

Presenting TWITTIRÒ-UD

An Italian Twitter Treebank in Universal Dependencies

Alessandra Teresa Cignarella^{a,b} Cristina Bosco^b and Paolo Rosso^a



- a. Universitat Politècnica de València
- b. Università degli Studi di Torino



Motivation

1. Sentiment Analysis and Opinion Mining

1. Sentiment Analysis and Opinion Mining

→ irony, sarcasm, stance, hate speech, misogyny...

1. Sentiment Analysis and Opinion Mining

→ irony, sarcasm, stance, hate speech, misogyny...

2. Dealing with social media texts

1. Sentiment Analysis and Opinion Mining

→ irony, sarcasm, stance, hate speech, misogyny...

2. Dealing with social media texts

→ **hard!!**

1. Sentiment Analysis and Opinion Mining

→ irony, sarcasm, stance, hate speech, misogyny...

2. Dealing with social media texts

→ **hard!!**

3. Syntax

1. Sentiment Analysis and Opinion Mining

→ irony, sarcasm, stance, hate speech, misogyny...

2. Dealing with social media texts

→ **hard!!**

3. Syntax

→ **Universal Dependencies are cool!**

Research Questions

Research Questions

1. How can we automatically detect irony ?

Research Questions

1. How can we automatically detect irony ?
2. Could syntax information help in the detection of irony?

Research Questions

1. How can we automatically detect irony ?
2. Could syntax information help in the detection of irony?
...and maybe help in other detection tasks too?

Research Questions

1. How can we automatically detect irony ?
2. Could syntax information help in the detection of irony?
...and maybe help in other detection tasks too?

Our approach:

Research Questions

1. How can we automatically detect irony ?
2. Could syntax information help in the detection of irony?
...and maybe help in other detection tasks too?

Our approach:

Let's build a corpus and find out!

What is TWITTIRÒ-UD ?

What is TWITTIRÒ-UD ?

Treebank

What is TWITTIRÒ-UD ?

Treebank

Italian



What is TWITTIRÒ-UD ?

Treebank

Italian



Twitter



What is TWITTIRÒ-UD ?

Treebank

Italian



Twitter



Universal Dependencies



What is TWITTIRÒ-UD ?

Treebank

Italian



Twitter



Universal Dependencies



Sarcasm

Irony



Related Work



















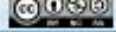



Social media & Twitter:

Social media & Twitter:


















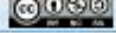



- **Tagging the Twittersverse** (Foster et al., 2011)
- **The French Social Media Bank** (Seddah et al., 2012)
- **TWEEBANK** (Kong et al., 2014)
- **TWEEBANK v2** (Liu et al., 2018)
- **Arabic** (Albogamy and Ramsay, 2017)
- **African-American English** (Blodgett et al., 2018)
- **Hindi English** (Bhat et al., 2018)

Related Work

Related Work





















▼	 Italian	5	781K	   W	IE, Romance
Italian treebanks					
▶	ISDT	298K		  W	 ★★★★★
▶	VIT	279K		 	 ★★★★★
▶	ParTUT	55K		  W	 ★★★★★
▶	PoSTWITA	124K			 ★★★★★
▶	PUD	23K		 W	 ★★★★★

Related Work

▼	 Italian	5	781K	   W	IE, Romance
Italian treebanks					
▶	ISDT	298K		  W	 ★★★★★
▶	VIT	279K			 ★★★★★
▶	ParTUT	55K		  W	 ★★★★★
▶	PoSTWITA	124K			 ★★★★★
▶	PUD	23K		 W	 ★★★★★

Two main references for our work:

Related Work

Italian		5	781K	   W	IE, Romance
Italian treebanks					
▶	ISDT	298K		  W	 ★★★★★
▶	VIT	279K			 ★★★★★
▶	ParTUT	55K		  W	 ★★★★★
▶	PoSTWITA	124K			 ★★★★★
▶	PUD	23K		 W	 ★★★★★

Two main references for our work:

- UD_Italian treebank (Simi et al., 2014)

Related Work

Italian		5	781K		IE, Romance	
Italian treebanks						
▶	ISDT	298K	(LFD)			★★★★★
▶	VIT	279K	(LF)			★★★★★
▶	ParTUT	55K	(LF)			★★★★★
▶	PoSTWITA	124K	(LF)			★★★★★
▶	PUD	23K	(LF)			★★★★★

Two main references for our work:

