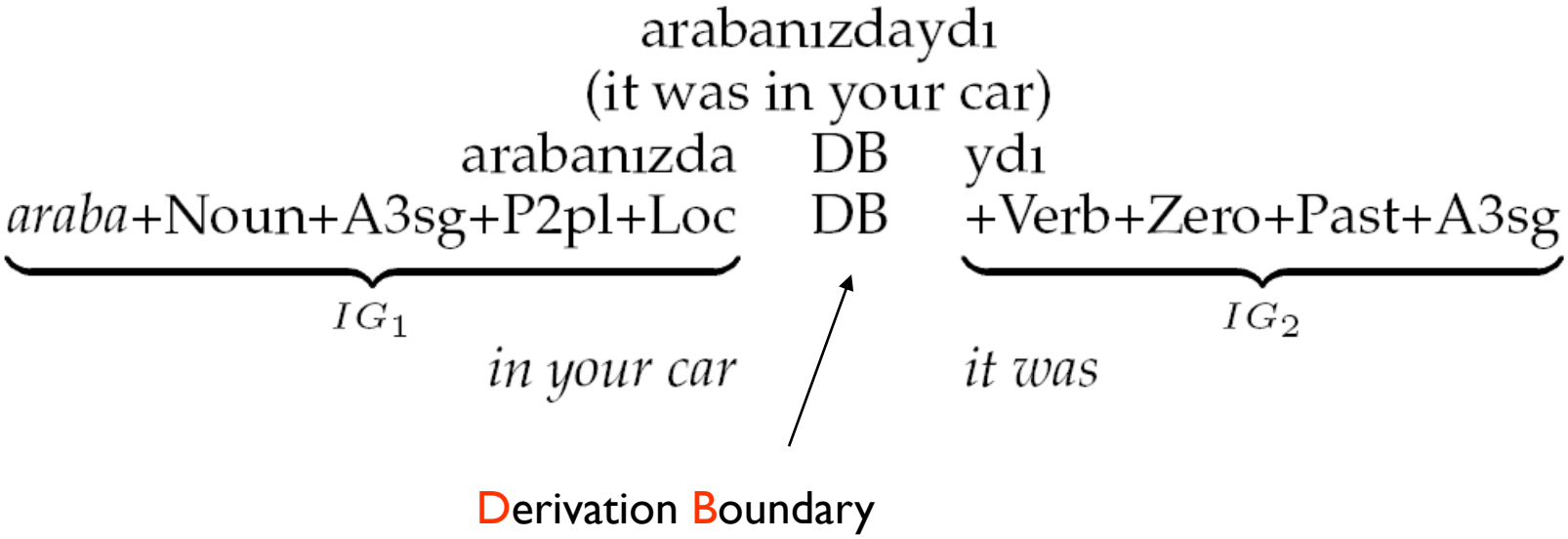
- The impact of extracting different types of multiword expressions (MWEs) in improving the accuracy of a data-driven dependency parser for a morphologically rich language (Turkish).

Turkish

- Free constituent order language
  - Most common forms are SOV and OSV
- Very productive agglutinative morphology
- Productive derivations
- Predominantly head final

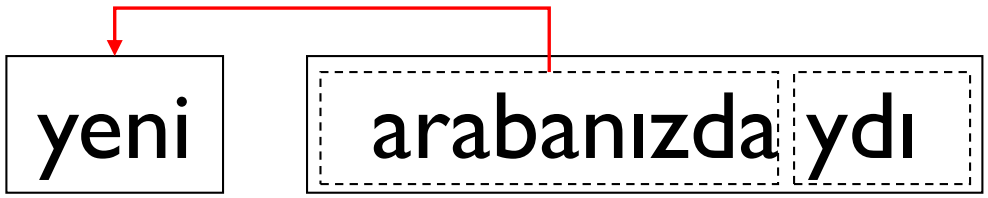
Turkish

- Representing morpholexical information- Inflectional Groups



Turkish

A dependency parser should extract IG-IG dependency links.



