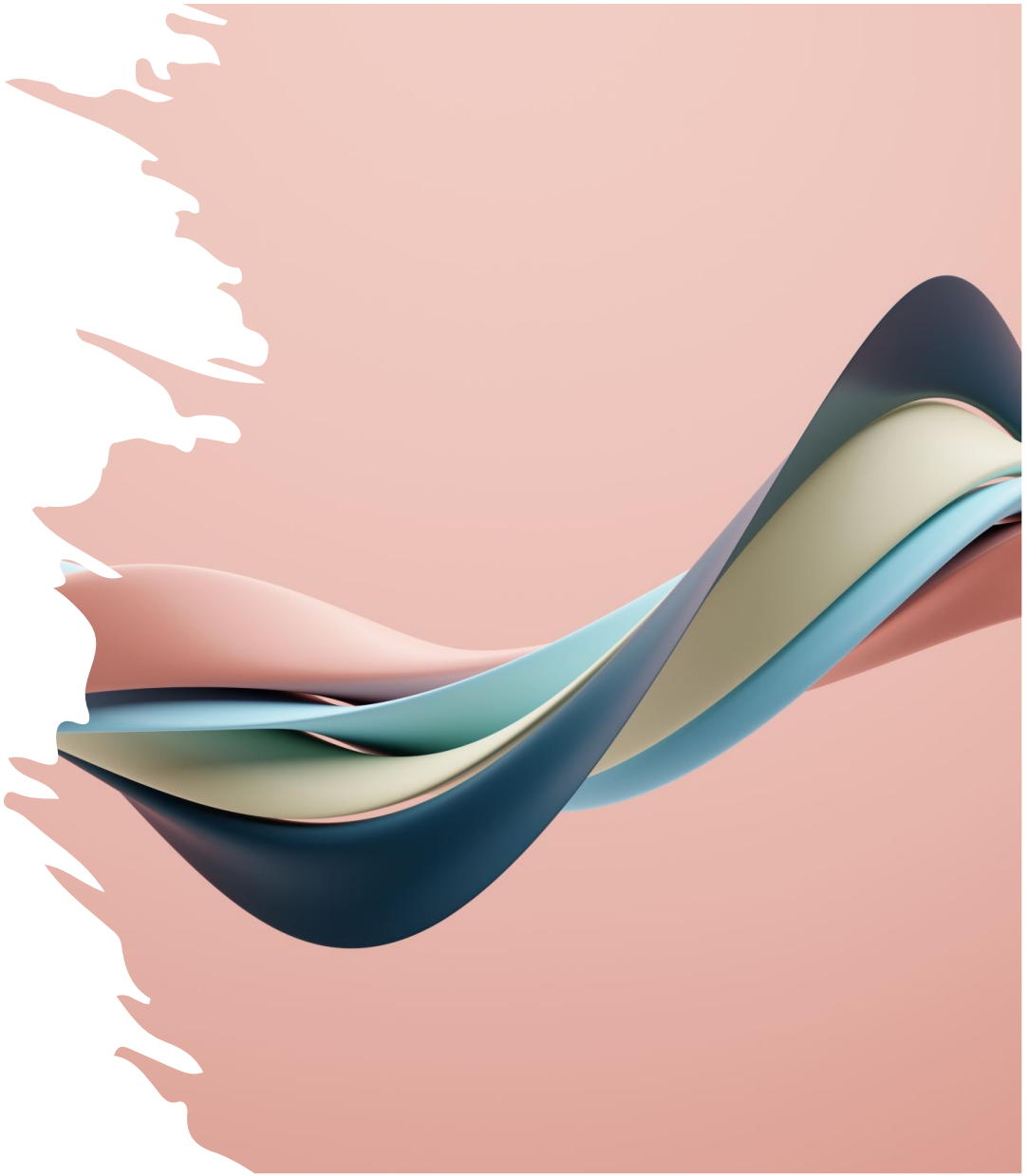
Journey of TreeBanking

STARLANG YAZILIM
DANIŞMANLIK



Outline

- Starlang TreeBanks
- Annotation Scheme
- Annotation Interface
- Future Work

TreeBanks

- Atis (Tr, En)
- KeNet
- Penn
- Tourism
- FrameNet

Atis TreeBank (2.9)

- Airline Travel Information System
- Parallel TreeBank: Atis English
- Mainly question and imperative sentences
- Small vocabulary, English and Turkish mixed
- 5.432 Sentences (4274 train, 586 test, 572 dev)
- 45.875 Tokens (61.879 Tokens in English)
- Contributors: Mehmet Köse[4], Olcay Taner Yıldız[2]

