

# Annotating Italian Social Media Texts in Universal Dependencies

M. Sanguinetti\*, C. Bosco\*, A. Mazzei\*, A. Lavelli\*\*, F. Tamburini\*\*\*

\*Dipartimento di Informatica,  
Università di Torino



\*\*Fondazione Bruno Kessler,  
Trento



\*\*\* FICLIT,  
Università di Bologna



# About

## PoSTWITA-UD:

a collection of Italian texts from Twitter annotated according to the **Universal Dependencies** format

### Goals:

- create a treebank of social media texts
- contribute to the wider debate about social media text processing and analysis

# Dataset

Developed by processing and further enriching the **PoSTWITA** corpus: the dataset used for the EVALITA 2016 task on Part-of-Speech tagging of social media

## Composition:

- dev set: 6,438 tweets (114,967 tokens)
- test set: 300 tweets (4,759 tokens)

# Trebank Preprocessing

## PoSTWITA

## PoSTWITA-UD

```
_____ 571947735294873600 _____  
@GScarcina MENTION  
Ciao INTJ  
@ EMO  
se SCONJ  
hai VERB  
posti NOUN  
liberi ADJ  
nel ADP_A  
gruppo NOUN  
sull' ADP_A  
app NOUN  
di ADP  
braccialetti PROPN  
Rossi PROPN  
puoi AUX  
aggiungermi VERB_CLIT  
? PUNCT
```

```
# sent_id = _____ 571947735294873600 _____  
# text = @GScarcina Ciao @ se hai posti liberi nel gruppo sull'app di braccialetti Rossi puoi aggiungermi?  
1 @GScarcina @GScarcina SYM SYM _ _ _ _  
2 Ciao ciao INTJ I _ _ _  
3 @ @ SYM SYM _ _ _ _  
4 se se SCONJ CS _ _ _ _  
5 hai avere VERB V Mood=Ind|Number=Sing|Person=2|Tense=Pres|VerbForm=Fin _ _ _ _  
6 posti posto NOUN S Gender=Masc|Number=Plur _ _ _ _  
7 liberi libero ADJ A Gender=Masc|Number=Plur _ _ _ _  
8-9 nel _ _ _ _ _ _ _ _  
8 in in ADP E _ _ _ _  
9 il il DET RD Definite=Def|Gender=Masc|Number=Sing|PronType=Art _ _ _ _  
10 gruppo gruppo NOUN S Gender=Masc|Number=Sing _ _ _ _  
11-12 sull' _ _ _ _ _ _ SpaceAfter=No  
11 su su ADP E _ _ _ _  
12 l' il DET RD Definite=Def|Number=Sing|PronType=Art _ _ _ _  
13 app app NOUN S _ _ _ _  
14 di di ADP E _ _ _ _  
15 braccialetti braccialetti PROPN SP _ _ _ _  
16 Rossi Rossi PROPN SP _ _ _ _  
17 puoi potere AUX VM Mood=Ind|Number=Sing|Person=2|Tense=Pres|VerbForm=Fin _ _ _ _  
18-19 aggiungermi _ _ _ _ _ _ SpaceAfter=No  
18 aggiunger aggiungere VERB V VerbForm=Inf _ _ _ _  
19 mi mi PRON PC Clitic=Yes|Number=Sing|Person=1|PronType=Prs _ _ _ _  
20 ? ? PUNCT FS _ _ _ _
```

all Internet-specific tags converted into SYM

multi-word tokens re-splitted

lemmas and morphological features added using AnIta\*

\*(Tamburini and Melandri, 2012)

# Parsing Experiments

- train:

UD\_Italian v.2 (11,699 sentences)

- test:

**1)** PoSTWITA test set (300 tweets)

*a)* with lemmas and language-specific tags (-LF)

*b)* with morphological features (-F)

**2)** UD\_Italian test set (489 sentences) (-UD)

- evaluation metric:  $LAS F_1$

# Parsing Experiments

- train:

UD\_Italian v.2 (11,699 sentences)

- test:

**1)** PoSTWITA test set (300 tweets)

*a)* with lemmas and language-specific tags (-LF)

*b)* with morphological features (-F)