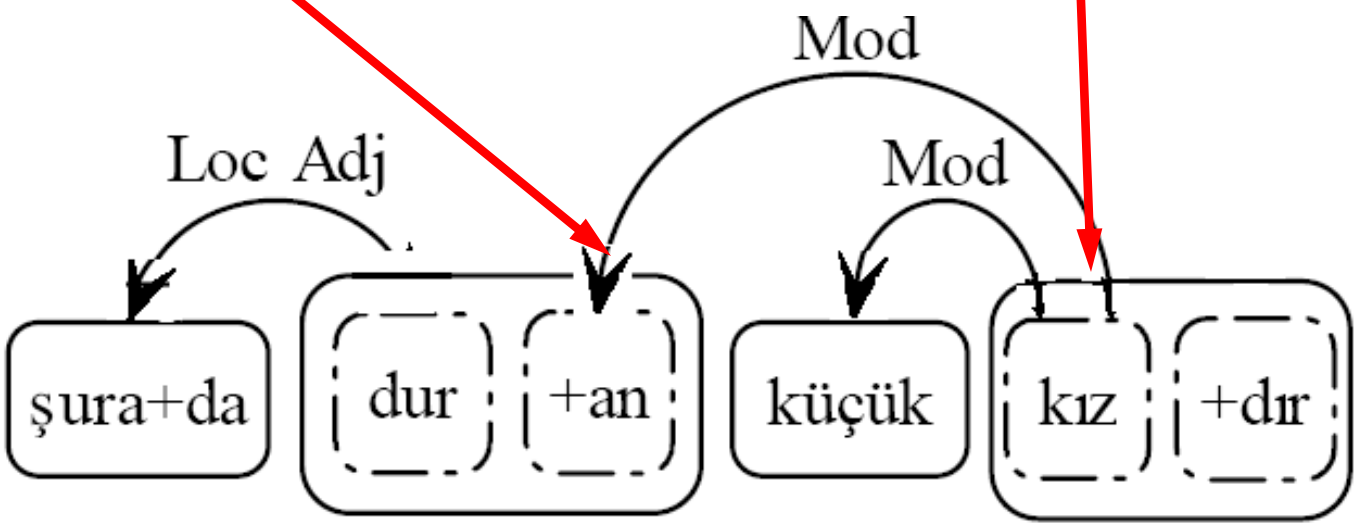
yeni+Adj araba+Noun+A3sg+P2sg+Loc
DB+Verb+Zero+Past+A3sg