KeNet TreeBank (2.8)

- Turkish WordNet KeNet
- Example sentences for synsets. Sentences are mainly taken from Turkish Language Association.
- Complex vocabulary
- 18.687 Sentences (9.345 train, 9.342 test)
- 178.658 Tokens
- Contributors: Aslı Kuzgun[1], Neslihan Cesur[3], Olcay Taner Yıldız[2], Oğuzhan Kuyrukçu[5], Arife Betül Yenice[6], Bilge Nas Arıcan[7], Ezgi Sanıyar[8]

Penn TreeBank (2.8)

- Largest Turkish TreeBank
- Translated sentences from subsample of Penn Constituency TreeBank
- Parallel TreeBank: Penn English
- Sentences of length < 20 words in English
- Complex vocabulary, English and Turkish mixed
- 16.396 Sentences
- 183.555 Tokens
- Contributors: Neslihan Cesur[3], Aslı Kuzgun[1], Olcay Taner Yıldız[2], Büşra Marşan[9], Neslihan Kara[10], Bilge Nas Arıcan[7], Merve Özçelik[11], Deniz Baran Aslan[12]

Tourism TreeBank (2.8)

- Customer reviews of a tourism company
- Small, sometimes irregular sentences, i.e. multiple sentences in one sentence, nominal sentences, etc.
- Small vocabulary, domain dataset.
- 19.833 Sentences
- 91.469 Tokens
- Contributors: Aslı Kuzgun[1], Neslihan Cesur[3], Olcay Taner Yıldız[2], Oğuzhan Kuyrukçu[5], Büşra Marşan[9], Bilge Nas Arıcan[7], Neslihan Kara[10], Deniz Baran Aslan[12], Ezgi Sanıyar[8], Cengiz Asmazoğlu[13]

FrameNet (2.8)

- Turkish FrameNet
- Example constructed sentences taken from Turkish FrameNet
- Complex vocabulary, basically one sentence per verb
- 2.698 Sentences
- 19.221 Tokens
- Contributors: Neslihan Cesur[3], Aslı Kuzgun[1], Olcay Taner Yıldız[2], Büşra Marşan[9], Oğuzhan Kuyrukçu[5], Bilge Nas Arıcan[7], Ezgi Sanıyar[8], Neslihan Kara[10], Merve Özçelik[11]

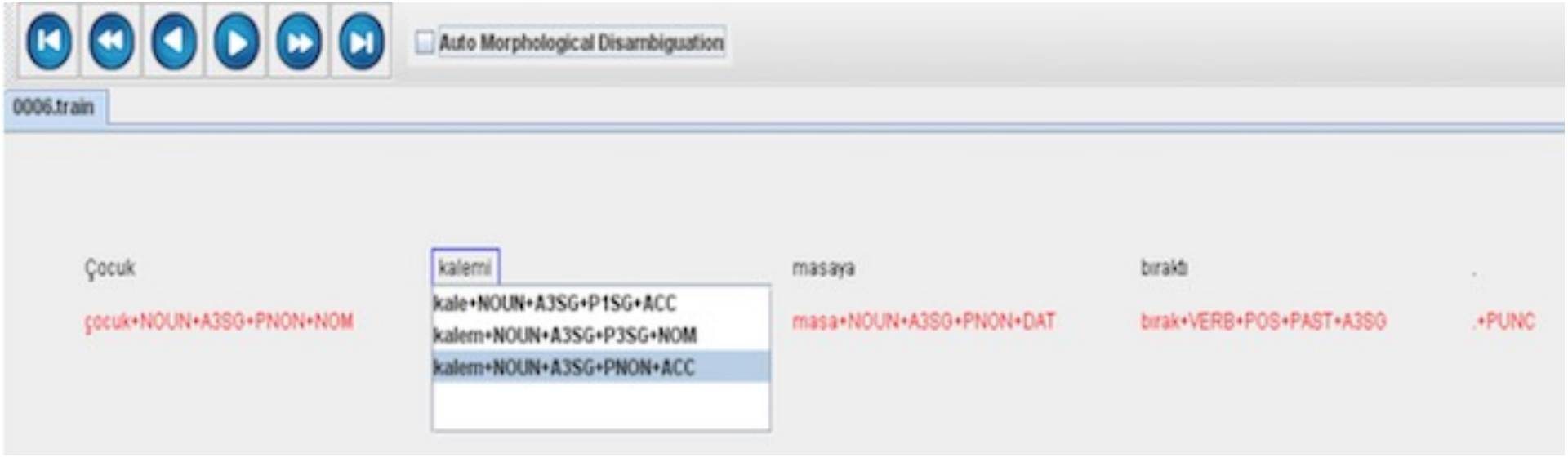
Annotation Scheme

- Morphological Disambiguation Process
- Dependency Annotation Process
- Backed by NlpToolkit Library (8 Programming Languges, over 1 million lines of code)
  - MorphologicalAnalysis, MorphologicalDisambiguation
  - WordNet
  - DependencyParser, UniversalDependencyParser
  - AnnotatedSentence
  - DataCollector

