VALICO-UD

Treebank Annotation Guidelines in Universal Dependencies

version 1.1

Elisa Di Nuovo (elisa.dinuovo@unito.it), 2021

Overview:

1. Introduction
2. Tokenization, lemmatization, morphology, PoS tagging
3. Dependency annotation

1. Introduction

These guidelines are meant to briefly outline how the annotation is applied to the learner part, called Learner Sentences (LS), of the VALICO-UD corpus, released within the framework of Universal Dependencies (UD) for the first time in version 2.8. For each LS we will also use the parallel corrected sentence, called Target Hypothesis (TH), to highlight differences between the learner language and that of a 'ideal' mother-tongue speaker that are also made more evident in the annotation (see also guidelines for error annotation).
Since the LS part of VALICO-UD contains texts written by non-native speakers, it is necessary to make some choices to deal with non-canonical forms. In these guidelines, we are going to illustrate how we performed tokenization, lemmatization, morphology and dependency annotation in the LS part, especially focusing on what differs from the UD standard scheme.

2. Tokenization, lemmatization, morphology, PoS tagging

Tokenization
If words are mistakenly tokenized, we follow the UD scheme[1], since this kind of error is sporadic.

Wrongly split word:

EXAMPLE (1a)
```
Ieri al parco ha suceso un distrasto per che una donna è andata al parco […]
```

```
9       per     perche  SCONJ   CS      _       14      mark    _       _
10      che     _       X       _       _       9       goeswith        _       _
```

Wrongly merged word:

EXAMPLE (1b)
```
Nell parco non c'era nessunosolo io.
```

```
7       nessuno nessuno PRON    PI      Gender=Masc|Number=Sing|PronType=Ind  6
                nsubj   _       SpaceAfter=No|CorrectSpaceAfter=Yes
8       solo    solo    ADV     B       _       9       advmod  _       _
```

We do not mark misspelled words[2] because they are highly frequent errors, and the correct spelling can be retrieved either in the err field of the LS CoNLL-U file (see also error annotation guidelines) or in the TH file.

Lemmatization
Lemmatization follows standard rules but maintains spelling errors, if present.
For example, in (1a), *perche* is lemmatized without the graphic accent. In example (2a) *perchè* maintains the wrong accent in the lemma, while *cominzava* and *sapiava*, although they do not

[1] https://universaldependencies.org/u/overview/typos.html#wrongly-split-word
https://universaldependencies.org/u/overview/typos.html#wrongly-merged-words
[2] https://universaldependencies.org/u/overview/typos.html#misspelled-word

exist in Italian[3], are lemmatized following the Italian rules for verb lemmatization. Note that it is not possible with auxiliaries because it is a closed set of words that can be auxiliaries in the Italian UD framework (e.g., *Per lui solo ha **viuto** salvarla*. → *viuto* AUX VM can be lemmatized only as **volere**).

EXAMPLE (2a)

```
Ma la donna cominzava a gridare e Io non sapiava perchè.
```

```
1       Ma      ma      CCONJ   […]
2       la      la      DET     […]
3       donna   donna   NOUN    […]
4       cominzava       cominzare       VERB    […]
5       a       a       ADP     […]
6       gridare gridare VERB    […]
7       e       e       CCONJ   […]
8       Io      io      PRON    […]
9       non     non     ADV     […]
10      sapiava sapiare VERB    […]
11      perchè  perchè  NOUN    […]
12      .       .       PUNCT   […]
```

When a foreign word occurs, if it is contextually plausible, we lemmatize it with its foreign lemma and add 'Foreign=Yes' in the MISC column, as shown in (2b).

EXAMPLE (2b)

```
Il uomo era alto, forte e molto musculuso, ma Io può derribarle salvare a la
donna.
```

```
14      derribar        derribar        VERB […] _      Foreign=Yes
```

If it is contextually implausible, we lemmatize it accordingly to the PoS and we do not add 'Foreign=Yes' in the MISC column, as shown in (2c).

Example (2c)

```
Mi h'offerto du chiamare la polizia […]
```

```
4       du      du      ADP     E       _       5       mark    _       _
5       chiamare        chiamare        VERB    V       […]
```

Morphology

As specified in UD annotation scheme, if some tokens' morphological features are invariant, we do not mark these features in the dedicated column. So, for example, in (2a) *perchè* is a masculine noun, which is number invariant, hence we mark only the gender in the dedicated column, as shown in (3).

EXAMPLE (3a)

```
Ma la donna cominzava a gridare e Io non sapiava perchè.
```

```
11      perchè  perchè  NOUN    S       Gender=Masc     10      obj     […]
```

Conversely, if an invariant token displays a non-canonical suffix (3b), we annotate it literally maintaining the features that the learner gave to it.