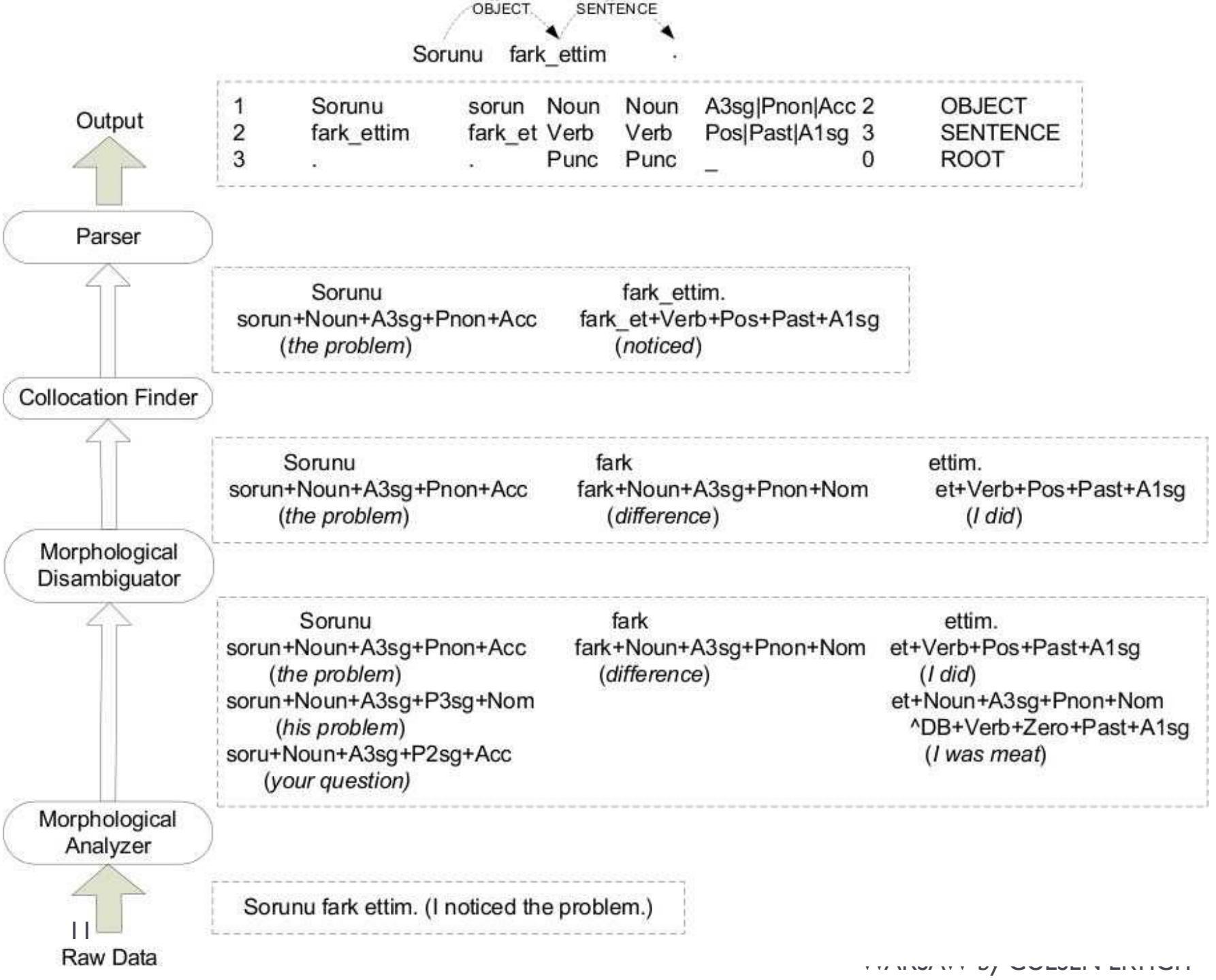
new + your car _ he was in =

(he was in your new car)

Turkish

- Dependency relations are between the last IG of the dependent and some IG of the head word.




OBJECT
SENTENCE
Sorunu fark_ettim .
1 Sorunu sorun Noun Noun A3sg|Pnon|Acc 2 OBJECT
2 fark_ettim fark_et Verb Verb Pos|Past|A1sg 3 SENTENCE
3 . . Punc Punc _ 0 ROOT
Output
Parser
Sorunu
sorun+Noun+A3sg+Pnon+Acc
(the problem)
fark_ettim.
fark_et+Verb+Pos+Past+A1sg
(noticed)
Collocation Finder
Sorunu
sorun+Noun+A3sg+Pnon+Acc
(the problem)
fark
fark+Noun+A3sg+Pnon+Nom
(difference)
ettim.
et+Verb+Pos+Past+A1sg
(I did)
Morphological Disambiguator
Sorunu
sorun+Noun+A3sg+Pnon+Acc
(the problem)
sorun+Noun+A3sg+P3sg+Nom
(his problem)
soru+Noun+A3sg+P2sg+Acc
(your question)
fark
fark+Noun+A3sg+Pnon+Nom
(difference)
ettim.
et+Verb+Pos+Past+A1sg
(I did)
et+Noun+A3sg+Pnon+Nom
^DB+Verb+Zero+Past+A1sg
(I was meat)
Morphological Analyzer
Sorunu fark ettim. (I noticed the problem.)
Raw Data

