Categorial Type Logics and Italian Corpora

RAFFAELLA BERNARDI

FREE UNIVERSITY OF BOLZANO-BOZEN

JOINT WORK WITH: A. BOLOGNESI, S. ROMAGNOLI, C. SEIDENARI, L. SURACE, F. TAMBURINI

Contents

1 Project: Italian Corpus Annotation 3
2 PoS tagging 4
3 Comparison 5
4 Distributional Method: Words 6
5 Distributional Method: Tags 7
6 Distributional Method: Structures 8
7 Proposal 9
8 Inducing and Clustering CTAs 10
9 Which linguistic information can we exploit? 11
10 From Treebank to PoS Classification 12
11 Turin University Treebank (TUT) 13
12 TUT representation format 14
13 TUT example 15
14 Grammatical Relation 16
15 Functor Argument (F-A) structures 17
16 Multimodal Composition 18
17 TUT simplified trees 19

18 Multimodal F-A structures 20
19 Type Resolution 21
20 Clustering 22
21 Further Research 23
22 Questions 24

1. Project: Italian Corpus Annotation

- **Project** carried out at the University of Bologna (CILTA);

1. Project: Italian Corpus Annotation

- **Project** carried out at the University of Bologna (CILTA);
- **Corpus** 100-million-words synchronic corpus of contemporary Italian (CORIS);

1. Project: Italian Corpus Annotation

- **Project** carried out at the University of Bologna (CILTA);
- **Corpus** 100-million-words synchronic corpus of contemporary Italian (CORIS);
- **Deliverables** part-of-speech tagging for the complete corpus, and (possibly) in a later stage syntactic analysis for a subcorpus;

1. Project: Italian Corpus Annotation

- **Project** carried out at the University of Bologna (CILTA);
- **Corpus** 100-million-words synchronic corpus of contemporary Italian (CORIS);
- **Deliverables** part-of-speech tagging for the complete corpus, and (possibly) in a later stage syntactic analysis for a subcorpus;
- **Period** 3 year project. Start of the linguistic annotation task: January 2004.

1. Project: Italian Corpus Annotation

- **Project** carried out at the University of Bologna (CILTA);
- **Corpus** 100-million-words synchronic corpus of contemporary Italian (CORIS);
- **Deliverables** part-of-speech tagging for the complete corpus, and (possibly) in a later stage syntactic analysis for a subcorpus;
- **Period** 3 year project. Start of the linguistic annotation task: January 2004.

1. Project: Italian Corpus Annotation

- **Project** carried out at the University of Bologna (CILTA);
- **Corpus** 100-million-words synchronic corpus of contemporary Italian (CORIS);
- **Deliverables** part-of-speech tagging for the complete corpus, and (possibly) in a later stage syntactic analysis for a subcorpus;
- **Period** 3 year project. Start of the linguistic annotation task: January 2004.

Corpora

- **Beauty** Real data vs. linguists'data;

1. Project: Italian Corpus Annotation

- **Project** carried out at the University of Bologna (CILTA);
- **Corpus** 100-million-words synchronic corpus of contemporary Italian (CORIS);
- **Deliverables** part-of-speech tagging for the complete corpus, and (possibly) in a later stage syntactic analysis for a subcorpus;
- **Period** 3 year project. Start of the linguistic annotation task: January 2004.

Corpora

- **Beauty** Real data vs. linguists'data;
- **Essential** tool for any study on Natural Languages to provide empirical support to theories and applications.

2. PoS tagging

- **Aim** Part to Speech (PoS) tagging of CORIS.

2. PoS tagging

- **Aim** Part to Speech (PoS) tagging of CORIS.
- **Question** Which PoS classification should we use?

2. PoS tagging

- **Aim** Part to Speech (PoS) tagging of CORIS.
- **Question** Which PoS classification should we use?
- **Other Projects**

2. PoS tagging

- **Aim** Part to Speech (PoS) tagging of CORIS.
- **Question** Which PoS classification should we use?
- **Other Projects**
  - Xerox, Grenoble (France)
  - Delmonte, Venezia (Italy)
  - TUT, Torino (Italy)

2. PoS tagging

- **Aim** Part to Speech (PoS) tagging of CORIS.
- **Question** Which PoS classification should we use?
- **Other Projects**
  - Xerox, Grenoble (France)
  - Delmonte, Venezia (Italy)
  - TUT, Torino (Italy)
- **Standards** EAGLES project, guidelines by Monachini.

2. PoS tagging

- **Aim** Part to Speech (PoS) tagging of CORIS.
- **Question** Which PoS classification should we use?
- **Other Projects**
  - Xerox, Grenoble (France)
  - Delmonte, Venezia (Italy)
  - TUT, Torino (Italy)
- **Standards** EAGLES project, guidelines by Monachini.
- **Question** How much do these classifications depend on linguistic-theories? Would the tagging satisfy the original purpose of Corpus annotation (to provide empirical support to NL applications)?