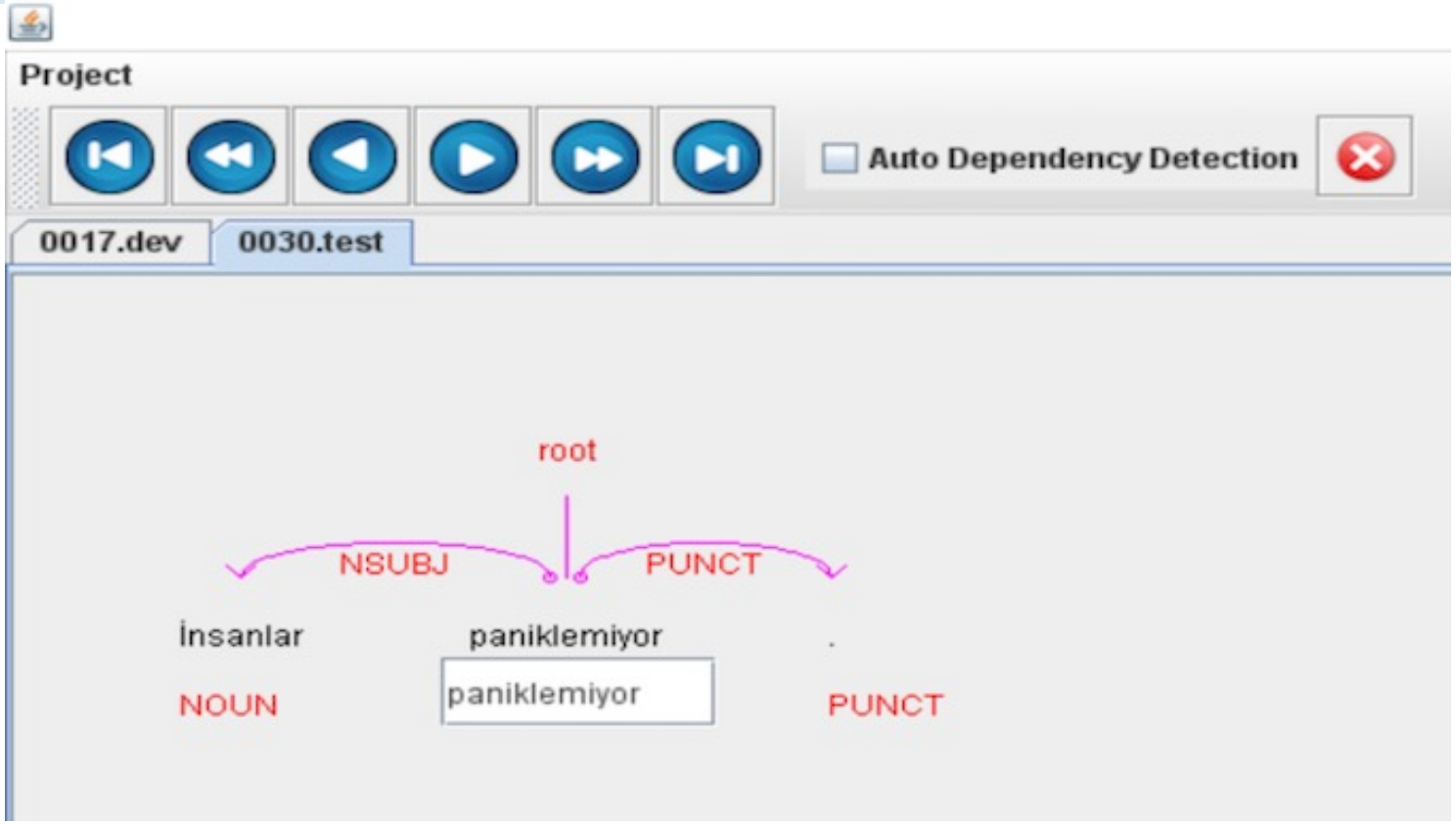
Features

- Layered architecture: surface form, morphological analysis, semantics, …
- Store intermediate representation, namely morphological analysis
- After morphological disambiguation, root, universal pos tag determined.
- Morphological analyzer is connected to Turkish WordNet, each root form is taken from Turkish WordNet (has a sense in WordNet)
- Universal features are extracted from morphological analysis.