Representation of MWEs

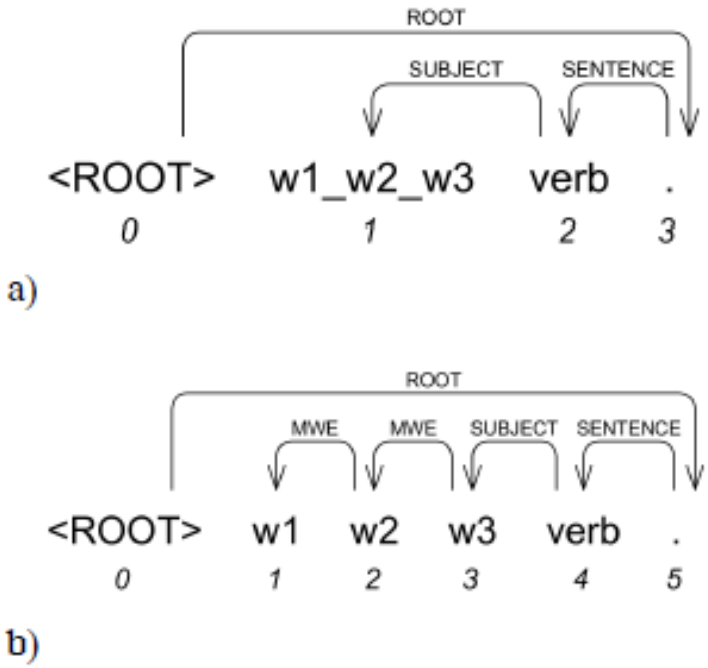


Figure 2: MWE representation in the Turkish Treebank

- 3 different versions of the treebanks:
- Vd (Detached Version):
  - MWEs are detached as in Figure 2-b.
  - Manually annotated postags of new MWE constituents (w1,w2)
  - The token indices and dependency links are renumbered in Conll Format. (a complex conversion for Turkish due to its IG structure)

Motivation-Parsing Raw Data

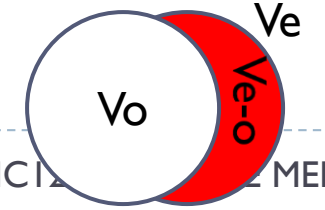
tested on	AS_U	AS_L	WW_U
Ery.(2008) Vo	$76.0_{\pm 0.2}$	$67.0_{\pm 0.3}$	$82.7_{\pm 0.5}$
Ery.(2008) Raw Data	$73.3_{\pm 0.3}$	$63.2_{\pm 0.3}$	$80.6_{\pm 0.7}$

Table 3: The parser's performance trained on the original treebank (Vo)

MWE type dependencies are missing in the training set

3 different versions of the treebanks.

- Vd (Detached Version)
- Ve (Enlarged Version):
  - First, the matching word groups with the MWEs in the dictionary are determined automatically.
  - Then, these matching word groups are manually checked and annotated and automatically combined into single units.
  - 1697 new MWEs (listed in the Turkish Dictionary) are added to the treebank;
- Ve-o:
  - is the treebank version where only the MWEs coming from the dictionary are annotated over the detached version.



Experiments & Results
Different Test Sets

tested on	AS_{IJ}	$AS_{I,}$	WW_{IJ}
Vo	76.1±0.2	67.4±0.3	83.0±0.2
Vd	74.7±0.2	63.3/66.5±0.3	81.8±0.2
Ve	75.5±0.2	66.7±0.3	82.5±0.2
Ve-o	74,0±0,2	62,4/65.7±0,3	81,1±0,1

- The MWEs coming from the dictionary have a harming effect on parsing accuracy
- The dependency counts are not the same in each version. Detailed analysis needed for differing dependencies.

Manual Classification of MWEs

- The current Turkish Treebank

MWE type	#of MWEs	#of Dep.	WW_U
Named ent.	618	941	83.7
Num. exp.	98	123	82.1
Comp. func.	54	54	5.6
Dup.	206	206	66.5
Comp. vn.	1061	1103	93.0

- The MWEs of the dictionary are (also) mostly from the last category (Compound verb and noun formations) where the parser is already very successful at finding.

Compound verb formations

fark	etmek
(difference)	*(to do)*

"fark etmek"

to notice

3 new versions of the treebank

- Subset 1 (S1)- Vo excluding MWEs of type compound verb and noun formations
- Subset 2 (S2)- S1 excluding MWEs of type duplications.
- Subset 3 (S3)- S2 excluding MWEs type compound function words.