- UD_Italian treebank (Simi et al., 2014)
- PoSTWITA-UD (Sanguinetti et al., 2018)

Data

- **1,424** tweets from **TWITTIRÒ** (Cignarella et al., 2018)

- **1,424** tweets from **TWITTIRÒ** (Cignarella et al., 2018)
- fine-grained irony annotation (Karoui et al. 2017)

- **1,424** tweets from **TWITTIRÒ** (Cignarella et al., 2018)
- fine-grained irony annotation (Karoui et al. 2017)
 1. EXPLICIT
 2. IMPLICIT

- **1,424** tweets from **TWITTIRÒ** (Cignarella et al., 2018)
- fine-grained irony annotation (Karoui et al. 2017)
 1. EXPLICIT
 1. ANALOGY
 2. EUPHEMISM
 3. RHETORICAL QUESTION
 4. OXYMORON or PARADOX
 5. FALSE ASSERTION
 6. CONTEXT SHIFT
 7. HYPERBOLE or EXAGGERATION
 8. OTHER
 2. IMPLICIT

- **1,424 tweets from TWITTIRÒ** (Cignarella et al., 2018)
- **fine-grained irony annotation** (Karoui et al. 2017)
 1. EXPLICIT
 2. IMPLICIT
 1. ANALOGY
 2. EUPHEMISM
 3. RHETORICAL QUESTION
 4. OXYMORON or PARADOX
 5. FALSE ASSERTION
 6. CONTEXT SHIFT
 7. HYPERBOLE or EXAGGERATION
 8. OTHER
- **sarcasm annotation** (EVALITA 2018)

Annotation

Annotation

```
# text = Presentato il nuovo iPhone. È già al 36%  
di batteria.
```

Annotation

```
# text = Presentato il nuovo iPhone. È già al 36%  
di batteria.
```

```
# irony = EXPLICIT OXYMORON/PARADOX
```

Annotation

```
# text = Presentato il nuovo iPhone. È già al 36%  
di batteria.
```

```
# irony = EXPLICIT OXYMORON/PARADOX
```

```
# sarcasm = 1
```

Annotation

```
# text = Presentato il nuovo iPhone. È già al 36%  
di batteria.
```

```
# irony = EXPLICIT OXYMORON/PARADOX
```

```
# sarcasm = 1
```

Translation:

The new iPhone has been launched. Battery is
already at 36%.

With the tool *UDPipe*:

With the tool *UDPipe*:

- tokenization
- lemmatization
- PoS-tagging
- dependency parsing

With the tool *UDPipe*:

- tokenization
- lemmatization
- PoS-tagging
- dependency parsing



With the tool *UDPipe*:

- tokenization
- lemmatization
- PoS-tagging
- dependency parsing



1,424 tweets!
(17,933 tokens)

With the tool *UDPipe*:

- tokenization
- lemmatization
- PoS-tagging
- dependency parsing



1,424 tweets!
(17,933 tokens)

Full release in the UD repository: **November 2019**

Data

sent_id = *****

twittiro = EXPLICIT OXYMORON/PARADOX

TEXT = Presentato il nuovo iPhone. È già al 36% di batteria.

1	Presentato	presentare	VERB	V	Gender=Masc Number=Sing Tense=Past VerbForm=Part	11	acl	_	_
2	il	il	DET	RD	Definite=Def Gender=Masc Number=Sing PronType=Art	4	det	_	_
3	nuovo	nuovo	ADJ	A	Gender=Masc Number=Sing	4	amod	_	_
4	iPhone	iPhone	PROPN	SP	_	1	nsubj	_	SpaceAfter=No
5	.	.	PUNCT	FS	_	1	punct	_	_
6	È	essere	AUX	V	Mood=Ind Number=Sing Person=3 Tense=Pres VerbForm=Fin	11	cop	_	_
7	già	già	ADV	B	_	11	advmod	_	_
8-9	al	_	_	_	_	_	_	_	_
8	a	a	ADP	E	_	11	case	_	_
9	il	il	DET	RD	Definite=Def Gender=Masc Number=Sing PronType=Art	11	det	_	_
10	36	36	NUM	N	NumType=Card	11	nummod	_	SpaceAfter=No
11	%	%	SYM	SYM	_	0	root	_	_
12	di	di	ADP	E	_	13	case	_	_
13	batteria	batteria	NOUN	S	Gender=Fem Number=Sing	11	nmod	_	SpaceAfter=No
14	.	.	PUNCT	FS	_	11	punct	_	SpaceAfter=ln

Data

sent_id = *****

twittiro = EXPLICIT OXYMORON/PARADOX

TEXT = Presentato il nuovo iPhone. E già al 36% di batteria.

