Annotating Italian Social Media Texts in Universal Dependencies

Manuela Sanguinetti
Cristina Bosco
Alessandro Mazzei
Università di Torino
Dipartimento di Informatica
Torino, Italy
{msanguin,bosco,mazzei}@di.unito.it

Alberto Lavelli
Fondazione Bruno Kessler
Trento, Italy
lavelli@fbk.eu

Fabio Tamburini
Università di Bologna
FICLIT
Bologna, Italy
fabio.tamburini@unibo.it

Abstract

Social media texts have been widely used in recent years for various tasks related to sentiment analysis and opinion mining; nevertheless, they still feature a wide range of linguistic phenomena that have proved to be particularly challenging for automatic processing, especially for syntactic parsing. In this paper, we describe a recently started project for the development of PoSTWITA-UD, a novel Italian Twitter treebank in Universal Dependencies. In particular, the paper focuses on its development steps, and on the challenges such work entails, both for automatic systems and human annotators, by discussing the errors produced, by parsers in particular, and the guidelines we adopted for manual revision of annotated tweets. Such guidelines aim to bring to the reader's attention the most critical cases (in themselves, but also in a UD perspective) encountered so far and stemming from the specific characteristics of the texts we are dealing with.

1 Introduction

In the last few years, the interest for automatic evaluation of social media texts has grown considerably; thanks to the various APIs available from the platform, Twitter in particular has been considered a valuable source of data that can be used for different computational linguistics studies and applications. Nevertheless, the annotation and exploitation of Twitter corpora are currently mainly referred to sentiment analysis and opinion mining or other semantic-oriented forms of processing, see e.g. tasks in SemEval 2017[1] and EVALITA (Barbieri et al., 2016). Only a few experiments have been done for developing treebanks and datasets from social media annotated with Part-of-Speech tags and other morphological features (see Section 2).

Regardless of the irregularities of Twitter language, human beings do not seem to find it excessively troubling to understand each other when communicating via social media. Therefore, among the research question that we would like to address, there is also how much this performance depends on human morpho-syntactic ability or on other parts of linguistic competence.

Considering that the availability of a full or partial syntactic analysis can improve the results of semantic and pragmatic-oriented techniques, we propose the development of PoSTWITA-UD, a collection of social media texts annotated according to a well-known dependency-based annotation format: the Universal Dependencies (Nivre et al., 2016)[2].

The goal of this work is twofold. On one hand, it consists in making available a resource currently missing, for Italian in particular, which can be exploited for training NLP systems in order to enhance their performance on social media texts. On the other hand, it may also contribute to the wider debate about social media texts and their analysis, for example by showing how much syntactic information can be helpful for a given NLP task or downstream application; we refer in particular to phenomena such as negation and coordination scope, which, if not correctly detected, can strongly undermine the results obtained e.g. by a sentiment analysis engine in classifying the polarity of a message (Bosco et al., 2013b).

From a methodological point of view, our choice to adopt the UD scheme stems from the interest in a dependency-based representation for-

[1] http://alt.qcri.org/semeval2017/task4/

[2] http://universaldependencies.org/

Proceedings of the Fourth International Conference on Dependency Linguistics (Depling 2017), pages 229-239,
Pisa, Italy, September 18-20 2017

mat that has gained full acceptance from the research community over a few years, especially regarding Italian resources. The goal of creating this resource goes hand in hand with that of sharing it and validating its annotation according to a shared standard, such as the one UD projects aims to provide. In addition, UD format allows to extend the inventory of morphological features and syntactic relations with further subtypes, according to the language, genre or linguistic construction peculiarities. For all these reasons Universal Dependencies proved to be the optimal representation choice.

This project benefits from the availability of a Twitter corpus used as dataset for the task of Part-of-Speech tagging on social media texts (PoSTWITA) held at the 2016 edition of EVALITA, the evaluation campaign for Italian NLP tools[3]. For our current purpose, we further enriched the corpus by adding the missing annotation layers, i.e. lemmas, morphological features and syntactic relations, all in compliance with the annotation scheme and principles of Universal Dependencies.

The content of this paper is thus organized as follows. Next section briefly surveys the literature on syntactic analysis of social media texts, and Section 3 introduces the dataset used for our project. Sections 4, 5 and 6 describe the various annotation steps, while in Section 7 we discuss the creation of the gold standard set. In particular, in Section 7.2 we discuss the annotation guidelines we followed for manual revision. Finally, Section 8 closes the paper with some considerations on the current state of the project.

2 Related Work

Considering their increasing importance in NLP, several efforts have been made to annotate, manually or semi-automatically, social media texts. However, the use of typical NLP tools and techniques has proved critical, essentially by virtue of the unconventional use of the language norms at all levels (orthography, lexicon, morphology and syntax) and the amount of noise such non-standard linguistic behaviors and meta-textual elements can bring about. Although various attempts to produce such kind of specialized resources and tools are described in literature (e.g. (Gimpel et al., 2011; Owoputi et al., 2013; Lynn et al., 2015; Rei et al., 2016)), most of these attempts mainly focus on PoS-tagged corpora, while few of them deal with syntactic annotation as well. One of such works is that of Foster et al. (2011), who built a dataset containing 1,000 sentences including tweets and forum posts, with the specific aim of investigating the problems of parsing social media texts. Later on, other works attempted to overcome such limits by creating *ah hoc* resources to be used as training data for parsing. This is the case of the French Social Media Bank (Seddah et al., 2012), a set of 1,700 sentences from various types of user-generated content (among those, tweets), annotated using an adapted version of the French Treebank (Abeillé et al., 2003) scheme, and TWEEBANK (Kong et al., 2014), built by manually adding dependency parses to tweets drawn from the PoS-tagged Twitter corpus of Owoputi et al. (2013).