Results with different MWE types

train on	test on	# of comb. MWEs	# of Dep.	Overall results AS_U	AS_L	WW_U
Vd	Vo	2040	43572	76.0±0.2	66.7±0.3	82.9±0.1
	S1	976 (1064)	44675	**76.0**±0.2	66.2/67.8±0.3	**82.9**±0.2
	S2	770	44881	75.9±0.2	66.1/67.8±0.3	82.8±0.2
	S3	716	44935	75.8±0.2	65.9/67.7±0.3	82.6±0.1

- It is just enough to find MWEs of the remaining types: named entities, numerical expressions, functional words, duplications

A Closer Look to the Results

3 different evaluation;

results on

- the overall dependencies
- the dependencies with "MWE" labels only

  (appearing after the detachment of MWE units)
- the dependencies excluding the ones with "MWE" labels

  (the surrounding structure in the sentence)

A Closer Look to the Results
(details given in the paper)

train on	test on	# of comb. MWEs	# of Dep.	Overall results / Results on MWE / Results Excl. MWE type dependencies(# of Dep= 43572) AS_U	AS_L	WW_U
Vo	Vo	2040	43572	n/c	n/c	n/c
	Vd	0	45999	75.9±0.2	67,2±0.3	82.8±0.1
Vd	Vo	2040	43572	n/c	n/c	n/c
	Vd	0	45999	76.0±0.2	66.6±0.3	82.8±0.2
	S1	976	44675	**76.0**±0.2	66.6±0.3	**82.9**±0.2

Experiments & Results

- A rule based dependency label chooser which assigns an appropriate label to the dependencies with MWE labels.
- Test are conducted by using these in training stage as well.

train.	test.	AS_U	AS_L	WW_U
Vo	Vo	76.1±0.2	67.4±0.3	83.0±0.2
	Vd*	74.7±0.2	66.1±0.2	81.8±0.2
S1*	S1*	76.1±0.2	67.6±0.3	82.9±0.2
	Vd*	75.3±0.2	66.7±0.2	81.9±0.2

Conclusions of the first section

- Collocating the MWEs of type compound noun and verb formations into single units increases the lexical scarcity and decreases the parsing performance.
- By using a MWE extractor for the remaining MWE types would improve the results to the level of gold standard treebank.

Content of the Talk

1. **Multiword Expressions in Statistical Dependency Parsing**
   (*SPRML 2011* at IWPT Dublin.)
2. **Named Entity Recognizer**

- "Initial explorations on using CRFs for Turkish Named Entity Recognition." ***COLING 2012***, Mumbai, India
- "Named Entity Recognition on Real Data: A Preliminary Investigation for Turkish." ***AICT2013***, Baku, Azarbeijan.

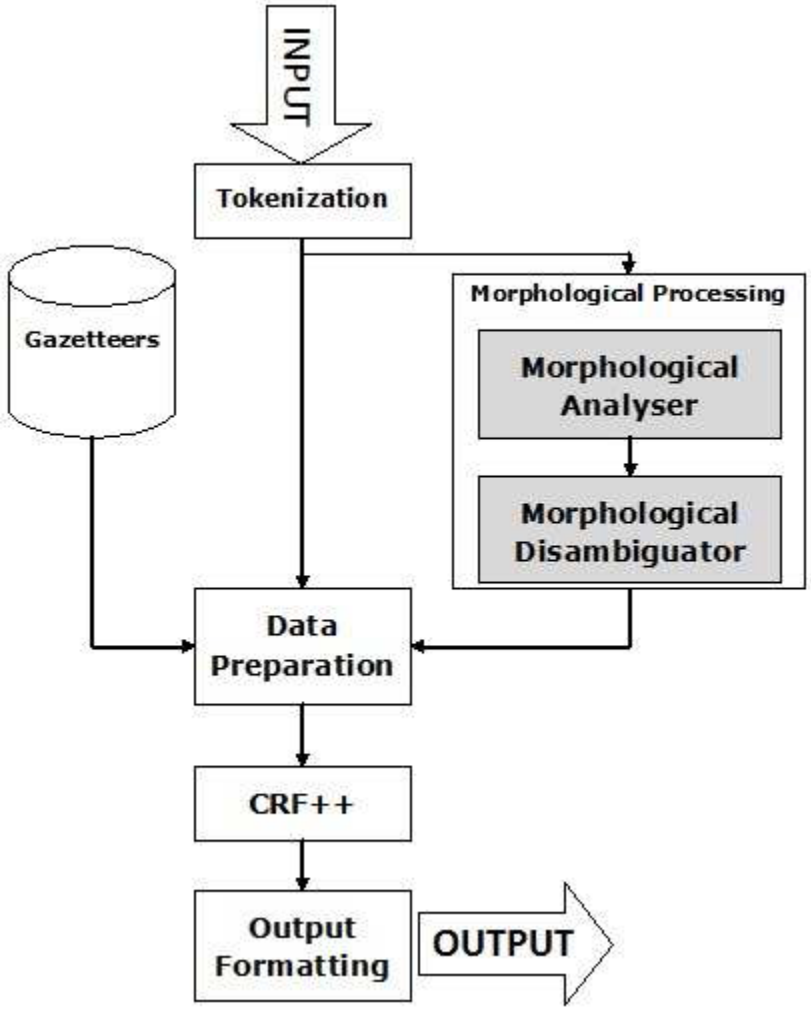
3. **Problems in Detecting MWEs in a free constituent order language**