**2)** UD\_Italian test set (489 sentences) (-UD)

- evaluation metric:  $LAS F_1$

Parser	-LX	-F	-UD
MATE graph-based	62.53	67.05	91.26
MATE transition-based	64.92	66.65	91.44
RBG full	64.36	67.07	90.16

# Annotation Guidelines

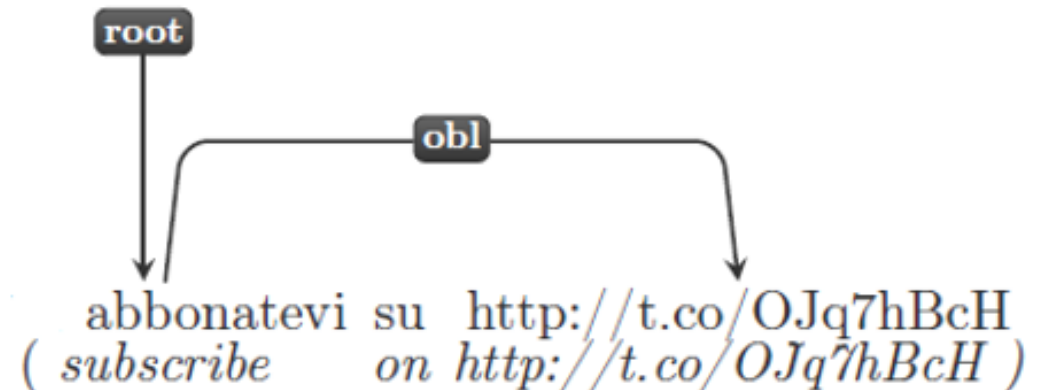
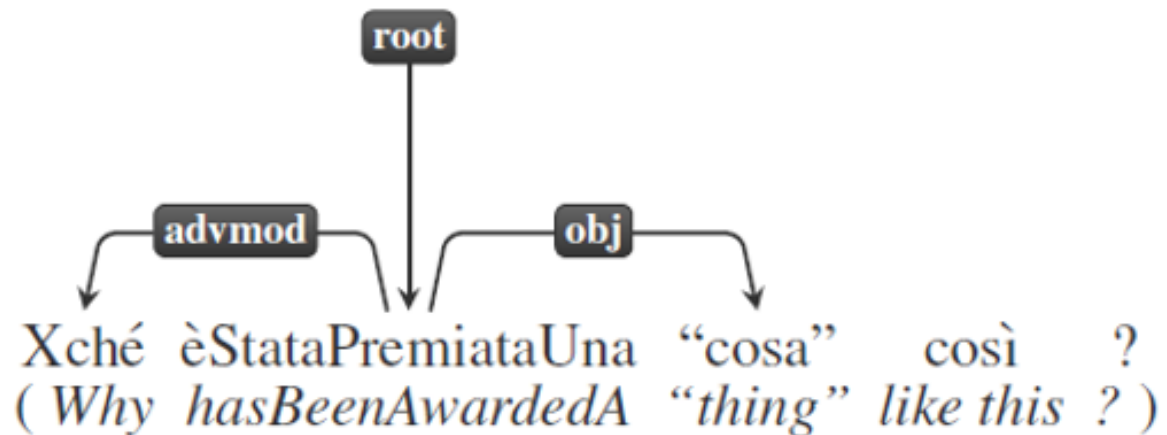
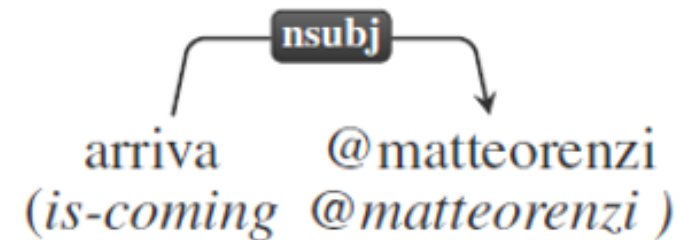
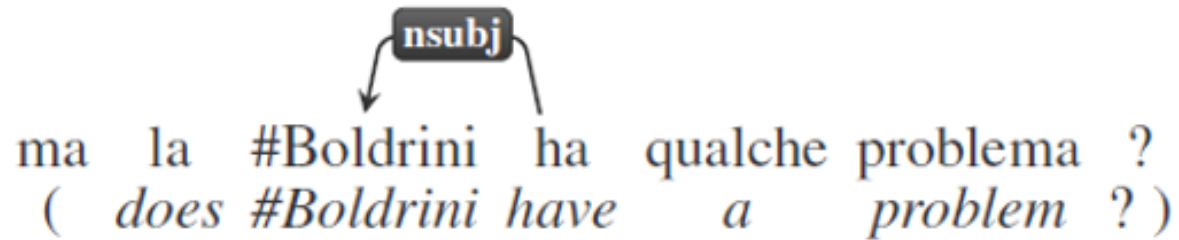
Challenging aspects of Twitter language and its analysis:

- continuous shift from written to spoken language, and *vice versa*
- hashtags, mentions/replies, emoticons/emojis, and other conventions of computer-mediated communication
- unconventional, even unintelligible, elements (e.g. unknown/mispelled words)

# Annotation Guidelines

How we dealt with them in annotation:

- (if syntactically integrated) assigning them their corresponding syntactic role



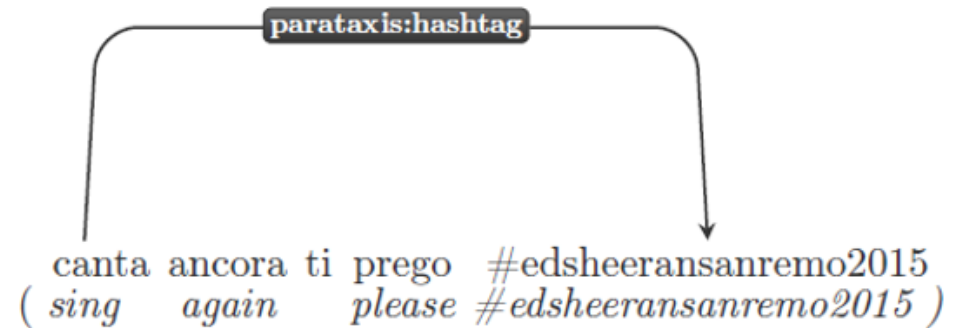
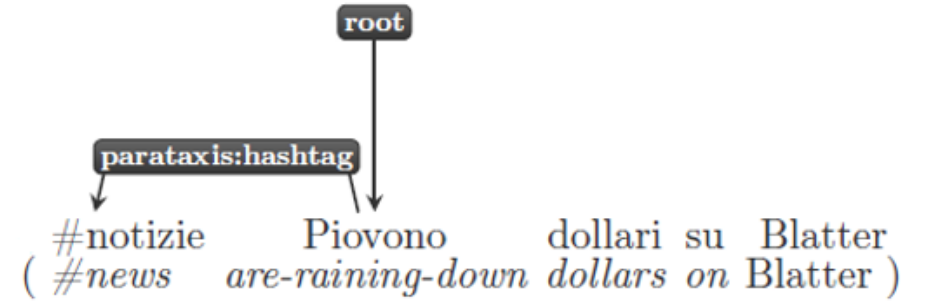
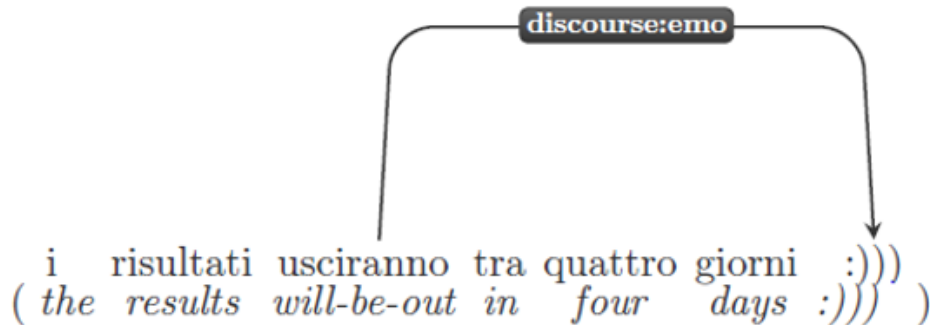
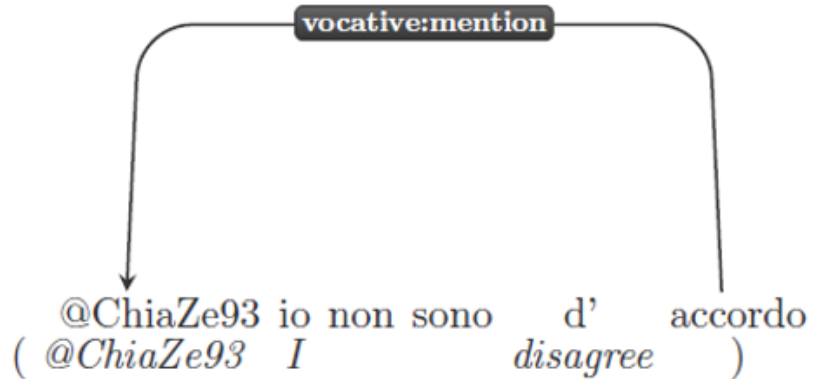