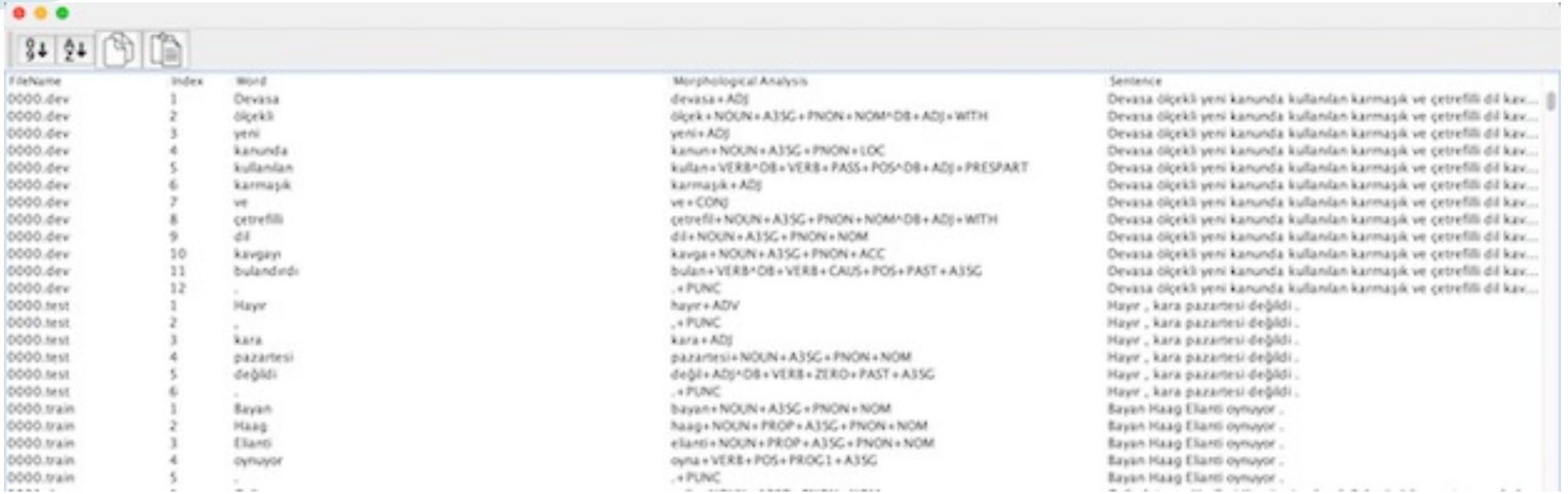
Morphological Disambiguation



Editing Words

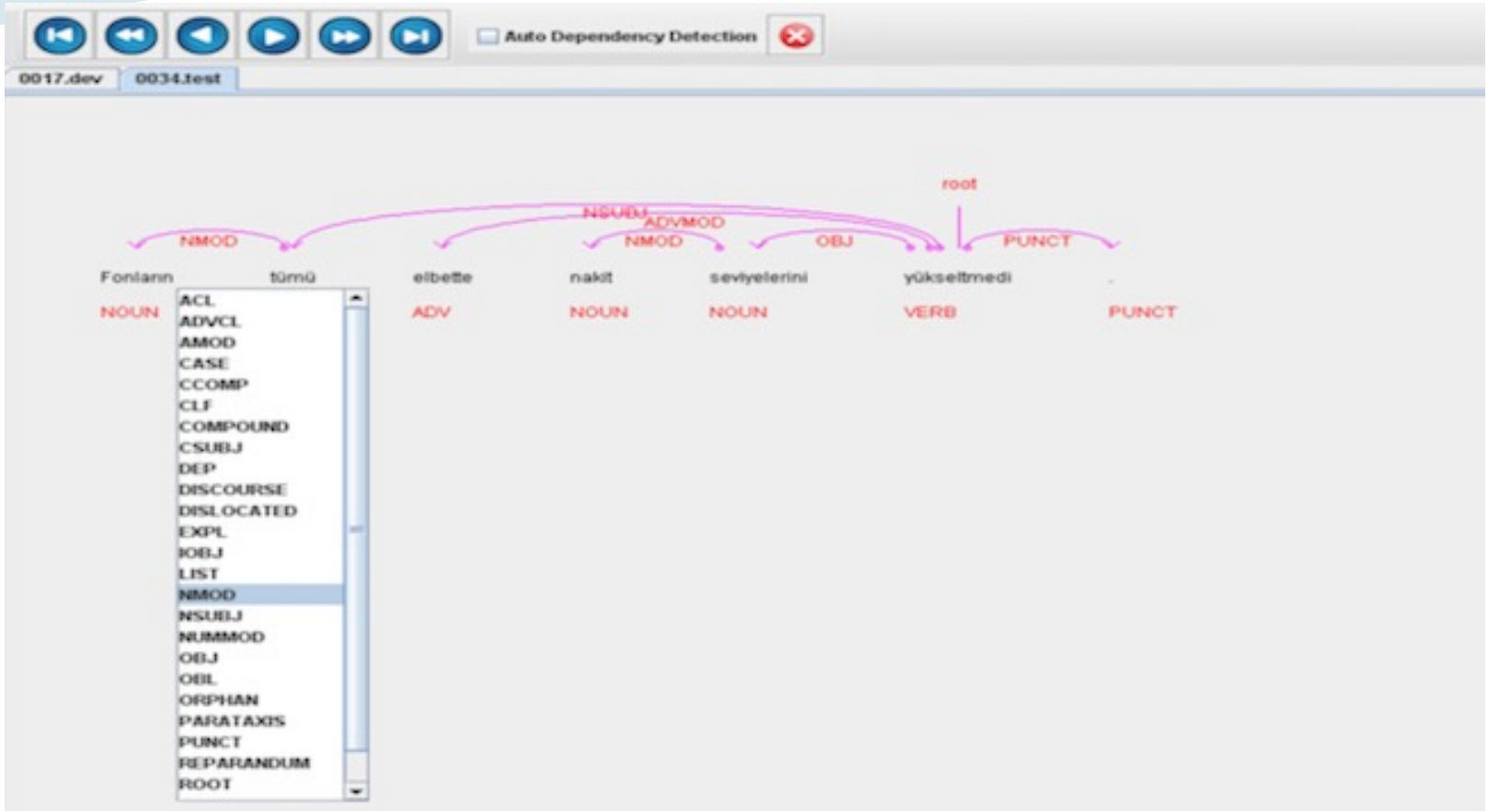


Annotation Comparison

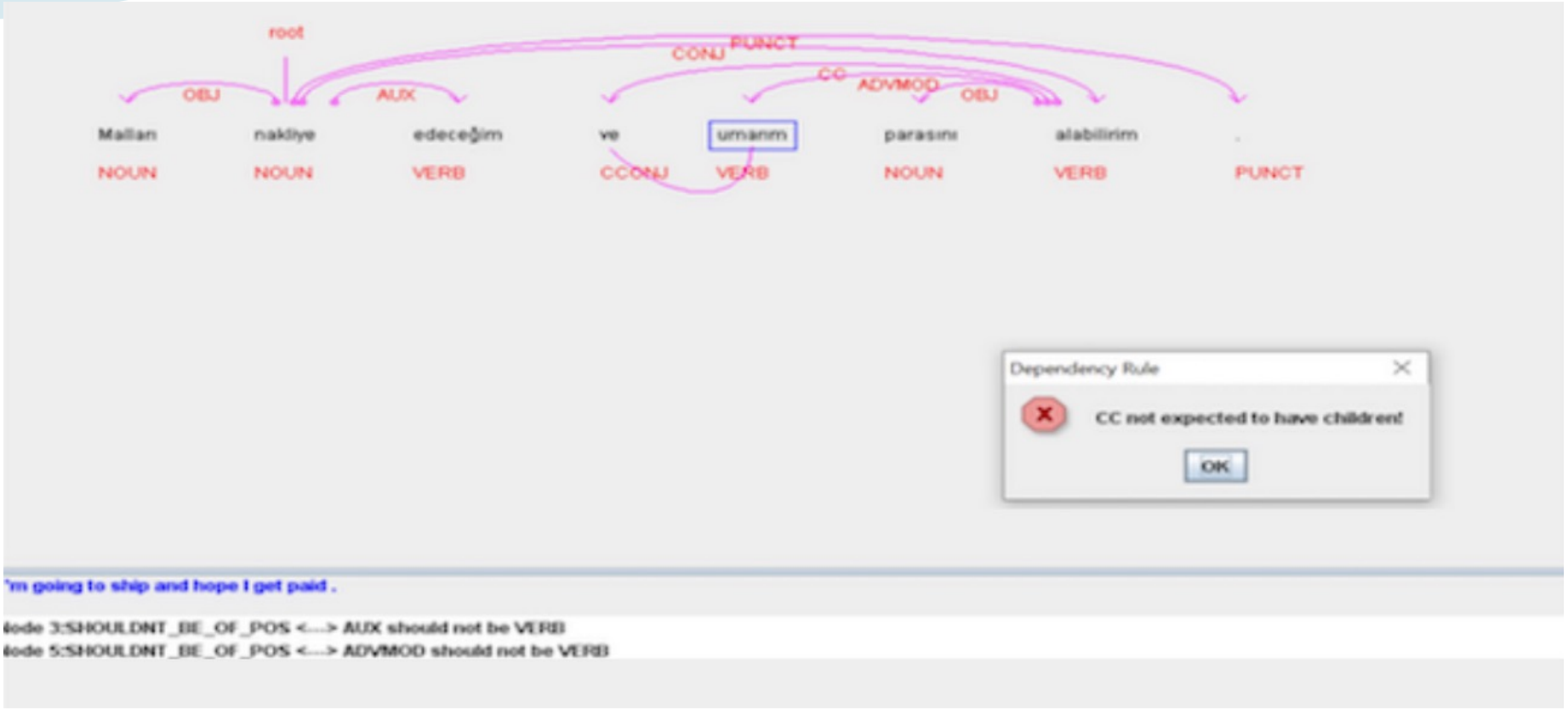


FileName	Index	Word	Morphological Analysis	Sentence
0000.dev	1	Devasa	devasa+ADJ	Devasa ölçekli yeni kanunda kullanılan karmaşık ve çetrefilli dil kav...
0000.dev	2	ölçekli	ölçek+NOUN+A3SG+PNON+NOM^DB+ADJ+WITH	Devasa ölçekli yeni kanunda kullanılan karmaşık ve çetrefilli dil kav...
0000.dev	3	yeni	yeni+ADJ	Devasa ölçekli yeni kanunda kullanılan karmaşık ve çetrefilli dil kav...
0000.dev	4	kanunda	kanun+NOUN+A3SG+PNON+LOC	Devasa ölçekli yeni kanunda kullanılan karmaşık ve çetrefilli dil kav...
0000.dev	5	kullanılan	kullan+VERB^DB+VERB+PASS+POS^DB+ADJ+PRESPART	Devasa ölçekli yeni kanunda kullanılan karmaşık ve çetrefilli dil kav...