Named Entity recognition using CRFs

NER

- Person
- Location
- Organization NAMES
- TIMEX – time expressions
- NUMEX – numerical expressions

USED FRAMEWORK



Morphological Features

Feature	
Stem	The stem information
Part of Speech Tag (POS)	The final part of speech category for each word. We assigned a special POS tag ("APOST") to the tokens separated by an apostrophe from the proper nouns
Noun Case (NCS)	"0" for non nominal tokens and one of [Nominative(NOM), Accusative/Objective(ACC), Dative (DAT), Ablative(ABL), Locative(LOC), Genitive(GEN), Instrumental(INS), Equative(EQU)] for nominals.
Proper Noun (PROP)	A binary feature indicating that the "+Prop" tag exists (1) in the selected morphological analysis or not (0).
All Inflectional Features (INF)	All inflectional tags after the POS category. If a derivation exists then the inflectional tags after the last derived POS category is used.

Lexical Features

- **Case Feature (CS) :** The information about lowercase and uppercase letters used in the current token. This feature takes 4 different values:
  - lowercase(0),
  - UPPERCASE(1),
  - Proper Name Case(2)
  - miXEd CaSe(3)
- **Start of the Sentence (SS) :** A binary feature indicating that the current token is the beginning of a sentence (1) or not (0).

Gazetteer Lookup Features

- Six different features used for each of the six gazetteers introduced before.
- Lookup features for base gazetteers (BG) have a 1 value if the token exists in the corresponding gazetteer and 0 otherwise.
- Generator gazetteer lookup features (GG) are binary features as well but this time the stem of the word is checked instead of the full surface form.

State of the art results for Turkish

Related work	Best Result	Ev.Metr.	Domain	NE Types
Özkaya and Diri (2011)	84.24	*n/a*	E-mail texts	ENAMEX
Küçük and Yazıcı (2012)	90.13	OTHER	General news	ENAMEX,TIMEX,NUMEX
Tür et al. (2003)	91.56	MUC	General news	ENAMEX
Bayraktar and Temizel (2008)	81.97	MUC	Financial Texts	PERSON NAMES
OURS	**94.59**	**MUC**	**General news**	**ENAMEX**
Tatar and Cicekli (2011)	91.08	CoNLL	Terrorism news	ENAMEX,TIMEX
Yeniterzi (2011)	88.94	CoNLL	General news	ENAMEX
OURS	**91.94**	**CoNLL**	**General news**	**ENAMEX**