3. Comparison

- **Agreement** on the main PoS tags: nouns, verbs, adjectives, determiners, articles, adverbs, prepositions, conjunctions, numerals, interjections, punctuation and a class of residual items.

3. Comparison

- **Agreement** on the main PoS tags: nouns, verbs, adjectives, determiners, articles, adverbs, prepositions, conjunctions, numerals, interjections, punctuation and a class of residual items.

- **Disagreement** on the classification within the main PoS tags. For instance, "molti luoghi diversi" - many different places- "molti" (many) is considered

3. Comparison

- **Agreement** on the main PoS tags: nouns, verbs, adjectives, determiners, articles, adverbs, prepositions, conjunctions, numerals, interjections, punctuation and a class of residual items.

- **Disagreement** on the classification within the main PoS tags. For instance, "molti luoghi diversi" - many different places- "molti" (many) is considered

  - an Indefinite DETERMINER in Monachini

3. Comparison

- **Agreement** on the main PoS tags: nouns, verbs, adjectives, determiners, articles, adverbs, prepositions, conjunctions, numerals, interjections, punctuation and a class of residual items.

- **Disagreement** on the classification within the main PoS tags. For instance, "molti luoghi diversi" - many different places- "molti" (many) is considered

  - an Indefinite DETERMINER in Monachini
  - a Plural QUANTIFIER in Xerox, and

3. Comparison

- **Agreement** on the main PoS tags: nouns, verbs, adjectives, determiners, articles, adverbs, prepositions, conjunctions, numerals, interjections, punctuation and a class of residual items.

- **Disagreement** on the classification within the main PoS tags. For instance, "molti luoghi diversi" - many different places- "molti" (many) is considered

  - an Indefinite DETERMINER in Monachini
  - a Plural QUANTIFIER in Xerox, and
  - Indefinite ADJECTIVE in Delmonte and TUT.

3. Comparison

- **Agreement** on the main PoS tags: nouns, verbs, adjectives, determiners, articles, adverbs, prepositions, conjunctions, numerals, interjections, punctuation and a class of residual items.
- **Disagreement** on the classification within the main PoS tags. For instance, "molti luoghi diversi" - many different places- "molti" (many) is considered
  - an Indefinite DETERMINER in Monachini
  - a Plural QUANTIFIER in Xerox, and
  - Indefinite ADJECTIVE in Delmonte and TUT.
- **Proposal** To follow a bottom-up approach and deduce the PoS classification from empirical data by considering the distributional behavior of words.

4. Distributional Method: Words

- **Aim** To examine the distributional behaviour of some target words we can compare the lexical distribution of their contexts [Harris (1951), Kiss (1973), Brill (1993)]:

4. Distributional Method: Words

- **Aim** To examine the distributional behaviour of some target words we can compare the lexical distribution of their contexts [Harris (1951), Kiss (1973), Brill (1993)]:

...	...	il	babbo	gioca	...	dad plays
...	macchina	del	babbo	...	...	car of dad

4. Distributional Method: Words

- **Aim** To examine the distributional behaviour of some target words we can compare the lexical distribution of their contexts [Harris (1951), Kiss (1973), Brill (1993)]:

...	...	il	babbo	gioca	...	dad plays
...	macchina	del	babbo	...	...	car of dad
...	...	il	nonno	gioca	...	grandfather plays
...	macchina	del	nonno	...	...	car of grandfather

4. Distributional Method: Words

- **Aim** To examine the distributional behaviour of some target words we can compare the lexical distribution of their contexts [Harris (1951), Kiss (1973), Brill (1993)]:

...	...	il	babbo	gioca	...	dad plays
...	macchina	del	babbo	...	...	car of dad
...	...	il	nonno	gioca	...	grandfather plays
...	macchina	del	nonno	...	...	car of grandfather

- **Result** Using this method on Italian four different categories are obtained: Verbs (V), Nouns (N), Adjectives (Adj) and Grammatical Words (GW). [Tamburini et. ali (2000)]

4. Distributional Method: Words

- **Aim** To examine the distributional behaviour of some target words we can compare the lexical distribution of their contexts [Harris (1951), Kiss (1973), Brill (1993)]:

...	...	il	babbo	gioca	...	dad plays
...	macchina	del	babbo	...	...	car of dad
...	...	il	nonno	gioca	...	grandfather plays
...	macchina	del	nonno	...	...	car of grandfather

- **Result** Using this method on Italian four different categories are obtained: Verbs (V), Nouns (N), Adjectives (Adj) and Grammatical Words (GW). [Tamburini et. ali (2000)]

- **Drawback** sparse data problem which inflates the GW category.