Finally, it is worth mentioning the English Web Treebank (Silveira et al., 2014), a collection of more than 16k sentences taken from various Web media, including blogs, emails, reviews and Yahoo! answers, and also available in UD format.

To the best of our knowledge, however, the one presented here is the first work devoted to create a Twitter treebank annotated according to UD specifications, and is almost certainly the first resource of this kind created for Italian.

3 The Dataset

PoSTWITA-UD was not built from scratch, but it has been developed by processing and further enriching an already existing resource, that is the dataset used for the EVALITA 2016 task on Part-of-Speech tagging of social media, i.e. PoSTWITA (Bosco et al., 2016). Therefore, data and content are the same as those of the PoSTWITA corpus released to the task participants, which includes a development set composed of 6,438 tweets (114,967 tokens), and a test set of 300 tweets (4,759 tokens).

Its content, in turn, comes from the SENTIPOLC corpus, i.e. the dataset used for the EVALITA SENTIment POLarity Classification (SENTIPOLC) task in 2014 (Basile et al., 2014) and 2016 (Barbieri et al., 2016). Furthermore, within the EVALITA 2016 campaign, the same core dataset was made available with semantic-oriented annotations for two other tasks as well: the Named Entity Recognition and Linking in Italian Tweets (NEEL-IT) (Basile et al., 2016)

[3] http://www.evalita.it

and the Event Factuality Annotation (FactA) task (Minard et al., 2016). Working on this treebank thus collocates our current activity in the perspective of the development of a benchmark where a full pipeline of NLP tools can be applied and tested in the future evaluation campaigns.

Considering its use for EVALITA, the PoSTWITA dataset has already been automatically preprocessed, tokenized and PoS tagged, as well as entirely revised by human annotators, in order to remove duplicate tweets and provide a gold annotation. Such gold set is the starting point of the PoSTWITA-UD project, whose development steps are described in the next sections.

4 Tokenization and Part-of-Speech Tags: from PoSTWITA to PoSTWITA-UD

For what concerns tokenization and tagging principles, the PoSTWITA task organizers followed the strategy proposed in the Italian section of the UD guidelines, though applying some minor changes. Assuming, as usual and more suitably in PoS tagging, a neutral perspective with respect to the solution of parsing problems (more relevant in building treebanks), PoSTWITA format differs from the one applied in UD, in that it leaves tokens unsplitted in the two following cases:

- articulated prepositions (e.g. *dalla* ('from-the [fem]'), *nell'* ('in-the'), *al* ('to-the'), ...)
- clitic clusters, which are composed by one or more clitic pronouns attached to the end of a verb form (e.g. *regalaglielo* ('offer-it-to-him'), *dandolo* ('giving-it'), ...)

For this reason, and according to the strategy assumed in previous EVALITA PoS tagging evaluations, two novel specific tags were assigned in these cases: `ADP_A` and `VERB_CLIT`, for articulated prepositions and verbs with clitics respectively.

Furthermore, all the Internet and Twitter-specific tokens that, according to UD specifications, should be classified as `SYM` (symbol) were further specified based on the token type. As a result, all the categories that typically occur in social media texts, like emoticons, Internet addresses, email addresses, hashtags and Twitter mentions had their own tag, i.e. `EMO`, `URL`, `EMAIL`, `HASHTAG` and `MENTION`.

For the development of PoSTWITA-UD, we had to restore the initial UD tokenization format, thus re-splitting all `ADP_A` and `VERB_CLIT` cases into the corresponding UD PoS tags (upos) `ADP+DET` and `VERB+PRON` respectively. We also had to restore all the Twitter-specific tags into `SYM`.

Finally, it should be pointed out that no modification on the sentence splitting has been carried out. Just like the original PoSTWITA dataset, the reference unit is always the tweet in its entirety – which may thus consist of multiple sentences – not the sentence alone.

5 Lemmas and Morphological Features

In order to produce a correctly formatted corpus in CoNLL-U format, we also inserted information about lemmas and morphological features associated to each word. To speed up the process, we relied on AnIta (Tamburini and Melandri, 2012), an Italian morphological analyzer based on a large lexicon (about 110,000 lemmas) able to analyze the various word forms and produce all the possible lemmas and morphological features for these forms. A two-step semi-automatic conversion between the different annotation schemes ensured a full compatibility with the UD specifications.
In the first step we added lemmas and language-specific PoS tags (xpos). As mentioned above, the insertion was done partly with a script that converts AnIta output into a UD-compatible form, and matches the word forms on the PoSTWITA-UD side with the lemmas provided by AnIta for the respective upos. While the xpos tags (the same used in UD_Italian) were added with *ad hoc* heuristics and manual disambiguation.
The insertion of lemmas was also performed manually, by revising the automatic results of the script and adding the missing lemmas. The choice we made in this manual stage represented a guiding principle for syntactic annotation as well (see Section 7.2), i.e what is understandable by a human should be annotated accordingly. With regard to lemmatization in particular, this means that whenever possible, we assigned to a non-standard form the lemma of the respective standard form (though leaving the word form unchanged). Following this principle, we thus assigned the corresponding lemma to the various cases of abbreviation, capitalization, typos and grammatical errors, and word

lengthening.
An exception is made for punctuation, non-intelligible word forms, dialectal forms and foreign words, in which cases the lemma remained the same as the word form.