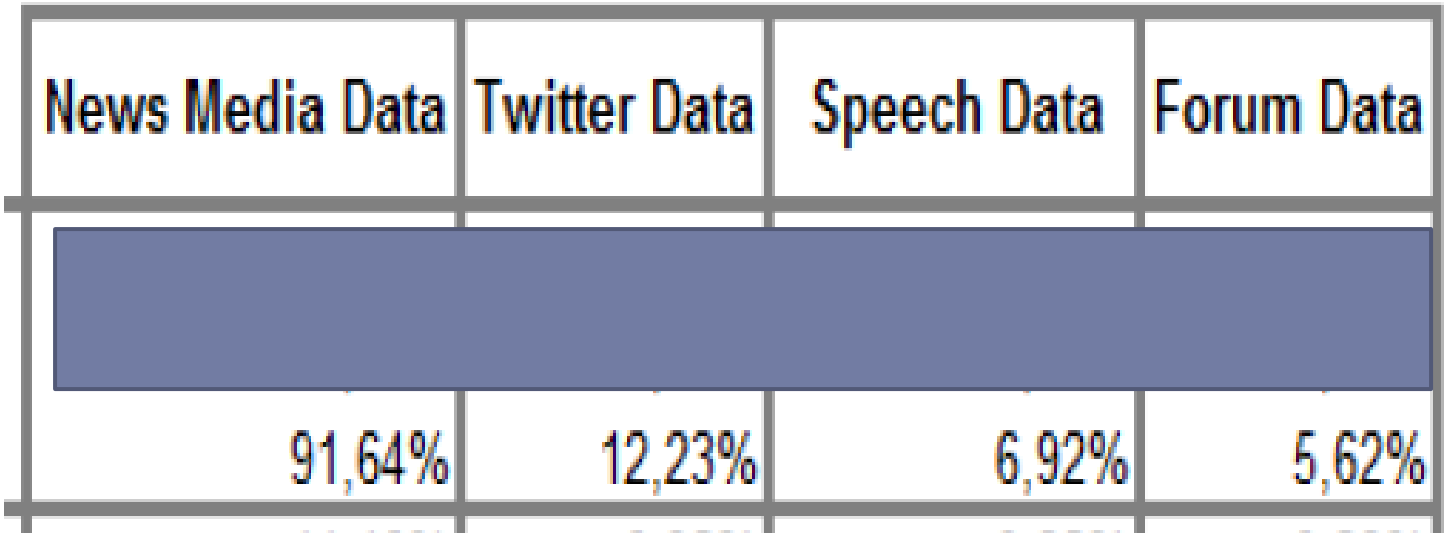
Content of the Talk

1. **Multiword Expressions in Statistical Dependency Parsing**
   (*SPRML 2011* at IWPT Dublin.)
2. **Named Entity Recognizer**

- "Initial explorations on using CRFs for Turkish Named Entity Recognition." ***COLING 2012***, Mumbai, India
- "Named Entity Recognition on Real Data: A Preliminary Investigation for Turkish." ***AICT2013***, Baku, Azarbeijan.

3. **Problems in Detecting MWEs in a free constituent order language**

What about real data?



NER RESULTS

What about real data?

Features	Model A	Model B	Model C
Word	x	x	x
Stem	x	x	x
[illegible]	[illegible]	[illegible]	x
			x
			x
			x
			x
			x

		News Media Data	Twitter Data	Speech Data	Forum Data
No Normalization	Model A	14,70%	0,93%	0,82%	0,40%
	Model B	88,56%	13,88%	50,69%	2,86%
	Model C	91,64%	12,23%	6,92%	5,62%
Normalized	Model A	14,10%	0,92%	0,82%	0,39%
	Model B	86,72%	15,27%	50,84%	2,41%
	Model C	82,59%	19,28%	6,90%	2,16%
Normalized + Capitalized	Model A	14,06%	1,33%	3,95%	0,38%
	Model B	83,60%	14,57%	50,26%	2,18%
	Model C	75,48%	15,00%	32,88%	1,49%

What about real data?

- Even the NER task that we consider accomplished is still there for this new domain?
- Meltem Yanık → a women name in Turkish
- meltem yanık → (gentle breeze) burn

- How to differentiate between these?

Problems of representing the remaining MWEs in Free Word Order languages.

How should we annotate them before or after the parsing stage?

Problems of representing MWEs in free Word order languages.

- Turkish is a free constituent order language.
- Since the morphological features mostly determines the syntactic role of the constituents, they may easily change position in the sentence according to different emphases.