5. Distributional Method: Tags

- **First Solution** To solve this problem Tamburini et ali. (2002) applied Brill's method on tags, obtaining a more fine-grained analysis of GW. [Brown, P. et. ali (1992)]

5. Distributional Method: Tags

- **First Solution** To solve this problem Tamburini et ali. (2002) applied Brill's method on tags, obtaining a more fine-grained analysis of GW. [Brown, P. et. ali (1992)]

. . . non vedo mai nessuno . . . I never see anyone

5. Distributional Method: Tags

- **First Solution** To solve this problem Tamburini et ali. (2002) applied Brill's method on tags, obtaining a more fine-grained analysis of GW. [Brown, P. et. ali (1992)]

...	non	vedo	mai	nessuno	...	I never see anyone
...	...	vedo	sempre	qualcuno	...	I always see someone

5. Distributional Method: Tags

- **First Solution** To solve this problem Tamburini et ali. (2002) applied Brill's method on tags, obtaining a more fine-grained analysis of GW. [Brown, P. et. ali (1992)]

		V		GW		
...	non	vedo	mai	nessuno	...	I never see anyone
...	...	vedo	sempre	qualcuno	...	I always see someone

6. Distributional Method: Structures

- Relying on limited distributional contexts ($\pm$ 2 words), the method fails to manage linguistic phenomena involving larger chunks of language such as conjunctions.

6. Distributional Method: Structures

- Relying on limited distributional contexts ($\pm$ 2 words), the method fails to manage linguistic phenomena involving larger chunks of language such as conjunctions.

			GW	N		GW	N
la	mamma	incarta	il	regalo	per	il	babbo. . . ..
(the)	mum	wraps	the	gift	for	(the)	dad

6. Distributional Method: Structures

- Relying on limited distributional contexts ($\pm$ 2 words), the method fails to manage linguistic phenomena involving larger chunks of language such as conjunctions.

			GW	N		GW	N			
la	mamma	incarta	il	regalo	per	il	babbo	. . .	. . .	. . .
(the)	mum	wraps	the	gift	for	(the)	dad			
la	mamma	incarta	il	regalo	e	il	babbo	scrive	il	biglietto
(the)	mum	wraps	the	gift	and	(the)	dad	writes	the	greetings card

6. Distributional Method: Structures

- Relying on limited distributional contexts ($\pm$ 2 words), the method fails to manage linguistic phenomena involving larger chunks of language such as conjunctions.

			GW	N		GW	N			
la	mamma	incarta	il	regalo	per	il	babbo	...	...	...
(the)	mum	wraps	the	gift	for	(the)	dad			
la	mamma	incarta	il	regalo	e	il	babbo	scrive	il	biglietto
(the)	mum	wraps	the	gift	and	(the)	dad	writes	the	greetings card

- Hence

6. Distributional Method: Structures

- Relying on limited distributional contexts ($\pm$ 2 words), the method fails to manage linguistic phenomena involving larger chunks of language such as conjunctions.

			GW	N		GW	N			
la	mamma	incarta	il	regalo	per	il	babbo	...	...	...
(the)	mum	wraps the		gift	for	(the)	dad			
la	mamma	incarta	il	regalo	e	il	babbo	scrive	il	biglietto
(the)	mum	wraps the		gift	and	(the)	dad	writes the greetings card		

- Hence
  - With limited context "e" seems to act as "per"

6. Distributional Method: Structures

- Relying on limited distributional contexts ($\pm$ 2 words), the method fails to manage linguistic phenomena involving larger chunks of language such as conjunctions.

			GW	N		GW	N			
la	mamma	incarta	il	regalo	per	il	babbo	. . .	. . .	. . .
(the)	mum	wraps	the	gift	for	(the)	dad			
la	mamma	incarta	il	regalo	e	il	babbo	scrive	il	biglietto
(the)	mum	wraps	the	gift	and	(the)	dad	writes	the	greetings card

- Hence
  - With limited context “e” seems to act as “per”
  - Conjunctions may be clustered with prepositions.

6. Distributional Method: Structures

- Relying on limited distributional contexts ($\pm$ 2 words), the method fails to manage linguistic phenomena involving larger chunks of language such as conjunctions.

			GW	N		GW	N			
la	mamma	incarta	il	regalo	per	il	babbo	...	...	...
(the)	mum	wraps	the	gift	for	(the)	dad			
la	mamma	incarta	il	regalo	e	il	babbo	scrive	il	biglietto
(the)	mum	wraps	the	gift	and	(the)	dad	writes	the	greetings card

- Hence
  - With limited context "e" seems to act as "per"
  - Conjunctions may be clustered with prepositions.
- Tags carrying structural information could help overcome this problem.

7. Proposal

We propose to exploit the structural information carried out by Categorial Type Assignments (CTAs).

7. Proposal

We propose to exploit the structural information carried out by Categorial Type Assignments (CTAs).

For example, in the sentence:

7. Proposal

We propose to exploit the structural information carried out by Categorial Type Assignments (CTAs).

For example, in the sentence:

la mamma incarta il regalo per il babbo - mum wraps the gift for dad

7. Proposal

We propose to exploit the structural information carried out by Categorial Type Assignments (CTAs).

For example, in the sentence:

la mamma incarta il regalo per il babbo - mum wraps the gift for dad

per is a functor (head) which has the type pp/np

7. Proposal

We propose to exploit the structural information carried out by Categorial Type Assignments (CTAs).