In the second step we then added the morphological features by following the same strategy described above for lemmatization, that is by matching the proper morphological features with a given word form based on its lemma, upos and xpos tag. The feature insertion step involved the following parts of speech: adjectives, adverbs, determiners, nouns, numerals, pronouns and verbs.

In order to preserve a higher consistency among resources, we also used the language-specific features introduced in UD_Italian for clitic pronouns (`Clitic=Yes`) and possessives (`Poss=Yes`).

6 Syntactic Analysis

The last step included the syntactic annotation of the tweets according to UD specifications. We carried out this task by running different parsers and developing proper annotation guidelines. In this and the next section we describe both aspects.

6.1 Data Parsing

Similar to the previous steps, we first automatically analyzed the texts with state-of-the-art dependency parsers, and then we manually revised the annotation.

As regards Italian UD-compliant resources, the only dataset that was suitable for training is UD_Italian (Bosco et al., 2013a)[4], version 2, which includes texts from newspapers, Wikipedia and legal Italian and European Community sources. Therefore, we performed an out-of-domain parsing experiment, by training different systems on this treebank, though being aware that the result would be undermined by the deep differences between the text types included in such resources.

For the automatic annotation we used some of the parsers that obtained the best performance in a recent comparative study concerning an Italian dependency treebank (Lavelli, 2016), in particular:

- the MATE tools, that include both a graph-based (Bohnet, 2010) and a transition-based parser (Bohnet and Nivre, 2012; Bohnet and Kuhn, 2012); they were run using standard parameters;
- RBG (Lei et al., 2014; Zhang et al., 2014b; Zhang et al., 2014a), which is based on a low-rank factorization method that enables to map high dimensional feature vectors into low dimensional representations; the full model was chosen.

[4]The other resource is the Italian section of the parallel treebank ParTUT (UD_Italian-ParTUT), but it has many overlapping sentences with UD_Italian, and it is much smaller.

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

Table 1: Results of the parsers after the different annotation stages, i.e. with lemmas and language-specific PoS tags (-LX), and with morphological features as well (-F). The parser outputs were evaluated against the gold standard of the test set (300 tweets, -LX and -F columns) but also against the UD_Italian test set (489 sentences, -UD column).

For the near future, we also plan to extend the experiment to other state-of-the-art parsers as well (namely TurboParser (Martins et al., 2013) and ZPar (Zhang and Nivre, 2011)), and to combine all the outputs produced to obtain an improved parsing quality (Hall et al., 2010).

In order to get an overall picture of the parsing results after each of the steps described in Section 5, we parsed both development and test set *a*) after the insertion of lemmas and language-specific PoS tags, and *b*) after the morphological features were also added.

To get a measure of how much parsing quality differs between standard and Twitter texts, in Table 1 we report also the results of the parser on the UD_Italian test set (489 sentences).
For the evaluation step we used the script made available for the CoNLL 2017 Shared Task[5] with the default setting (i.e. by reporting the Labelled Attachment Score, *LAS* F_1 score only).
The overall parsing results are discussed in the next section.

6.2 Results and Discussion

The reported results actually show what we were already expecting: the performance of the three

[5]`http://universaldependencies.org/conll17/evaluation.html`

parsers improves when we add linguistic information. Overall, however, the parsing quality for the PoSTWITA test set is relatively poor considering both the results on UD_Italian test set and the fact that the systems start from partially annotated and corrected texts, rather than raw ones[6]. The explanation we can give is also the most obvious, that is, parsers have to deal with texts from a different domain than those of the training set, and what is more, having very specific - and challenging - features. As a proof of this, we observed the behavior of the three parsers on single relations, assuming that their performance would remain stable on well-known cases and decrease on poorly-covered phenomena in the training set.

To verify this assumption, we observed the F-score obtained by parsers on two sub-sets of relations that reflect two different, though in a sense complementary, aspects: the first one includes the 10 most frequent relations in UD_Italian[7], and the second one comprises three of the relations where parsers get the lowest results, i.e. `discourse`, `parataxis` and `vocative`. These relations are summarized in Table 2, along with their F-score averaged over the three parsers and their distribution both in UD_Italian training set and on PoSTWITA-UD test set.

As it can be seen, just three relations exceed the 90% threshold (`advmod`, `amod` and `cc`), and just one is between 80% and 90%, i.e. the relation linking the direct object to its predicate (`obj`).

The relation with the lowest F-score, among the most frequent in UD_Italian, is the one representing adverbial clauses (`advcl`). This can be explained by the fact that most of the relations labeled by the parsers as adverbial clauses were rather considered as paratactic constructions in the gold set.