0000.dev	6	karmaşık	karmaşık+ADJ	Devasa ölçekli yeni kanunda kullanılan karmaşık ve çetrefilli dil kav...
0000.dev	7	ve	ve+CONJ	Devasa ölçekli yeni kanunda kullanılan karmaşık ve çetrefilli dil kav...
0000.dev	8	çetrefilli	çetrefil+NOUN+A3SG+PNON+NOM^DB+ADJ+WITH	Devasa ölçekli yeni kanunda kullanılan karmaşık ve çetrefilli dil kav...
0000.dev	9	dil	dil+NOUN+A3SG+PNON+NOM	Devasa ölçekli yeni kanunda kullanılan karmaşık ve çetrefilli dil kav...
0000.dev	10	kavgayı	kavga+NOUN+A3SG+PNON+ACC	Devasa ölçekli yeni kanunda kullanılan karmaşık ve çetrefilli dil kav...
0000.dev	11	bulandırdı	bulan+VERB^DB+VERB+CAUS+POS+PAST+A3SG	Devasa ölçekli yeni kanunda kullanılan karmaşık ve çetrefilli dil kav...
0000.dev	12	.	.+PUNC	Devasa ölçekli yeni kanunda kullanılan karmaşık ve çetrefilli dil kav...
0000.test	1	Hayır	hayır+ADV	Hayır , kara pazartesi değildi .
0000.test	2	,	,+PUNC	Hayır , kara pazartesi değildi .
0000.test	3	kara	kara+ADJ	Hayır , kara pazartesi değildi .
0000.test	4	pazartesi	pazartesi+NOUN+A3SG+PNON+NOM	Hayır , kara pazartesi değildi .
0000.test	5	değildi	değil+ADJ^DB+VERB+ZERO+PAST+A3SG	Hayır , kara pazartesi değildi .
0000.test	6	.	.+PUNC	Hayır , kara pazartesi değildi .
0000.train	1	Bayan	bayan+NOUN+A3SG+PNON+NOM	Bayan Haag Elanti oynuyor .
0000.train	2	Haag	haag+NOUN+PROP+A3SG+PNON+NOM	Bayan Haag Elanti oynuyor .
0000.train	3	Elanti	elanti+NOUN+PROP+A3SG+PNON+NOM	Bayan Haag Elanti oynuyor .
0000.train	4	oynuyor	oyna+VERB+POS+PROG1+A3SG	Bayan Haag Elanti oynuyor .
0000.train	5	.	.+PUNC	Bayan Haag Elanti oynuyor .