[3] In the correspondent TH, we provide one of the possible interpretations, which in this case is *cominciare* and *sapere*, respectively.

EXAMPLE (3b)

```
Ieri al parco, un uomo con dei grossi muscoli avevano una fragila donna sulla
spalla.
```

```
14      fragila fragilo ADJ     A       Gender=Fem|Number=Sing 15      amod    […]
```

As a rule, if the orthographical and morphological form of the word respects its canonical form, then we mark its features in a standard way. This applies to nouns, verbs, and other parts of speech (PoS) which are not context dependent.

In (3c) we show a sentence in which we literally annotated the morphological features of a verb and a noun, without considering their distributional features, because we interpret this error as an agreement error involving the article and the noun.

EXAMPLE (3c)

```
La dona ringraziava suo salvatore con un braccio e chiusa le occhi.
```

```
2       dona    dona    NOUN […] Gender=Fem|Number=Sing […]
10      chiusa […] Gender=Fem|Number=Sing|Tense=Past|VerbForm=Part […]
11      le […] Definite=Def|Gender=Fem|Number=Plur|PronType=Art […]
12      occhi occhio NOUN S Gender=Masc|Number=Plur 10 obj _ SpaceAfter=No
```

We break this rule when we have overextension errors, such as in (3d), in which we have a gender invariant adjective which is likely used as a feminine plural (overextension of feminine plural suffix -e).

EXAMPLE (3d)

```
Avevo sentito delle parole forte, una donna sta gridando […]
```

```
5       forte   forto   ADJ     A       Gender=Fem|Number=Plur 4       amod    […]
```

In the same text, we find also another error involving the same suffix, reported in (3e). In this case, it is likely an agreement error. In fact, *fiore* in French—the learner's mother tongue—is a feminine noun, so the adjective's feminine suffix could be due to a negative interference of the L1. However, another interpretation could be that the learner uses the suffix -e as a marker of plural with no gender distinction, but the occurrences of this suffix in the text made us reject this hypothesis (e.g., *le ore piccole* [3-05_fr-3]). Also note the lemmatization of *uccele* in example 3e, which not only maintains the spelling errors but also the feminine gender (even though *uccele* likely refers to *uccelli*, noun masculine and plural).

EXAMPLE (3e)

```
La natura del parco mi sembra piu verde, i fiori piu aperte, le uccele
cantarono.
```

```
14      aperte  aperto  ADJ     A       Gender=Fem|Number=Plur 12      […]
17      uccele  uccela  NOUN    S       Gender=Fem|Number=Plur 18      […]
```

PoS tagging

As shown in the previous examples, we try to always annotate token's distributional PoS tags. We annotate the distributional tag for spelling errors, also those resulting in real words, as in (3c).

In cases of wrongly split words, we tag the first part with the distributional tag, the second with X, as shown in example (1a).

We annotate literally (morphologically, not distributional-based) if the word belongs to the closed-class set and the original PoS is inconsistent with the distributional one, as in (4a) in which a preposition is used instead of a conjunction.

EXAMPLE (4a)

```
Durante un ragazzo è passato.
```

1	Durante	durante	**ADP**	E	[…]
2	un	un	DET	RI	[…]
3	ragazzo	ragazzo	NOUN	S	[…]
4	è	essere	AUX	VA	[…]
5	passato	passare	VERB	V	[…]

Since our PoS tags are distributional based, we label cancelled words (marked with X in the text) with their distributional PoS tags[4], as in (4b).

EXAMPLE (4b)

```
L'uomo era brutto e ha la ragazza sull X.
```

6	ha	avere	VERB	V	[…]				
7	la	la	DET	RD	[…]				
8	ragazza	ragazza	NOUN	S	[…]				
9-10	sull	_	_	_	_	_	_	_	_
9	su	su	ADP	E	[…]				
10	l	lo	DET	RD	[…]				
11	X	_	**NOUN**	S	[…]				
12	.	.	PUNCT	FS	[…]				

3. Dependency annotation

For this level of annotation too, when a non-canonical form appears, we annotate it according to the co-text (considering the distributional slot) but keeping in consideration the morphological features and the annotation possibilities offered by the standard language.

Reusing the example 2b, in Figure 3.1a, we show its dependency tree, and the attributes of the node *le*. *Derribar* is a Spanish transitive verb, which is semantically appropriate in the sentence context, then we annotate *le* as direct object of *derribar*, but since *le* in Italian, when direct object refers to feminine plural nouns, we mark these features in the appropriate column as shown in **feats,** on the left of Figure 3.1a. If compared with its correspondent TH tree, shown in figure 3.1b, it is possible to infer the provided sentence interpretation, in which *lo* correspond to *le* and the antecedent of *lo* can be easily identified with *uomo*.