Interestingly enough, a quite low F-score is reported for the `nsubj` relation, and the cases where it was erroneously annotated are quite systematic on all three parsers. They correspond to cases of nouns that in the gold set we have chosen to consider as the root of the whole tweet, because they are followed by paratactic elements (see Section 7.2), or as addressees of a given utterance (hence as `vocative`). This aspect, in turn, raises the issue of the use, within the gold set, of labels such as `parataxis`, `vocative` and `discourse`. For the reasons outlined in Section 7.2, these three relations are much more frequent in the PoSTWITA-UD gold set than in UD_Italian, as also reported in Table 2. The far lower frequency of these relations in the training set and, as a result, in parsers outputs, compared to the gold set, leads to the extremely poor parsing quality with respect to these three phenomena.

UD relation	F **score**	% train	% test
acl	58.00	1.18	0.46
advcl	50.98	1.26	1.05
advmod	96.85	3.52	6.22
amod	90.27	5.45	2.25
cc	97.27	2.74	2.43
conj	66.74	3.39	3.26
obj	82.75	3.41	4.72
obl	72.46	5.74	4.23
nmod	72.62	8.06	5.23
nsubj	65.37	4.26	3.62
discourse	0	0.02	3.18
parataxis	11.18	0.14	5.29
vocative	0	0.07	3.83

Table 2: Averaged results of the three parsers, in terms of F-score, along with the relative frequency in UD_Italian training set ('train') and PoSTWITA-UD test set ('test'), of individual relations: the 10 most frequent relations in UD_Italian (upper part), and three of the relations with poorer parsing results (lower part).

7 Towards the gold standard

In this section we describe the creation of a fully corrected PoSTWITA-UD, from the manual revision of parsing output to the definition of the guidelines for the annotators. The annotation methodology, as conceived and tested so far for the test set only, will also be applied to the development set in the next project phase.

7.1 Manual revision and Inter-Annotator Agreement

The manual post-processing of annotated texts, while it was useful for parsers evaluation, represented the first step towards the goal of our work: obtaining a reference gold standard for the further manual annotation, for the current evaluation

[6]This is also true for the UD_Italian test set, which was parsed starting from the CoNLL-U files with gold PoS tags, rather than from raw texts.

[7]Excluding `punct`, `det` and `case`, which are poorly indicative of the challenging aspects of this out-of-domain parsing experiment.

of parsers and for their future training on Twitter texts.

The revision was made by two trained annotators who were familiar with the UD format and using DGAnnotator[8] as tree editor. Although their work proceeded independently, some particularly critical phenomena were previously discussed. This allowed to come up with shared guidelines (see Section 7.2). In order to take into account the fact that the outputs of the different parsers can be affected by different errors, the two annotators used as starting dataset the output files from two (of the three used) different parsers, randomly selected.

As a result of the first correction phase, the degree of inter-annotator agreement (on relations alone) was calculated, using Cohen's kappa as the reference index (Carletta, 1996). The agreement at this stage was $k = 0.83$.
Based on this result, and in particular on cases with higher disagreement, a consistency check on the application of the guidelines and a further revision were made (after which the agreement went up to $k = 0.92$); finally, the corrections of both annotators were merged into a single final file.

7.2 Annotation Guidelines

Several phenomena featuring social media texts are poorly treated by existing morphological analyzers and parsing systems. In fact, it can be quite difficult to decide their collocation within a single layer of analysis (syntax, semantics or pragmatics), since they better collocate in the broader area of communication dynamics taking place in social media conversation. In computer-mediated communication, and specifically on Twitter, users often resort to a language type that is closer to speech, rather than written language. Narrowing it down to Italian, this is found at various levels, from orthography, with forms and expressions that imitate the verbal face-to-face conversation, to lexis (colloquialisms and vulgar language) and syntax, with the prevalence of simple sentences or paratactic forms, clefting, dislocations and syntactic structures that do not respect the typical SVO order of constituents (Zaga, 2012). The continuous shift from written to spoken language and *vice versa*, on the other hand, is also found in the absence (at least in our corpus) of those typical mechanisms of spoken language, such as repairs and restarts.
The absence of these phenomena, and, at the same time, the presence of others (mentioned later in this section) that are typical of the medium used, make Twitter language a unique, for which - unlike written and speech treebanks[9] - we were not able to find clear and shared guidelines.

For the purposes of our project, we had to face the challenge of classifying all these Twitter-specific phenomena within a syntactic framework – rather than within pragmatics or semantics – more specifically the one conceived for Universal Dependencies. For that purpose, we drafted some tentative guidelines and followed them while preparing the gold standard.

In the remainder of this section we briefly discuss these principles by showing some practical annotation examples[10].

Emoticons, emojis and similar aspects. As regards these iconic elements, and emojis in particular, a wide debate has opened on whether they should be considered as an emerging language in itself[11] or just a powerful communication tool that does not substitute language, but rather complements it. While going into the substance of this debate is well beyond the scope of this paper, and of our project in general, we equally had to face the issue on what status we should attribute to these so-called pictograms, in an attempt to draw a line between what should or should not be annotated on the syntactic layer. In fact, emoticons and emojis are typically used to express feelings and emotions, reproduce facial expressions or even convey the intonation of spoken language. Although performing on a more pragmatic, than merely syntactic level, they seem to function in a language-like fashion[12]. In this sense, they could then be compared to interjections and other discourse particles. Bearing in mind what UD guidelines suggest for

[8] http://medialab.di.unipi.it/Project/QA/Parser/DgAnnotator/

[9] Regarding, in particular, UD-based speech treebanks, we mention here the resource available for Slovenian (Dobrovoljc and Nivre, 2016), and that for French (upcoming) (Gerdes and Kahane, 2017).