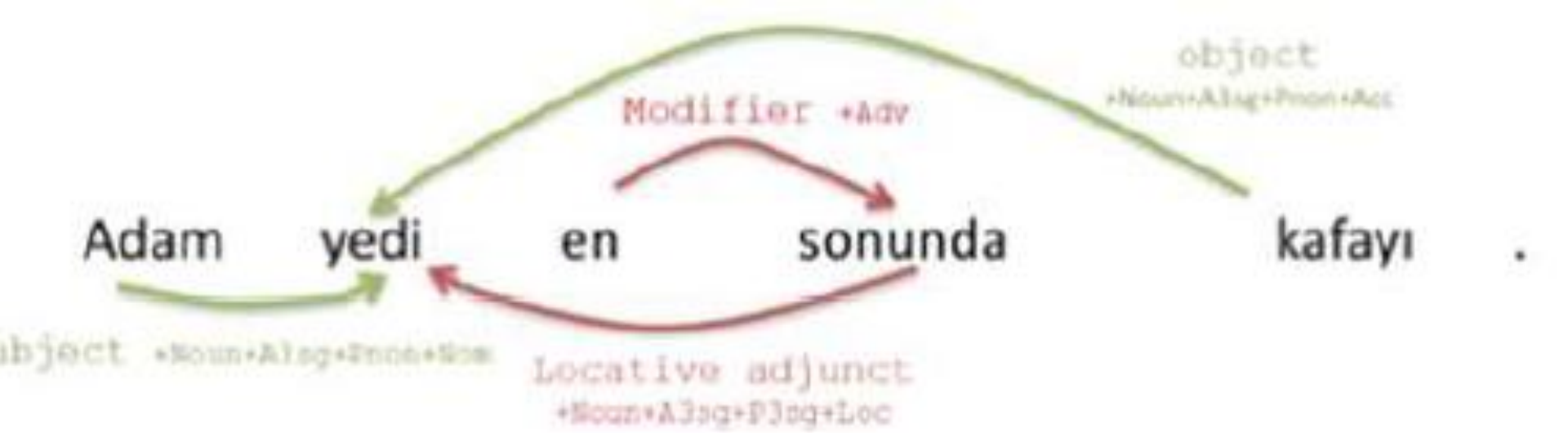
- kafayı_ye+Verb...
  "get mentally deranged" (literally "eat the head")

- Adam en sonunda **kafayı yedi.**
- The men finally became deranged.

- Adam **kafayı** en sonunda **yedi**.
- Adam **kafayı yedi** en sonunda.
- En sonunda adam **yedi kafayı**.
- **Kafayı yedi** adam en sonunda.

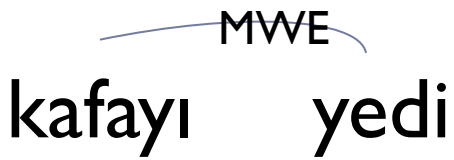
How should we represent the MWEs in the treebanks.

- Two answers: 1) unification 2) a special dependency type
- Both of the answers are relevant according to different MWE categories



Decisions so far …

- Unification for Named Entities, ….
  Asst._Prof._Dr._Gülşen_Eryiğit
- Specific dependencies for verb constructions.



- These results conforms with the empirical results of our first section:

Results with different MWE types

train on	test on	# of comb. MWEs	# of Dep.	Overall results AS_U	AS_L	WW_U
Vd	Vo	2040	43572	76.0±0.2	66.7±0.3	82.9±0.1
	S1	976 (1064)	44675	**76.0**±0.2	66.2/67.8±0.3	**82.9**±0.2
	S2	770	44881	75.9±0.2	66.1/67.8±0.3	82.8±0.2
	S3	716	44935	75.8±0.2	65.9/67.7±0.3	82.6±0.1

- It is just enough to find MWEs of the remaining types: named entities, numerical expressions, functional words, duplications

Decisions so far …

- Unification for Named Entities, ….

Give example

- Specific dependencies for verb constructions.
- Give example

- These results conforms with the empirical results of our first section:
- Of course the verb constructions are very valuable (e.g. İn Machine translation), but we **couldn't prove yet** their impact on parsing performances.

2 open questions from yesterday

- «compound recognition before or after parsing?» by Mathieu

My answer :

Some categories before and some after.

- The lexicon representation by Shully Winter.

The syntactic slots are very impressive.

But is it enough flexible to apply to free word order languages?

Thank you for listening