1	Presentato	presentare	VERB	V	Gender=Masc Number=Sing Tense=Past VerbForm=Part	11	acl	_	_
2	il	il	DET	RD	Definite=Def Gender=Masc Number=Sing PronType=Art	4	det	_	_
3	nuovo	nuovo	ADJ	A	Gender=Masc Number=Sing	4	amod	_	_
4	iPhone	iPhone	PROPN	SP	_	1	nsubj	_	SpaceAfter=No
5	.	.	PUNCT	FS	_	1	punct	_	_
6	È	essere	AUX	V	Mood=Ind Number=Sing Person=3 Tense=Pres VerbForm=Fin	11	cop	_	_
7	già	già	ADV	B	_	11	advmod	_	_
8-9	al	_	_	_	_	_	_	_	_
8	a	a	ADP	E	_	11	case	_	_
9	il	il	DET	RD	Definite=Def Gender=Masc Number=Sing PronType=Art	11	det	_	_
10	36	36	NUM	N	NumType=Card	11	nummod	_	SpaceAfter=No
11	%	%	SYM	SYM	_	0	root	_	_
12	di	di	ADP	E	_	13	case	_	_
13	batteria	batteria	NOUN	S	Gender=Fem Number=Sing	11	nmod	_	SpaceAfter=No
14	.	.	PUNCT	FS	_	11	punct	_	SpaceAfter=In

1. Fine-grained annotation for irony

sent_id = *****

twittiro = EXPLICIT OXYMORON/PARADOX

TEXT = Presentato il nuovo iPhone. E già al 36% di batteria.

1	Presentato	presentare	VERB	V	Gender=Masc Number=Sing Tense=Past VerbForm=Part	11	acl	–	–
2	il	il	DET	RD	Definite=Def Gender=Masc Number=Sing PronType=Art	4	det	–	–
3	nuovo	nuovo	ADJ	A	Gender=Masc Number=Sing	4	amod	–	–
4	iPhone	iPhone	PROPN	SP	–	1	nsubj	–	SpaceAfter=No
5	.	.	PUNCT	FS	–	1	punct	–	–
6	È	essere	AUX	V	Mood=Ind Number=Sing Person=3 Tense=Pres VerbForm=Fin	11	cop	–	–
7	già	già	ADV	B	–	11	advmod	–	–
8-9	al	–	–	–	–	–	–	–	–
8	a	a	ADP	E	–	11	case	–	–
9	il	il	DET	RD	Definite=Def Gender=Masc Number=Sing PronType=Art	11	det	–	–
10	36	36	NUM	N	NumType=Card	11	nummod	–	SpaceAfter=No
11	%	%	SYM	SYM	–	0	root	–	–
12	di	di	ADP	E	–	13	case	–	–
13	batteria	batteria	NOUN	S	Gender=Fem Number=Sing	11	nmod	–	SpaceAfter=No
14	.	.	PUNCT	FS	–	11	punct	–	SpaceAfter=In

1. Fine-grained annotation for irony

sent_id = *****

twittiro = EXPLICIT OXYMORON/PARADOX

TEXT = Presentato il nuovo iPhone. E già al 36% di batteria.

1	Presentato	presentare	VERB	V	Gender=Masc Number=Sing Tense=Past VerbForm=Part	11	acl	–	–
2	il	il	DET	RD	Definite=Def Gender=Masc Number=Sing PronType=Art	4	det	–	–
3	nuovo	nuovo	ADJ	A	Gender=Masc Number=Sing	4	amod	–	–
4	iPhone	iPhone	PROPN	SP	–	1	nsubj	–	SpaceAfter=No
5	.	.	PUNCT	FS	–	1	punct	–	–
6	È	essere	AUX	V	Mood=Ind Number=Sing Person=3 Tense=Pres VerbForm=Fin	11	cop	–	–
7	già	già	ADV	B	–	11	advmod	–	–
8-9	al	–	–	–	–	–	–	–	–
8	a	a	ADP	E	–	11	case	–	–
9	il	il	DET	RD	Definite=Def Gender=Masc Number=Sing PronType=Art	11	det	–	–
10	36	36	NUM	N	NumType=Card	11	nummod	–	SpaceAfter=No
11	%	%	SYM	SYM	–	0	root	–	–
12	di	di	ADP	E	–	13	case	–	–
13	batteria	batteria	NOUN	S	Gender=Fem Number=Sing	11	nmod	–	SpaceAfter=No
14	.	.	PUNCT	FS	–	11	punct	–	SpaceAfter=ln

1. Fine-grained annotation for irony

2. Morpho-syntactic information

sent_id = *****

twittiro = EXPLICIT OXYMORON/PARADOX

TEXT = Presentato il nuovo iPhone. E già al 36% di batteria.