For example, in the sentence:

la mamma incarta il regalo per il babbo - mum wraps the gift for dad

per is a functor (head) which has the type pp/np

In the sentence:

7. Proposal

We propose to exploit the structural information carried out by Categorial Type Assignments (CTAs).

For example, in the sentence:

la mamma incarta il regalo per il babbo - mum wraps the gift for dad

per is a functor (head) which has the type pp/np

In the sentence:

la mamma incarta il regalo e il babbo scrive il biglietto
mum wraps the gift and dad writes the greetings card

7. Proposal

We propose to exploit the structural information carried out by Categorial Type Assignments (CTAs).

For example, in the sentence:

la mamma incarta il regalo per il babbo - mum wraps the gift for dad

per is a functor (head) which has the type pp/np

In the sentence:

la mamma incarta il regalo e il babbo scrive il biglietto
mum wraps the gift and dad writes the greetings card

e is a functor which has the type $(s\backslash s)/s$

7. Proposal

We propose to exploit the structural information carried out by Categorial Type Assignments (CTAs).

For example, in the sentence:

la mamma incarta il regalo per il babbo - mum wraps the gift for dad

per is a functor (head) which has the type pp/np

In the sentence:

la mamma incarta il regalo e il babbo scrive il biglietto
mum wraps the gift and dad writes the greetings card

e is a functor which has the type $(s\backslash s)/s$

Therefore, categorial types clustering will properly distinguish prepositions from conjunction.

8. Inducing and Clustering CTAs

We need to

- Induce Categorial Type Assignments from "raw" data

8. Inducing and Clustering CTAs

We need to

- Induce Categorial Type Assignments from "raw" data
- Or better, from data enriched with linguistically neutral information

8. Inducing and Clustering CTAs

We need to

- Induce Categorial Type Assignments from “raw” data
- Or better, from data enriched with linguistically neutral information
- Apply the clustering algorithm on the obtained CTAs.

8. Inducing and Clustering CTAs

We need to

- Induce Categorial Type Assignments from “raw” data
- Or better, from data enriched with linguistically neutral information
- Apply the clustering algorithm on the obtained CTAs.

Note, a rather small number of highly frequent words should suffice for the present task [Brill (1993)].

9. Which linguistic information can we exploit?

- The only PoS tags could be the ones clustered via of the distributional approach.

9. Which linguistic information can we exploit?

- The only PoS tags could be the ones clustered via of the distributional approach.
- The grammatical relations that are less theory-driven are Head-Dependent (H-D) and Functor-Argument (F-A) relations.

9. Which linguistic information can we exploit?

- The only PoS tags could be the ones clustered via of the distributional approach.
- The grammatical relations that are less theory-driven are Head-Dependent (H-D) and Functor-Argument (F-A) relations.
- They way H-D and F-A relate can be used to identify different modes of composition. In particular, we have founded the following main classes of dependents

9. Which linguistic information can we exploit?

- The only PoS tags could be the ones clustered via of the distributional approach.
- The grammatical relations that are less theory-driven are Head-Dependent (H-D) and Functor-Argument (F-A) relations.
- They way H-D and F-A relate can be used to identify different modes of composition. In particular, we have founded the following main classes of dependents

    1. Arguments (ARG), H-D coincides with F-A;

9. Which linguistic information can we exploit?

- The only PoS tags could be the ones clustered via of the distributional approach.
- The grammatical relations that are less theory-driven are Head-Dependent (H-D) and Functor-Argument (F-A) relations.
- They way H-D and F-A relate can be used to identify different modes of composition. In particular, we have founded the following main classes of dependents

  1. Arguments (ARG), H-D coincides with F-A;
  2. Modifiers (RMOD), H-D does not coincide with F-A. RMOD are optional –they return the same category they compose with;

9. Which linguistic information can we exploit?

- The only PoS tags could be the ones clustered via of the distributional approach.
- The grammatical relations that are less theory-driven are Head-Dependent (H-D) and Functor-Argument (F-A) relations.
- They way H-D and F-A relate can be used to identify different modes of composition. In particular, we have founded the following main classes of dependents

    1. Arguments (ARG), H-D coincides with F-A;
    2. Modifiers (RMOD), H-D does not coincide with F-A. RMOD are optional –they return the same category they compose with;
    3. Auxiliaries (AUX), H-D does not coincide with F-A. AUX are indispensable for grammaticality since they modify the head verb;

9. Which linguistic information can we exploit?

- The only PoS tags could be the ones clustered via of the distributional approach.
- The grammatical relations that are less theory-driven are Head-Dependent (H-D) and Functor-Argument (F-A) relations.
- They way H-D and F-A relate can be used to identify different modes of composition. In particular, we have founded the following main classes of dependents

  1. Arguments (ARG), H-D coincides with F-A;
  2. Modifiers (RMOD), H-D does not coincide with F-A. RMOD are optional –they return the same category they compose with;
  3. Auxiliaries (AUX), H-D does not coincide with F-A. AUX are indispensable for grammaticality since they modify the head verb;
  4. Coordination (COORD), they are polymorphic ternary relations.