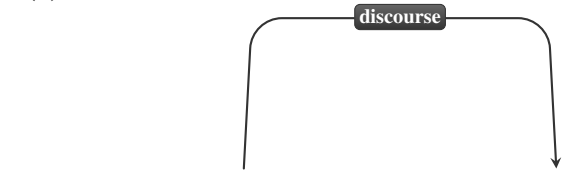
[10] For the sake of readability, we kept just the more relevant dependency edges and the corresponding relations.

[11] See, for example, the study on emojis in Italian language (Chiusaroli, 2015) and the EmojitalianoBot and EmojiWorldBot experiments (Monti et al., 2016)

[12] http://blog.oxforddictionaries.com/2015/11/emoji-language/

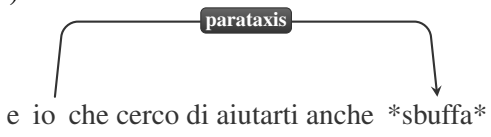
such particles[13], we labelled also emoticons and emojis as `discourse` items, as in example (1).

(1)



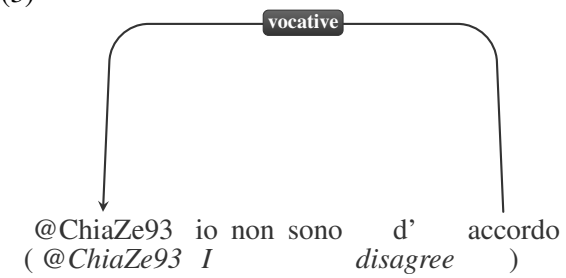
On the other hand, we also found few cases where the tweet ends with an expression (typically a verb) between asterisks that, conversely, substitutes an iconic element (perhaps an emoji). Despite the similar pragmatic function these verbs seem to have with respect to emoticons and emojis, we considered them as independent clauses, therefore as paratactic elements, and annotated as shown in (2).

(2)



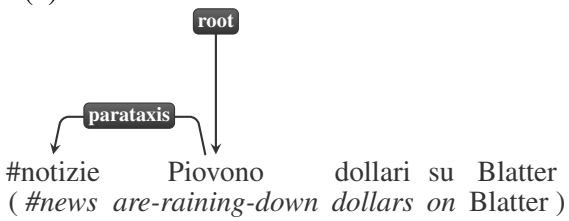
Hashtags, mentions and replies. These are meta-language items with manifold purposes.
The @ symbol that characterizes the so-called mentions and replies is used to call out usernames in tweets. Usernames preceded by the sign become links to the respective Twitter profiles, and can be used mainly in two ways: to just mention another user or to reply another user/s' tweet[14]. The act of addressing to other users by resorting to such conventions can be compared to a typical vocative function, which made us lean on annotating these cases with the `vocative` relation, by attaching the addressee to its host sentence, as in example (3).

(3)

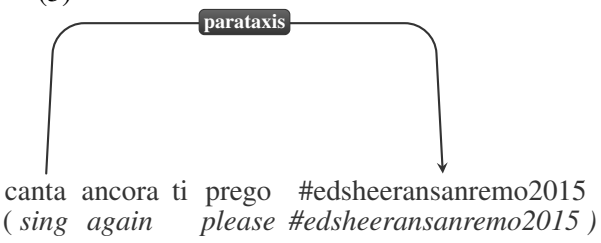


Hashtags are key words or phrases preceded by the # symbol. They serve different purposes, often depending on their position within the tweet. When placed in prefix (example (4)) or suffix position (example (5)), they are mainly used to describe and/or comment the main topic of the tweet, making it more intelligible to other users; in most cases they do not modify any word in particular, nor they reflect any explicit coordination, subordination, or argument relation with a given head word. Similar to other run-on sentences, hashtags are are not integrated into the sentence, rather being joined to the latter without any conjunction or punctuation mark; therefore, we consider them as paratactic elements

(4)



(5)



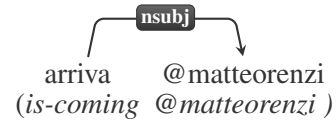
Hashtags and mentions, however, can also be placed in infix position, i.e. by adding their respective sign to the word/phrase or username within the tweet, even just to keep it simple and save character space. In these cases they can be considered as fully syntactically-integrated elements, whose removal could potentially make the sentence ungrammatical (Chiusaroli, 2014); we thus assign them their corresponding syntactic role. In tweets

[13] `http://universaldependencies.org/u/dep/discourse.html`

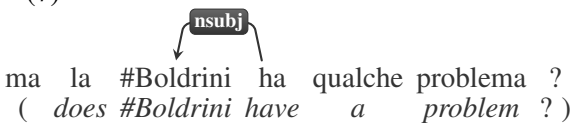
[14] `https://support.twitter.com/articles/464314`

(6) and (7), shown below, the tokens *@matteorenzi* and *#Boldrini* are the actual subjects of the predicates *arriva* and *ha*, respectively.