1	Presentato	presentare	VERB	V	Gender=Masc Number=Sing Tense=Past VerbForm=Part	11	acl	_	_
2	il	il	DET	RD	Definite=Def Gender=Masc Number=Sing PronType=Art	4	det	_	_
3	nuovo	nuovo	ADJ	A	Gender=Masc Number=Sing	4	amod	_	_
4	iPhone	iPhone	PROPN	SP	_	1	nsubj	_	SpaceAfter=No
5	.	.	PUNCT	FS	_	1	punct	_	_
6	È	essere	AUX	V	Mood=Ind Number=Sing Person=3 Tense=Pres VerbForm=Fin	11	cop	_	_
7	già	già	ADV	B	_	11	advmod	_	_
8-9	al	_	_	_	_	_	_	_	_
8	a	a	ADP	E	_	11	case	_	_
9	il	il	DET	RD	Definite=Def Gender=Masc Number=Sing PronType=Art	11	det	_	_
10	36	36	NUM	N	NumType=Card	11	nummod	_	SpaceAfter=No
11	%	%	SYM	SYM	_	0	root	_	_
12	di	di	ADP	E	_	13	case	_	_
13	batteria	batteria	NOUN	S	Gender=Fem Number=Sing	11	nmod	_	SpaceAfter=No
14	.	.	PUNCT	FS	_	11	punct	_	SpaceAfter=ln

Issues Encountered and Lessons Learned

Issues Encountered and Lessons Learned

- Tokenization errors depending on misspelled words

Issues Encountered and Lessons Learned

- Tokenization errors depending on misspelled words

xkè → perché

Issues Encountered and Lessons Learned

- Tokenization errors depending on misspelled words

xkè → perché

- Punctuation irregularly used

Issues Encountered and Lessons Learned

- Tokenization errors depending on misspelled words

xkè → perché

- Punctuation irregularly used
- Twitter marks

Issues Encountered and Lessons Learned

- Tokenization errors depending on misspelled words

xkè → perché

- Punctuation irregularly used
- Twitter marks

#HASHTAG

Issues Encountered and Lessons Learned

- Tokenization errors depending on misspelled words

xkè → perché

- Punctuation irregularly used
- Twitter marks

#HASHTAG

@mention

Issues Encountered and Lessons Learned

- Tokenization errors depending on misspelled words

xkè → perché

- Punctuation irregularly used
- Twitter marks
- No sentence splitting

#HASHTAG

@mention

Issues Encountered and Lessons Learned

- Tokenization errors depending on misspelled words

xkè → perché

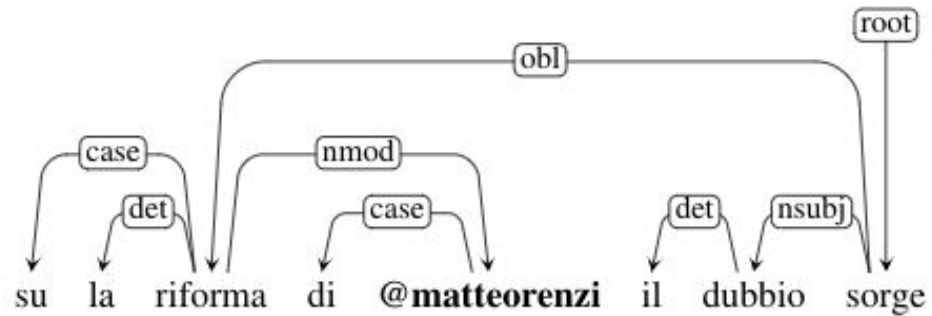
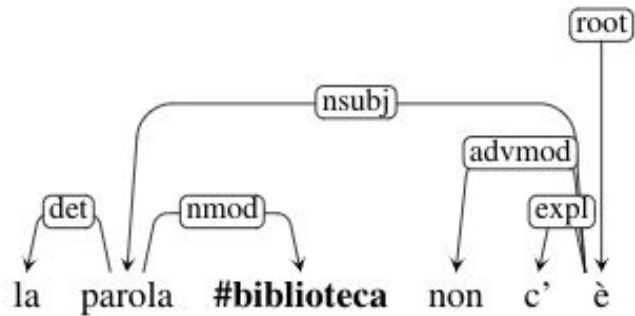
- Punctuation irregularly used
- Twitter marks
- No sentence splitting
- Single-root constraint