9. Which linguistic information can we exploit?

- The only PoS tags could be the ones clustered via of the distributional approach.
- The grammatical relations that are less theory-driven are Head-Dependent (H-D) and Functor-Argument (F-A) relations.
- They way H-D and F-A relate can be used to identify different modes of composition. In particular, we have founded the following main classes of dependents

  1. Arguments (ARG), H-D coincides with F-A;
  2. Modifiers (RMOD), H-D does not coincide with F-A. RMOD are optional –they return the same category they compose with;
  3. Auxiliaries (AUX), H-D does not coincide with F-A. AUX are indispensable for grammaticality since they modify the head verb;
  4. Coordination (COORD), they are polymorphic ternary relations.

- Based on these observations, information on H-D and F-A can be extracted from (dependency) treebanks.

10. From Treebank to PoS Classification

Given an Italian treebank

10. From Treebank to PoS Classification

Given an Italian treebank

1. **H-D relations**: we extract only the linguistically-neutral information on the Heads and their Dependents (ARG, RMOD, AUX, COORD);

10. From Treebank to PoS Classification

Given an Italian treebank

1. **H-D relations**: we extract only the linguistically-neutral information on the Heads and their Dependents (ARG, RMOD, AUX, COORD);
2. **F-A structures** we extract the functor-argument (F-A) structures;

10. From Treebank to PoS Classification

Given an Italian treebank

1. **H-D relations**: we extract only the linguistically-neutral information on the Heads and their Dependents (ARG, RMOD, AUX, COORD);

2. **F-A structures** we extract the functor-argument (F-A) structures;

3. **Types** we apply the Type Resolution algorithm proposed in [van Emden 88] and [Buszkowski, Penn 90] obtaining Categorial Type Assignments (CTAs);

10. From Treebank to PoS Classification

Given an Italian treebank

1. **H-D relations**: we extract only the linguistically-neutral information on the Heads and their Dependents (ARG, RMOD, AUX, COORD);
2. **F-A structures** we extract the functor-argument (F-A) structures;
3. **Types** we apply the Type Resolution algorithm proposed in [van Emden 88] and [Buszkowski, Penn 90] obtaining Categorial Type Assignments (CTAs);
4. **Clusters** we apply a distributional-syntactic clustering method on CTAs, obtaining empirical suggestions to PoS TAG sets.

10. From Treebank to PoS Classification

Given an Italian treebank

1. **H-D relations**: we extract only the linguistically-neutral information on the Heads and their Dependents (ARG, RMOD, AUX, COORD);
2. **F-A structures** we extract the functor-argument (F-A) structures;
3. **Types** we apply the Type Resolution algorithm proposed in [van Emden 88] and [Buszkowski, Penn 90] obtaining Categorial Type Assignments (CTAs);
4. **Clusters** we apply a distributional-syntactic clustering method on CTAs, obtaining empirical suggestions to PoS TAG sets.

$$\begin{array}{ccccc} \text{Treebank} & \xrightarrow[meta-data]{extraction\ of} & \text{Meta-Treebank} & \xrightarrow{conversion} & \text{F-A structures} \\ & & & & \downarrow{\scriptstyle typeresolution} \\ & & \text{TAG set} & \xleftarrow[clustering]{} & \text{Categorial Types} \end{array}$$

11. Turin University Treebank (TUT)

The only available Italian Treebank is TUT.

11. Turin University Treebank (TUT)

The only available Italian Treebank is TUT.

- It is a collection of syntactically annotated Italian sentences;

11. Turin University Treebank (TUT)

The only available Italian Treebank is TUT.

- It is a collection of syntactically annotated Italian sentences;
- it's rather small. It consists of 38,653 words and 1,500 sentences;

11. Turin University Treebank (TUT)

The only available Italian Treebank is TUT.

- It is a collection of syntactically annotated Italian sentences;
- it's rather small. It consists of 38,653 words and 1,500 sentences;
- it's a dependency treebank.

11. Turin University Treebank (TUT)

The only available Italian Treebank is TUT.

- It is a collection of syntactically annotated Italian sentences;
- it's rather small. It consists of 38,653 words and 1,500 sentences;
- it's a dependency treebank.

There is also ISST

11. Turin University Treebank (TUT)

The only available Italian Treebank is TUT.

- It is a collection of syntactically annotated Italian sentences;
- it's rather small. It consists of 38,653 words and 1,500 sentences;
- it's a dependency treebank.

There is also ISST

- It's a multi-layered corpus, annotated at the syntactic and lexico-semantic levels;
- it has a user interface to explore the corpus;

11. Turin University Treebank (TUT)

The only available Italian Treebank is TUT.

- It is a collection of syntactically annotated Italian sentences;
- it's rather small. It consists of 38,653 words and 1,500 sentences;
- it's a dependency treebank.