(6)

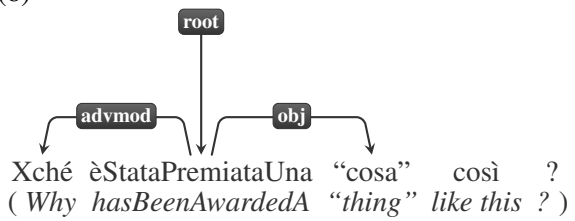


(7)



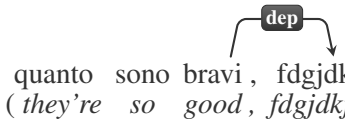
Unknown or mispelled words. Sometimes tweets can also contain a whole host of unconventional elements that substitute actual words: abbreviations, homophones, conflations, or just spelling errors. Whenever we can guess what that element stands for, we assign it the corresponding syntactic role. In the example tweet (8), the adverb *perché* is abbreviated to *Xché* (which is a quite common form in any kind of informal communication), while the two auxiliaries *è stata*, the predicate *premiata* and the determiner *una* were capitalized and conflated into a single token. Considering, however, that among these words, there is one, i.e. the predicate, that can be promoted as the head of the remaining words, we took this item as the sentence root.

(8)



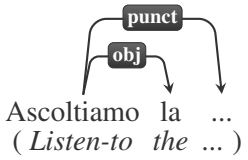
There are cases, however, where we cannot determine which category the word belongs to, nor its syntactic or pragmatic role: in the absence of such information, the word is attached to the nearest head with the `dep` relation, as shown in tweet (9).

(9)



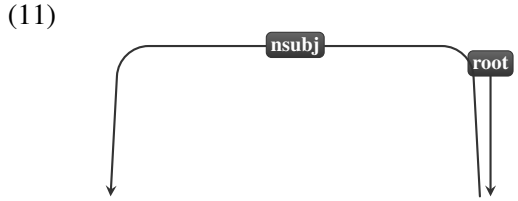
Incomplete tweets. Because of the 140-character limit imposed to Twitter users, it often appears that tweets are incomplete, and the elided part is represented by an ellipsis ("..."). In such cases, the full text can usually be read by clicking on the URL that is appended to the tweet; however, once the tweet is collected and processed, the elided part is lost. Despite this, most of the times, such part is quite predictable by the reader/annotator, either because of the way the remaining sentence is structured or because even one word was partially replaced by the ellipsis, as in example (11). We treat these two cases a bit differently, though. In sentence (10), for example, the fact that the ellipsis points occur after the predicate *Ascoltiamo* and the determiner *la* suggests that there may be a noun depending on that predicate, and, in turn, representing the head of the determiner. We then treat cases like this as typical noun ellipsis, by promoting one of its overt dependents (such as the determiner, in the example) as head word, following the order suggested in UD guidelines[15].

(10)



However, if the suspension ellipsis is used to replace also part of a word that has been cut off, and considering that - in these cases - the dots are part of the word itself[16], the word is treated as it is, without any head promotion of its dependents.

(11)



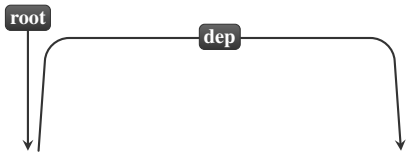
In tweet (11), for example, the word *comme...* is likely to stand for the predicate *commenta* (*(he) comments*); therefore we annotate it as the head.

[15] `http://universaldependencies.org/u/overview/specific-syntax.html/ellipsis`

[16] This is a tokenization principle adopted from the beginning of the corpus development for the PoSTWITA task and that were left unchanged.

URLs. Another common practice in microblogging, and in Twitter posts in particular, is to incorporate links to Web pages, blog entries or even other tweets. These links are usually appended at the end of the tweet and they are not part of its syntactic structure. Therefore, we always consider them as generic dependents of the root, using the `dep` relation (see tweet (12)).

(12)



il programma dettagliato : http://t.co/OJq7hBcH
(*the program detailed : http://t.co/OJq7hBcH*)

On the other hand, a URL may also happen to occur within the sentence, as a syntactically-integrated element. Although we have not encountered similar cases in our treebank yet, we consider the URL as a proper noun and apply the same annotation criteria described above for hashtags and mentions, i.e. we assign the proper syntactic label according to the actual role the URL plays within the sentence.

As mentioned before, these guidelines are preliminary and refer to the trickiest phenomena encountered in the test set. It is not to be ruled out, however, that in the manual revision of the development set there will be other cases that will lead us either to revise the criteria adopted so far or to extend the inventory of uncertain cases. A final version of the guidelines, to be considered as an integration of those conceived for UD_Italian, will be released in the UD repository along with the fully annotated PoSTWITA-UD treebank, by November 2017, that is with the release of UD version 2.1.

8 Conclusion and Future Work

In this paper we presented a recently started project of an Italian Twitter treebank in Universal Dependencies to be released as gold standard for training and testing NLP tools on social media texts. What we achieved so far is the complete annotation of the entire corpus on morphological and syntactic levels, and the manual revision of the test set (300 tweets) by two independent annotators. In parallel with the annotation correction, we also developed some guidelines to properly deal with the genre-specific most critical issues.

As stated above, the project is at an early stage, therefore much work has to be done. First of all, the complete revision of the development set as well (approximately 6,000 tweets), which is planned to be ready for the next release of Universal Dependencies, with a further revision and/or extension of the annotation manual, if necessary. Then we aim to train statistical parsers using this newly-created gold standard and compare their results with the ones obtained in other similar experiments (see, e.g. Petrov and McDonald (2012)).