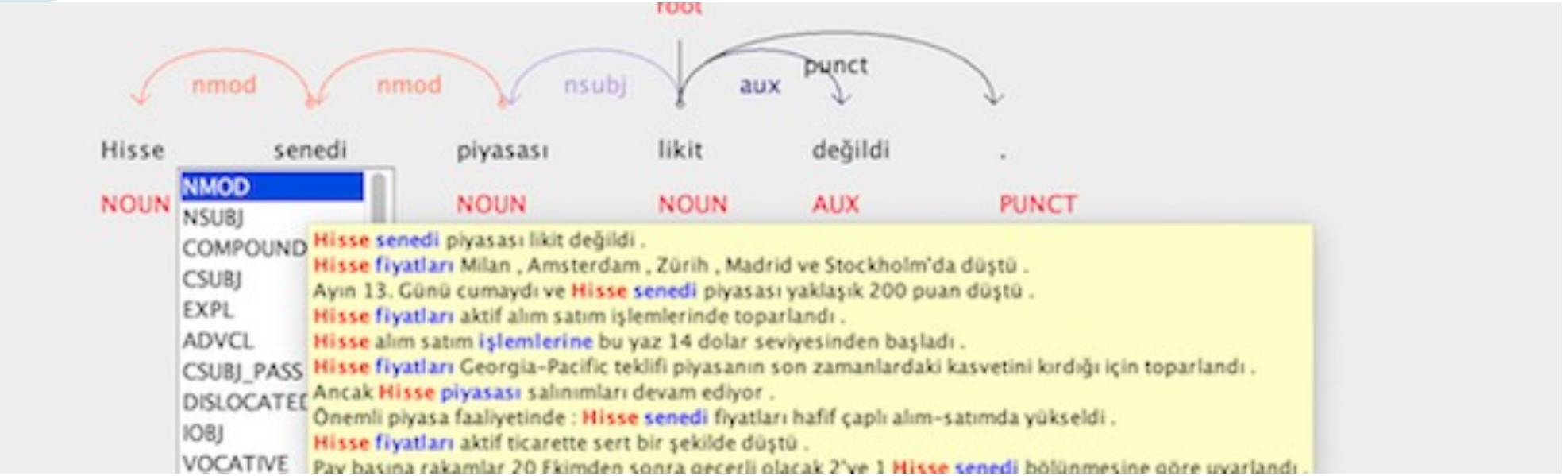
Dependency Annotation



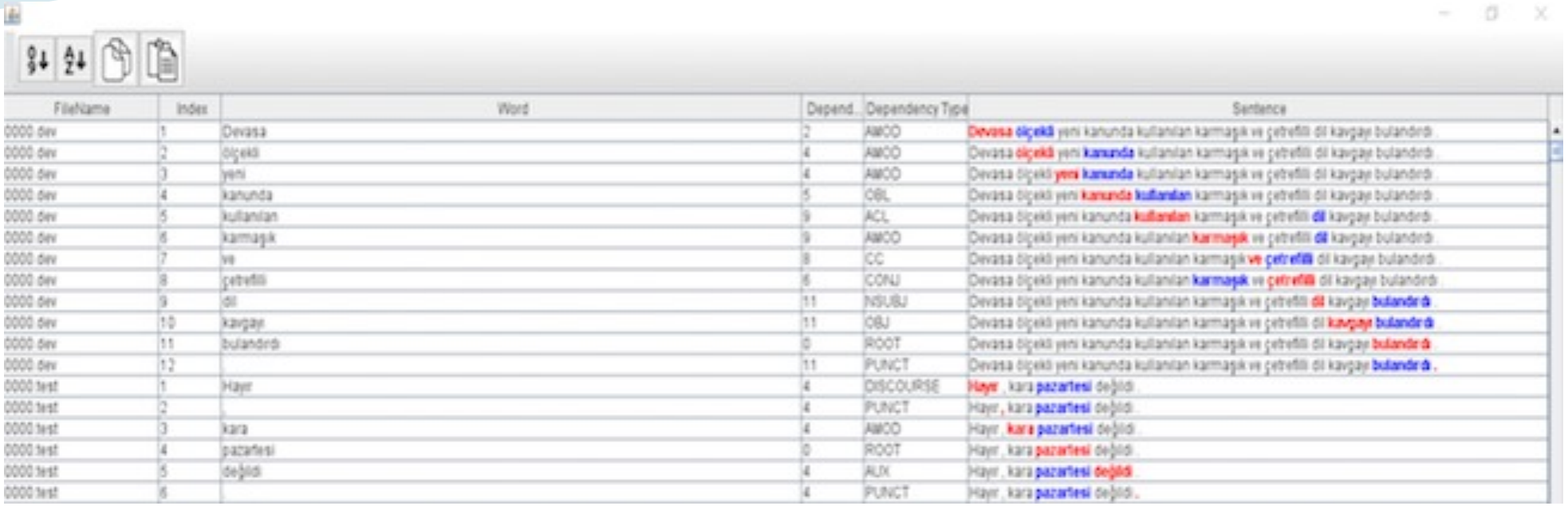
Error Check


root
CONJ
PUNCT
CC
ADVMOD
OBJ
OBJ
AUX
Malları
nakliye
edeceğim
ve
umarım
parasını
alabilirim
.
NOUN
NOUN
VERB
CCONJ
VERB
NOUN
VERB
PUNCT
Dependency Rule
CC not expected to have children!
OK
'm going to ship and hope I get paid .
lode 3:SHOULDNT_BE_OF_POS <---> AUX should not be VERB
lode 5:SHOULDNT_BE_OF_POS <---> ADVMOD should not be VERB

Annotation Examples



Annotation Comparison



FileName	Index	Word	Depend.	Dependency Type	Sentence
0000 dev	1	Devasa	2	AMOD	Devasa **ölçekli** yeni kanunda kullanılan karmaşık ve çetrefilli dil kavgayı bulandırdı .
0000 dev	2	ölçekli	4	AMOD	Devasa **ölçekli** yeni **kanunda** kullanılan karmaşık ve çetrefilli dil kavgayı bulandırdı .
0000 dev	3	yeni	4	AMOD	Devasa ölçekli **yeni kanunda** kullanılan karmaşık ve çetrefilli dil kavgayı bulandırdı .
0000 dev	4	kanunda	5	OBL	Devasa ölçekli yeni **kanunda kullanılan** karmaşık ve çetrefilli dil kavgayı bulandırdı .
0000 dev	5	kullanılan	9	ACL	Devasa ölçekli yeni kanunda **kullanılan** karmaşık ve çetrefilli **dil** kavgayı bulandırdı .
0000 dev	6	karmaşık	9	AMOD	Devasa ölçekli yeni kanunda kullanılan **karmaşık** ve çetrefilli **dil** kavgayı bulandırdı .
0000 dev	7	ve	8	CC	Devasa ölçekli yeni kanunda kullanılan karmaşık **ve** **çetrefilli** dil kavgayı bulandırdı .
0000 dev	8	çetrefilli	6	CONJ	Devasa ölçekli yeni kanunda kullanılan **karmaşık** ve **çetrefilli** dil kavgayı bulandırdı .