#HASHTAG

@mention

Issues Encountered and Lessons Learned

Issues Encountered and Lessons Learned



Issues Encountered and Lessons Learned

		PoSTWITA-UD	TWITTERÒ-UD
hashtags	parataxis:hashtag	40.89%	54.79%
	nmod	19.64%	11.55%
	nsubj	13.48%	8.59%
	<i>other</i>	25.99%	25.07%
mentions	vocative:mention	92.37%	87.41%
	<i>other</i>	7.63%	12.59%

Issues Encountered and Lessons Learned

		PoSTWITA-UD	TWITTIRÒ-UD
hashtags	parataxis:hashtag	40.89%	54.79%
	nmod	19.64%	11.55%
	nsubj	13.48%	8.59%
	<i>other</i>	25.99%	25.07%
mentions	vocative:mention	92.37%	87.41%
	<i>other</i>	7.63%	12.59%

Issues Encountered and Lessons Learned

		PoSTWITA-UD	TWITTIRÒ-UD
hashtags	parataxis:hashtag	40.89%	54.79%
	nmod	19.64%	11.55%
	nsubj	13.48%	8.59%
	<i>other</i>	25.99%	25.07%
mentions	vocative:mention	92.37%	87.41%
	<i>other</i>	7.63%	12.59%

Issues Encountered and Lessons Learned

		PoSTWITA-UD	TWITTIRÒ-UD
hashtags	parataxis:hashtag	40.89%	54.79%
	nmod	19.64%	11.55%
	nsubj	13.48%	8.59%
	<i>other</i>	25.99%	25.07%
mentions	vocative:mention	92.37%	87.41%
	<i>other</i>	7.63%	12.59%

Other Highlights

Other Highlights

- **Punctuation** is indeed exploited more extensively in the two social media datasets rather than in UD_Italian.

Other Highlights

- **Punctuation** is indeed exploited more extensively in the two social media datasets rather than in UD_Italian.
- **Mentions** and **hashtags** have a similar distribution in the two social media datasets.

Other Highlights

- **Punctuation** is indeed exploited more extensively in the two social media datasets rather than in UD_Italian.
- **Mentions** and **hashtags** have a similar distribution in the two social media datasets.
- The use of **passive voices** (`aux:pass`) is low in PoSTWITA-UD and in TWITTIRÒ-UD, indicating a preference for the exploitation of **active voices**, as it happens in spoken language.

A Parsing Experiment

A Parsing Experiment

We performed an evaluation of UDPipe using the TWITTIRÒ-UD gold corpus as a test set.

A Parsing Experiment

We performed an evaluation of UDPipe using the TWITTIRÒ-UD gold corpus as a test set.

The following settings were exploited:

A Parsing Experiment

We performed an evaluation of UDPipe using the TWITTIRÒ-UD gold corpus as a test set.

The following settings were exploited:

1. training UDPipe using only UD_Italian

A Parsing Experiment

We performed an evaluation of UDPipe using the TWITTIRÒ-UD gold corpus as a test set.

The following settings were exploited:

1. training UDPipe using only UD_Italian
2. training UDPipe using only PoSTWITA-UD

A Parsing Experiment

We performed an evaluation of UDPipe using the TWITTIRÒ-UD gold corpus as a test set.

The following settings were exploited:

1. training UDPipe using only UD_Italian
2. training UDPipe using only PoSTWITA-UD
3. training UDPipe using both resources

A Parsing Experiment

A Parsing Experiment

	UD_It			PoSTW			UD_It+PoSTW		
	P	R	F1	P	R	F1	P	R	F1
Tokens	66.85	67.28	67.07	66.50	65.15	65.82	67.62	67.63	67.62
Sentences	66.18	66.18	66.18	66.18	66.18	66.18	66.18	66.18	66.18
Words	66.73	67.12	66.92	66.36	65.01	65.67	67.54	67.56	67.55
UPOS	57.10	57.44	57.27	62.71	61.44	62.07	65.75	65.77	65.76
XPOS	56.30	56.63	56.47	62.23	60.97	61.59	65.59	65.61	65.60
Feats	59.35	59.70	59.52	62.17	60.91	61.53	65.64	65.66	65.65
AllTags	55.11	55.43	55.27	60.59	59.36	59.97	65.04	65.06	65.05
Lemmas	60.88	61.23	61.05	62.17	60.91	61.53	65.48	65.50	65.49
UAS	66.73	67.12	66.92	66.36	65.01	65.67	67.54	67.56	67.55
LAS	50.12	50.42	50.27	54.07	52.97	53.51	56.84	56.85	56.85