There is also ISST

- It's a multi-layered corpus, annotated at the syntactic and lexico-semantic levels;
- it has a user interface to explore the corpus;
- it counts 305,547 word tokens. But

11. Turin University Treebank (TUT)

The only available Italian Treebank is TUT.

- It is a collection of syntactically annotated Italian sentences;
- it's rather small. It consists of 38,653 words and 1,500 sentences;
- it's a dependency treebank.

There is also ISST

- It's a multi-layered corpus, annotated at the syntactic and lexico-semantic levels;
- it has a user interface to explore the corpus;
- it counts 305,547 word tokens. But
- it's not (freely) available.

12. TUT representation format

In TUT trees:

12. TUT representation format

In TUT trees:

- each node is labelled by a word;

12. TUT representation format

In TUT trees:

- each node is labelled by a word;
- each arch is labelled by a grammatical relation.

12. TUT representation format

In TUT trees:

- each node is labelled by a word;
- each arch is labelled by a grammatical relation.

The information concerning a single node word is given as below:

12. TUT representation format

In TUT trees:

- each node is labelled by a word;
- each arch is labelled by a grammatical relation.

The information concerning a single node word is given as below:

$$n \text{ word } (f_1 \ f_2 \ldots f_n) \ [H;MORPH - SYNT - SEM]$$

12. TUT representation format

In TUT trees:

- each node is labelled by a word;
- each arch is labelled by a grammatical relation.

The information concerning a single node word is given as below:

$$n \text{ word } (f_1\ f_2\ \ldots f_n)\ [H;MORPH - SYNT - SEM]$$

- n is the number of the linear order of the word occurrence;

12. TUT representation format

In TUT trees:

- each node is labelled by a word;
- each arch is labelled by a grammatical relation.

The information concerning a single node word is given as below:

$$n \text{ word } (f_1\ f_2 \ldots f_n)\ [H; MORPH - SYNT - SEM]$$

- n is the number of the linear order of the word occurrence;
- f_i are morphological features associated with the word itself;

12. TUT representation format

In TUT trees:

- each node is labelled by a word;
- each arch is labelled by a grammatical relation.

The information concerning a single node word is given as below:

$$n \text{ word } (f_1\ f_2 \ldots f_n)\ [H; MORPH - SYNT - SEM]$$

- n is the number of the linear order of the word occurrence;
- f_i are morphological features associated with the word itself;
- $MORPH - SYNT - SEM$ are the grammatical relations concerning the dependency edge linking the word with its syntactic head (H).

13. TUT example

```
**************  FRASE ALB-71  **************
1  I (IL ART DEF M PL)
                [6;VERB-SUBJ]
2  primi (PRIMO ADJ ORDIN M PL)
                [3;ADJC+ORDIN-RMOD]
3  approcci (APPROCCIO NOUN COMMON M PL)
                [1;DET+DEF-ARG]
4  non (NON ADV NEG)
                [6;ADVB-RMOD]
5  sono (ESSERE VERB AUX IND PRES INTR 3 PL)
                [6;AUX+TENSE]
6  stati (ESSERE VERB MAIN PART PAST INTR PL M)
                [0;TOP-VERB]
7  esaltanti (ESALTANTE ADJ QUALIF ALLVAL PL)
                [6;VERB-PREDCOMPL+SUBJ]
8  . (#\. PUNCT) [6;END]
```

14. Grammatical Relation

- **Aim** We want to extract from TUT only (as far as possible) linguistically neutral information.

14. Grammatical Relation

- **Aim** We want to extract from TUT only (as far as possible) linguistically neutral information.
- **Basic H-D relation** We can focus on the SYNT (functional-syntactic) component of the TUT annotation.

14. Grammatical Relation

- **Aim** We want to extract from TUT only (as far as possible) linguistically neutral information.
- **Basic H-D relation** We can focus on the SYNT (functional-syntactic) component of the TUT annotation.
- **Hierarchy of Dependents** Dependents are divided into a hierarchy reducing to a few main ones. ARG (e.g. sublabels: SUBJ, OBJ, INDOBJ, INDCOMPL, PREDCOMPL) and RMOD on the one hand, and AUX, COORD [see Bosco 2003].

15. Functor Argument (F-A) structures

We want to convert the meta-treebank into F-A structures [Buszkowski, Penn '90].

15. Functor Argument (F-A) structures

We want to convert the meta-treebank into F-A structures [Buszkowski, Penn '90].

- F-A structures are binary branching trees;

15. Functor Argument (F-A) structures

We want to convert the meta-treebank into F-A structures [Buszkowski, Penn '90].

- F-A structures are binary branching trees;
- The leaf nodes are labelled by lexical expressions (words);

15. Functor Argument (F-A) structures

We want to convert the meta-treebank into F-A structures [Buszkowski, Penn '90].

- F-A structures are binary branching trees;
- The leaf nodes are labelled by lexical expressions (words);
- The internal nodes are labelled by $\triangleleft$ (for structures with the functor as the left daughter) or $\triangleright$ (for structures with the functor as the right daughter).