We are aware of the debate on the nature of NLP results obtained with Twitter-based datasets and their poor generalization with other social media texts (Darling et al., 2012; Eisenstein, 2013). Therefore, in the future we could also attempt to incorporate texts from different social media sources and provide a more balanced resource.

Finally, we would also like to widen the debate on social media text processing by opening this work to a multilingual comparison, which would be made possible by the UD format, specifically designed for that purpose. This would allow us to asses the applicability of our annotation proposal to other languages, thus further encouraging cross-linguistic studies on social media communication.

References

Anne Abeillé, Lionel Clément, and François Toussenel. 2003. Building a treebank for French. In Anne Abeillé, editor, *Treebanks: Building and Using Parsed Corpora*, pages 165–187. Springer Netherlands, Dordrecht.

Francesco Barbieri, Valerio Basile, Danilo Croce, Malvina Nissim, Nicole Novielli, and Viviana Patti. 2016. Overview of the Evalita 2016 SENTIment POLarity Classification Task. In *Proceedings of Evalita 2016*.

Valerio Basile, Andrea Bolioli, Viviana Patti, Paolo Rosso, and Malvina Nissim. 2014. Overview of the Evalita 2014 SENTIment POLarity Classification Task. In *Proceedings of Evalita 2014*.

Pierpaolo Basile, Annalina Caputo, Anna Lisa Gentile, and Giuseppe Rizzo. 2016. Overview of the EVALITA 2016 Named Entity rEcognition and Linking in Italian Tweets (NEELIT) task. In *Proceedings of Evalita 2016*.

Bernd Bohnet and Jonas Kuhn. 2012. The best of both worlds – a graph-based completion model for

transition-based parsers. In *Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics*, pages 77–87, Avignon, France. Association for Computational Linguistics.

Bernd Bohnet and Joakim Nivre. 2012. A transition-based system for joint part-of-speech tagging and labeled non-projective dependency parsing. In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, pages 1455–1465, Jeju Island, Korea. Association for Computational Linguistics.

Bernd Bohnet. 2010. Top accuracy and fast dependency parsing is not a contradiction. In *Proceedings of the 23rd International Conference on Computational Linguistics (Coling 2010)*, pages 89–97, Beijing, China. Coling 2010 Organizing Committee.

Cristina Bosco, Simonetta Montemagni, and Maria Simi. 2013a. Converting Italian treebanks: Towards an Italian Stanford Dependency treebank. In *Proceedings of the 7th Linguistic Annotation Workshop and Interoperability with Discourse*, pages 61–69.

Cristina Bosco, Viviana Patti, and Andrea Bolioli. 2013b. Developing corpora for sentiment analysis: The case of irony and Senti-TUT. *IEEE Intelligent Systems*, 28(2):55–63.

Cristina Bosco, Fabio Tamburini, Andrea Bolioli, and Alessandro Mazzei. 2016. Overview of the EVALITA 2016 Part Of Speech on TWitter for ITAlian task. In *Proceedings of Evalita 2016*.

Jean Carletta. 1996. Assessing agreement on classification tasks: The kappa statistic. *Computational Linguistics*, 22(2):249–254, June.

Francesca Chiusaroli. 2014. Sintassi e semantica dell'hashtag: studio preliminare di una forma di scritture brevi. In *Proceedings of the 1st Italian Conference on Computational Linguistics (CLiC-it 2014)*, pages 117–121, Pisa, Italy.

Francesca Chiusaroli. 2015. La scrittura in emoji tra dizionario e traduzione. In *Proceedings of the 2nd Italian Conference on Computational Linguistics (CLIC-It 2015)*, pages 88–92, Trento, Italy.

William M. Darling, Michael J. Paul, and Fei Song. 2012. Unsupervised part-of-speech tagging in noisy and esoteric domains with a syntactic-semantic bayesian hmm. In *Proceedings of the Workshop on Semantic Analysis in Social Media*, pages 1–9, Stroudsburg, PA, USA. Association for Computational Linguistics.

Kaja Dobrovoljc and Joakim Nivre. 2016. The Universal Dependencies treebank of spoken Slovenian. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation LREC 2016, Portorož, Slovenia, May 23-28*, pages 1566–1571. European Language Resources Association (ELRA).

Jacob Eisenstein. 2013. What to Do About Bad Language on the Internet. *Proceedings of the North American Chapter of the Association for Computational Linguistics : Human Language Technologies (NAACL-HLT)*, pages 359–369.

Jennifer Foster, Özlem Çetinoglu, Joachim Wagner, Joseph Le Roux, Stephen Hogan, Joakim Nivre, Deirdre Hogan, and Josef van Genabith. 2011. #hardtoparse: POS tagging and parsing the twitterverse. In *Analyzing Microtext, Papers from the 2011 AAAI Workshop, San Francisco, California, USA, August 8, 2011*.

Kim Gerdes and Sylvain Kahane. 2017. Trois schémas d'annotation syntaxique en dépendance pour un même corpus de français oral: le cas de la macrosyntaxe. In *Actes de l'atelier "ACor4French - Les corpus annotés du français"*, pages 1–9, Orléans, France.