# Annotation Guidelines

How we dealt with them in annotation:

- extending the already existing relations with specific subtypes

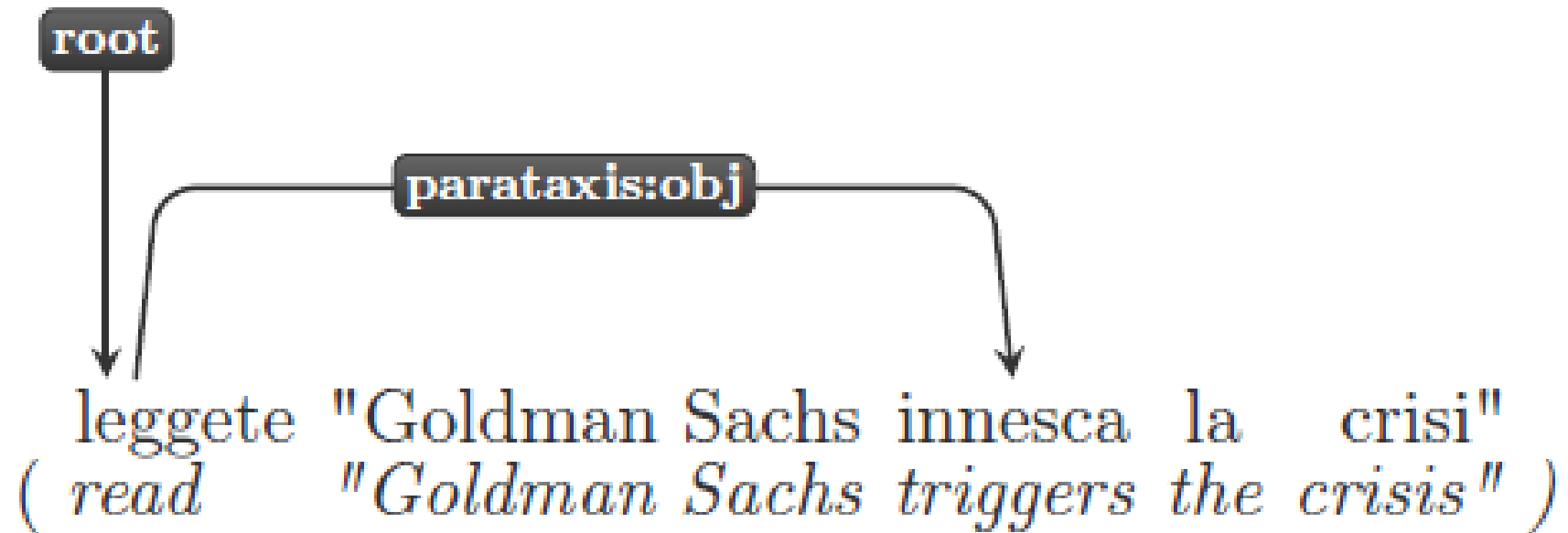
either new



# Annotation Guidelines

How we dealt with them in annotation:

- extending the already existing relations with specific subtypes  
... or mutated from other treebanks



# Future Work

## Short-term goals:

- Complete annotation and guidelines (first release expected: November 2017 – UD v.2.1)
- Extend parsing experiments using the resource as *training* set
- Use the resource for Sentiment Analysis applications

## Long-term goals:

- Enrich the resource with texts from other social media
- Open this work to a multilingual comparison

Thank you!