16. Multimodal Composition

Following [Moortgat and Morrill (1991)] we treat functor-argument and head-dependency relations as orthogonal dimensions of linguistic composition and use different modes.

16. Multimodal Composition

Following [Moortgat and Morrill (1991)] we treat functor-argument and head-dependency relations as orthogonal dimensions of linguistic composition and use different modes.

	ad	ah	fd	fh
fh	$\blacktriangleleft$			
fd		$\triangleleft$		
ah			$\triangleright$	
ad				$\blacktriangleright$

17. TUT simplified trees

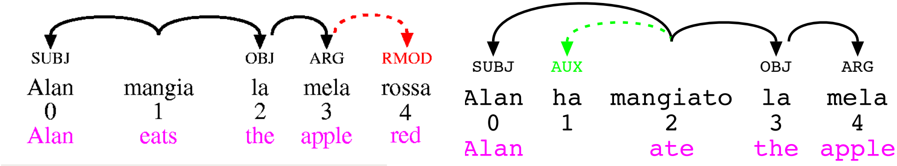

Figure 1: MOD and AUX: Functors as Dependents

18. Multimodal F-A structures

From TUT trees we obtain

18. Multimodal F-A structures

From TUT trees we obtain

- Allen $\blacktriangleright$ (mangia $\blacktriangleleft$ (la $\blacktriangleleft$ mela)

18. Multimodal F-A structures

From TUT trees we obtain

- Allen $\blacktriangleright$ (mangia $\blacktriangleleft$ (la $\blacktriangleleft$ mela)
- Allen $\blacktriangleright$ (mangia $\blacktriangleleft$ (la $\blacktriangleleft$ (mela $\triangleright$ rossa))

18. Multimodal F-A structures

From TUT trees we obtain

- Allen $\blacktriangleright$ (mangia $\blacktriangleleft$ (la $\blacktriangleleft$ mela)
- Allen $\blacktriangleright$ (mangia $\blacktriangleleft$ (la $\blacktriangleleft$ (mela $\triangleright$ rossa))
- Allen $\blacktriangleright$ ((ha $\triangleleft$ mangiato) $\blacktriangleleft$ (la $\blacktriangleleft$ mela))

19. Type Resolution

- To adapt the type resolution algorithm to the multimodal system is rather straightforward.

19. Type Resolution

- To adapt the type resolution algorithm to the multimodal system is rather straightforward.
- In the above example, fixing the goal type for these examples as s and using as only clustered set the ones of NOUN (n), we obtain the following type assignments:

19. Type Resolution

- To adapt the type resolution algorithm to the multimodal system is rather straightforward.

- In the above example, fixing the goal type for these examples as s and using as only clustered set the ones of NOUN (n), we obtain the following type assignments:

Allan	A
mangia	$(A \multimap\!\!\bullet\, s) \bullet\!\!\multimap B$
la	$B \bullet\!\!\multimap n$
mela	n
rossa	$n \multimap\!\!\circ\, n$
ha	$((A \multimap\!\!\bullet\, s) \bullet\!\!\multimap B) \circ\!\!\multimap D$
mangiato	D

20. Clustering

CTAs are trees. To cluster them we can apply a Tree Pattern Matching Algorithm [Shasha and Zhang '97]

20. Clustering

CTAs are trees. To cluster them we can apply a Tree Pattern Matching Algorithm [Shasha and Zhang '97]

- Tree-rewriting:

20. Clustering

CTAs are trees. To cluster them we can apply a Tree Pattern Matching Algorithm [Shasha and Zhang '97]

- Tree-rewriting:
  - Renaming;

20. Clustering

CTAs are trees. To cluster them we can apply a Tree Pattern Matching Algorithm [Shasha and Zhang '97]

- Tree-rewriting:
  - Renaming;
  - Deletion;

20. Clustering

CTAs are trees. To cluster them we can apply a Tree Pattern Matching Algorithm [Shasha and Zhang '97]

- Tree-rewriting:
  - Renaming;
  - Deletion;
  - Edit.

20. Clustering

CTAs are trees. To cluster them we can apply a Tree Pattern Matching Algorithm [Shasha and Zhang '97]

- Tree-rewriting:
  - Renaming;
  - Deletion;
  - Edit.
- Which is the weight of each operation?