Kevin Gimpel, Nathan Schneider, Brendan O'Connor, Dipanjan Das, Daniel Mills, Jacob Eisenstein, Michael Heilman, Dani Yogatama, Jeffrey Flanigan, and Noah A. Smith. 2011. Part-of-speech tagging for twitter: Annotation, features, and experiments. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies: Short Papers - Volume 2*, HLT '11, pages 42–47, Stroudsburg, PA, USA. Association for Computational Linguistics.

Johan Hall, Jens Nilsson, and Joakim Nivre. 2010. Single malt or blended? a study in multilingual parser optimization. In Harry Bunt, Paola Merlo, and Joakim Nivre, editors, *Trends in Parsing Technology: Dependency Parsing, Domain Adaptation, and Deep Parsing*, pages 19–33. Springer Netherlands.

Lingpeng Kong, Nathan Schneider, Swabha Swayamdipta, Archna Bhatia, Chris Dyer, and Noah A. Smith. 2014. A dependency parser for tweets. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1001–1012, Doha, Qatar. Association for Computational Linguistics.

Alberto Lavelli. 2016. Comparing state-of-the-art dependency parsers on the Italian Stanford Dependency Treebank. In *Proceedings of the Third Italian Computational Linguistics Conference (CLiC-it 2016)*.

Tao Lei, Yu Xin, Yuan Zhang, Regina Barzilay, and Tommi Jaakkola. 2014. Low-rank tensors for scoring dependency structures. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1381–1391, Baltimore, Maryland. Association for Computational Linguistics.

Teresa Lynn, Kevin Scannell, and Eimear Maguire. 2015. Minority language twitter: Part-of-speech tagging and analysis of Irish tweets. In *Workshop on Noisy User-generated Text*, Beijing, China.

Andre Martins, Miguel Almeida, and Noah A. Smith. 2013. Turning on the turbo: Fast third-order non-projective turbo parsers. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 617–622, Sofia, Bulgaria. Association for Computational Linguistics.

Anne-Lyse Minard, Manuela Speranza, and Tommaso Caselli. 2016. The EVALITA 2016 Event Factuality Annotation Task (FactA). In *Proceedings of Evalita 2016*.

Johanna Monti, Federico Sangati, Francesca Chiusaroli, Martin Benjamin, and Sina Mansour. 2016. Emojitalianobot and emojiworldbot - new online tools and digital environments for translation into emoji. In *Proceedings of Third Italian Conference on Computational Linguistics (CLiC-it 2016) & Fifth Evaluation Campaign of Natural Language Processing and Speech Tools for Italian. Final Workshop (EVALITA 2016)*, volume 1749 of *CEUR Workshop Proceedings*, Napoli, Italy. CEUR-WS.org.

Joakim Nivre, Marie-Catherine de Marneffe, Filip Ginter, Yoav Goldberg, Jan Hajič, Christopher D. Manning, Ryan T. McDonald, Slav Petrov, Sampo Pyysalo, Natalia Silveira, Reut Tsarfaty, and Daniel Zeman. 2016. Universal dependencies v1: A multilingual treebank collection. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation LREC 2016, Portorož, Slovenia, May 23-28, 2016*.

Olutobi Owoputi, Brendan OConnor, Chris Dyer, Kevin Gimpel, Nathan Schneider, and Noah A Smith. 2013. Improved part-of-speech tagging for online conversational text with word clusters. In *Proceedings of NAACL 2013*, pages 380–390.

Slav Petrov and Ryan McDonald. 2012. Overview of the 2012 Shared Task on Parsing the Web. Notes of the First Workshop on Syntactic Analysis of Non-Canonical Language (SANCL).

Luis Rei, Dunja Mladenić, and Simon Krek. 2016. A multilingual social media linguistic corpus. In *Proceedings of the 4th Conference on CMC and Social Media Corpora for the Humanities*, Ljubljana, Slovenia.

Djamé Seddah, Benoît Sagot, Marie Candito, Virginie Mouilleron, and Vanessa Combet. 2012. The French social media bank: a treebank of noisy user generated content. In *COLING 2012, 24th International Conference on Computational Linguistics, Proceedings of the Conference: Technical Papers, 8-15 December 2012, Mumbai, India*, pages 2441–2458.

Natalia Silveira, Timothy Dozat, Marie-Catherine De Marneffe, Samuel Bowman, Miriam Connor, John Bauer, and Chris Manning. 2014. A gold standard dependency corpus for English. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, Reykjavik, Iceland. European Language Resources Association (ELRA).

Fabio Tamburini and Matias Melandri. 2012. AnIta: a powerful morphological analyser for Italian. In *Proceedings of Language Resources and Evaluation Conference 2012*, pages 941–947.

Cristina Zaga. 2012. Twitter: un'analisi dell'italiano nel micro blogging. *Italiano LinguaDue*, 4(1):167–210.

Yue Zhang and Joakim Nivre. 2011. Transition-based dependency parsing with rich non-local features. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 188–193, Portland, Oregon, USA. Association for Computational Linguistics.

Yuan Zhang, Tao Lei, Regina Barzilay, and Tommi Jaakkola. 2014a. Greed is good if randomized: New inference for dependency parsing. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1013–1024, Doha, Qatar. Association for Computational Linguistics.

Yuan Zhang, Tao Lei, Regina Barzilay, Tommi Jaakkola, and Amir Globerson. 2014b. Steps to excellence: Simple inference with refined scoring of dependency trees. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 197–207, Baltimore, Maryland. Association for Computational Linguistics.