0000 dev	9	dil	11	NSUBJ	Devasa ölçekli yeni kanunda kullanılan karmaşık ve çetrefilli **dil** kavgayı **bulandırdı** .
0000 dev	10	kavgayı	11	OBJ	Devasa ölçekli yeni kanunda kullanılan karmaşık ve çetrefilli dil **kavgayı** **bulandırdı** .
0000 dev	11	bulandırdı	0	ROOT	Devasa ölçekli yeni kanunda kullanılan karmaşık ve çetrefilli dil kavgayı **bulandırdı** .
0000 dev	12	.	11	PUNCT	Devasa ölçekli yeni kanunda kullanılan karmaşık ve çetrefilli dil kavgayı **bulandırdı** **.**
0000 test	1	Hayır	4	DISCOURSE	**Hayır** , kara **pazartesi** değildi .
0000 test	2	,	4	PUNCT	Hayır **,** kara **pazartesi** değildi .
0000 test	3	kara	4	AMOD	Hayır , **kara** **pazartesi** değildi .
0000 test	4	pazartesi	0	ROOT	Hayır , kara **pazartesi** değildi .
0000 test	5	değildi	4	AUX	Hayır , kara **pazartesi** **değildi** .
0000 test	6	.	4	PUNCT	Hayır , kara **pazartesi** değildi **.**

Planned TreeBanks (2.13 or 2.14)

- QuestionBank
  - 10.000 Questions taken from SQUAD data set.
  - Parallel TreeBank with English also annotated.
  - Will be largest Question Treebank in UD.
  - 73.579 tokens in Turkish, 99.834 tokens in English
- Blogs
  - 11.128 Blog Sentences in 28 Topics (About 400 sentences per topic)
  - 141.567 Tokens

Future Work

- Extending dependency with extra layers:
  - Semantic layer
  - PropBank layer
  - FrameNet layer
  - NER layer
  - Shallow Parse layer
- Conversion to other representations
  - AMR (Abstract Meaning Representation)
  - Constituency treebanks

Questions?