A Parsing Experiment

	UD_It			PoSTW			UD_It+PoSTW		
	P	R	F1	P	R	F1	P	R	F1
Tokens	66.85	67.28	67.07	66.50	65.15	65.82	67.62	67.63	67.62
Sentences	66.18	66.18	66.18	66.18	66.18	66.18	66.18	66.18	66.18
Words	66.73	67.12	66.92	66.36	65.01	65.67	67.54	67.56	67.55
UPOS	57.10	57.44	57.27	62.71	61.44	62.07	65.75	65.77	65.76
XPOS	56.30	56.63	56.47	62.23	60.97	61.59	65.59	65.61	65.60
Feats	59.35	59.70	59.52	62.17	60.91	61.53	65.64	65.66	65.65
AllTags	55.11	55.43	55.27	60.59	59.36	59.97	65.04	65.06	65.05
Lemmas	60.88	61.23	61.05	62.17	60.91	61.53	65.48	65.50	65.49
UAS	66.73	67.12	66.92	66.36	65.01	65.67	67.54	67.56	67.55
LAS	50.12	50.42	50.27	54.07	52.97	53.51	56.84	56.85	56.85

A Parsing Experiment

	UD_It			PoSTW			UD_It+PoSTW		
	P	R	F1	P	R	F1	P	R	F1
Tokens	66.85	67.28	67.07	66.50	65.15	65.82	67.62	67.63	67.62
Sentences	66.18	66.18	66.18	66.18	66.18	66.18	66.18	66.18	66.18
Words	66.73	67.12	66.92	66.36	65.01	65.67	67.54	67.56	67.55
UPOS	57.10	57.44	57.27	62.71	61.44	62.07	65.75	65.77	65.76
XPOS	56.30	56.63	56.47	62.23	60.97	61.59	65.59	65.61	65.60
Feats	59.35	59.70	59.52	62.17	60.91	61.53	65.64	65.66	65.65
AllTags	55.11	55.43	55.27	60.59	59.36	59.97	65.04	65.06	65.05
Lemmas	60.88	61.23	61.05	62.17	60.91	61.53	65.48	65.50	65.49
UAS	66.73	67.12	66.92	66.36	65.01	65.67	67.54	67.56	67.55
LAS	50.12	50.42	50.27	54.07	52.97	53.51	56.84	56.85	56.85

Results in-line with state of the art
(PoSTWITA-UD, Sanguinetti et al., 2018)

Conclusions

Conclusions

- We discuss the annotation of this resource which encompasses a fine-grained representation of **irony** and the **UD** morpho-syntactic analysis

Conclusions

- We discuss the annotation of this resource which encompasses a fine-grained representation of **irony** and the **UD** morpho-syntactic analysis
- **Release of the complete resource** (1,424 tweets) to be accomplished in November 2019

Conclusions

- We discuss the annotation of this resource which encompasses a fine-grained representation of **irony** and the **UD** morpho-syntactic analysis
- **Release of the complete resource** (1,424 tweets) to be accomplished in November 2019
- It enriches the scenario of available resources for a text genre which is especially hard to parse (**social media** texts)

Future Work

Future Work

- Investigation of possible **relationships between syntax and semantics** of the uses of figurative language (irony in particular)

- Investigation of possible **relationships between syntax and semantics** of the uses of figurative language (irony in particular)
→ ongoing experiments...

- Investigation of possible **relationships between syntax and semantics** of the uses of figurative language (irony in particular)
→ ongoing experiments...
- A resource whose annotation encompasses both UD relations and a fine-grained description of irony may indeed pave the way for the investigation of whether **syntactic knowledge might help in SA** and other related tasks

- Investigation of possible **relationships between syntax and semantics** of the uses of figurative language (irony in particular)
→ ongoing experiments...
- A resource whose annotation encompasses both UD relations and a fine-grained description of irony may indeed pave the way for the investigation of whether **syntactic knowledge might help in SA** and other related tasks
→ new NLP features for Sentiment Analysis?

Thank you!

`cigna@di.unito.it`