20. Clustering

CTAs are trees. To cluster them we can apply a Tree Pattern Matching Algorithm [Shasha and Zhang '97]

- Tree-rewriting:
  - Renaming;
  - Deletion;
  - Edit.
- Which is the weight of each operation?
  - Renaming: Changing H-D relation cost more than changing F-A order;

20. Clustering

CTAs are trees. To cluster them we can apply a Tree Pattern Matching Algorithm [Shasha and Zhang '97]

- Tree-rewriting:
  - Renaming;
  - Deletion;
  - Edit.
- Which is the weight of each operation?
  - Renaming: Changing H-D relation cost more than changing F-A order;
  - Renaming: Replacing Variables/Constants, Con/Con, Var/Var;

20. Clustering

CTAs are trees. To cluster them we can apply a Tree Pattern Matching Algorithm [Shasha and Zhang '97]

- Tree-rewriting:
  - Renaming;
  - Deletion;
  - Edit.
- Which is the weight of each operation?
  - Renaming: Changing H-D relation cost more than changing F-A order;
  - Renaming: Replacing Variables/Constants, Con/Con, Var/Var;
  - Deletion and Edit: deleting (editing) a connective costs is tied to deleting (editing) a con/var. But

20. Clustering

CTAs are trees. To cluster them we can apply a Tree Pattern Matching Algorithm [Shasha and Zhang '97]

- Tree-rewriting:
  - Renaming;
  - Deletion;
  - Edit.
- Which is the weight of each operation?
  - Renaming: Changing H-D relation cost more than changing F-A order;
  - Renaming: Replacing Variables/Constants, Con/Con, Var/Var;
  - Deletion and Edit: deleting (editing) a connective costs is tied to deleting (editing) a con/var. But
  - Deletion and Edit: How do they relate to renaming?

21. Further Research

- On the conversion from TUT:

21. Further Research

- On the conversion from TUT:
  - TUT uses traces. Should we remove them? How much would the resulting clusters differ?

21. Further Research

- On the conversion from TUT:
  - TUT uses traces. Should we remove them? How much would the resulting clusters differ?
  - In TUT anything can be a Top-formula. Should we leave it like this?

21. Further Research

- On the conversion from TUT:
  - TUT uses traces. Should we remove them? How much would the resulting clusters differ?
  - In TUT anything can be a Top-formula. Should we leave it like this?
  - What is the role of Higher Order Types in this procedure? Can we use them to see how long distance dependency triggers gather together?

21. Further Research

- On the conversion from TUT:
  - TUT uses traces. Should we remove them? How much would the resulting clusters differ?
  - In TUT anything can be a Top-formula. Should we leave it like this?
  - What is the role of Higher Order Types in this procedure? Can we use them to see how long distance dependency triggers gather together?
- On the Clustering:

21. Further Research

- On the conversion from TUT:
  - TUT uses traces. Should we remove them? How much would the resulting clusters differ?
  - In TUT anything can be a Top-formula. Should we leave it like this?
  - What is the role of Higher Order Types in this procedure? Can we use them to see how long distance dependency triggers gather together?
- On the Clustering:
  - Does the tree clustering algorithm reduce to weight structural rules?

21. Further Research

- On the conversion from TUT:
  - TUT uses traces. Should we remove them? How much would the resulting clusters differ?
  - In TUT anything can be a Top-formula. Should we leave it like this?
  - What is the role of Higher Order Types in this procedure? Can we use them to see how long distance dependency triggers gather together?
- On the Clustering:
  - Does the tree clustering algorithm reduce to weight structural rules?
  - Can derivability relations among types help cleaning up clusters and reach the right level of similarity trees?

21. Further Research

- On the conversion from TUT:
  - TUT uses traces. Should we remove them? How much would the resulting clusters differ?
  - In TUT anything can be a Top-formula. Should we leave it like this?
  - What is the role of Higher Order Types in this procedure? Can we use them to see how long distance dependency triggers gather together?
- On the Clustering:
  - Does the tree clustering algorithm reduce to weight structural rules?
  - Can derivability relations among types help cleaning up clusters and reach the right level of similarity trees?
  - Elementary trees of TAG have been induced by TUT [A. Mazzei]. Would it make sense to compare clustering of TAGs trees and CTAs?

21. Further Research

- On the conversion from TUT:
  - TUT uses traces. Should we remove them? How much would the resulting clusters differ?
  - In TUT anything can be a Top-formula. Should we leave it like this?
  - What is the role of Higher Order Types in this procedure? Can we use them to see how long distance dependency triggers gather together?
- On the Clustering:
  - Does the tree clustering algorithm reduce to weight structural rules?
  - Can derivability relations among types help cleaning up clusters and reach the right level of similarity trees?
  - Elementary trees of TAG have been induced by TUT [A. Mazzei]. Would it make sense to compare clustering of TAGs trees and CTAs?
  - Is the rather small size of the treebank a limit for this study?

22. Questions

On the approach:

22. Questions

On the approach:

- Can this study help reaching a further understanding of structural rules in natural language analysis?

22. Questions

On the approach:

- Can this study help reaching a further understanding of structural rules in natural language analysis?
- Can this study help investigating the role of surface vs. deep structures? (vd. traces)

22. Questions

On the approach:

- Can this study help reaching a further understanding of structural rules in natural language analysis?
- Can this study help investigating the role of surface vs. deep structures? (vd. traces)
- How much are the result still empirically founded?

22. Questions

On the approach:

- Can this study help reaching a further understanding of structural rules in natural language analysis?
- Can this study help investigating the role of surface vs. deep structures? (vd. traces)
- How much are the result still empirically founded?
- What would we really learn from this study at the end?