[4] Currently, we have only one occurrence of this phenomenon.

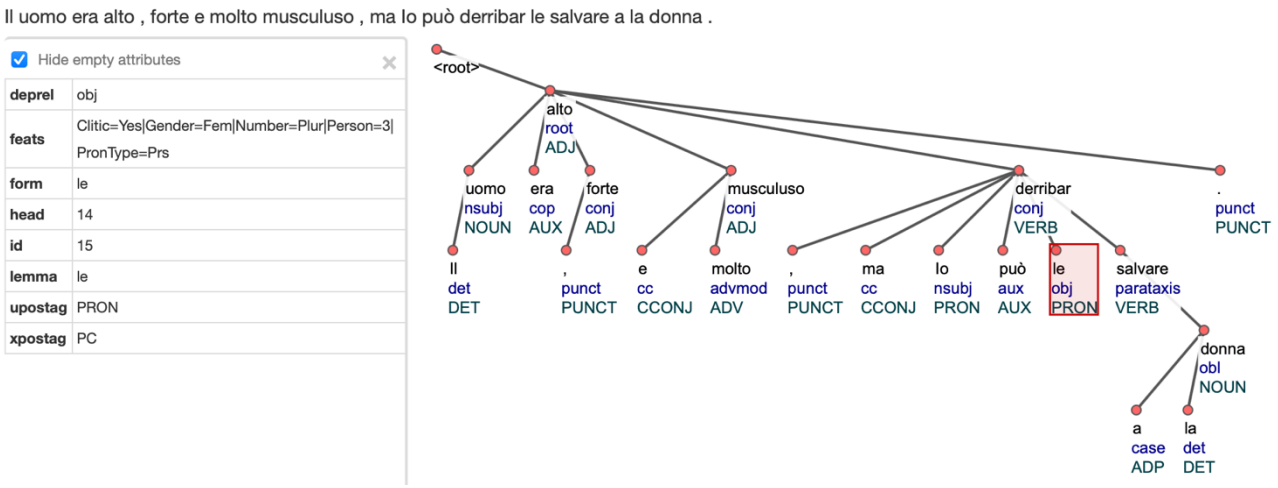

Figure 3.1a: Dependency tree of Example (2b).

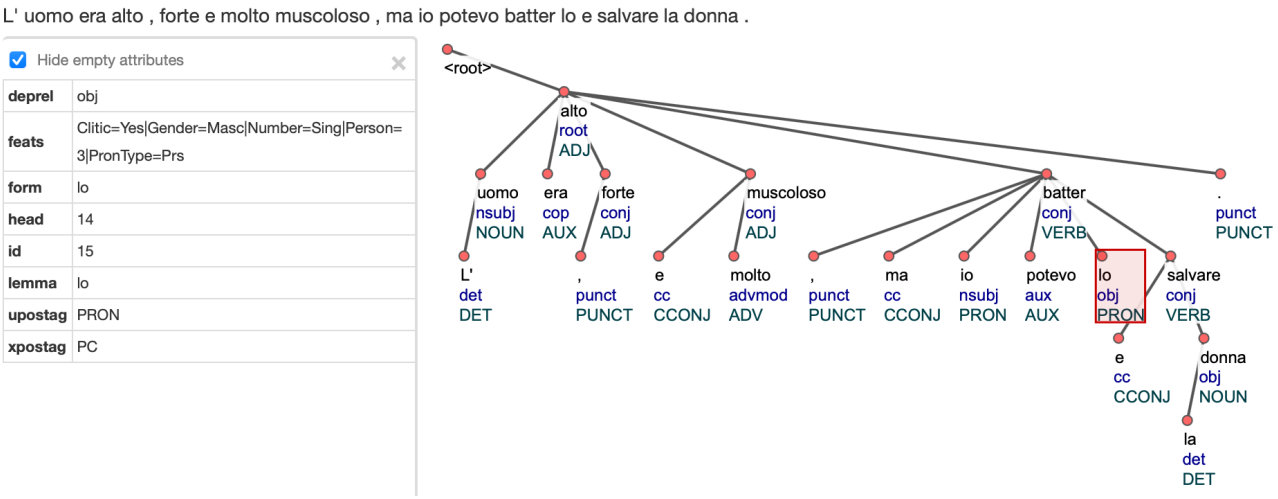

Figure 3.1b: TH version of dependency tree 3.1a.

Sometimes, learners modify verb argument structures adding or omitting some elements which can change the tree. For example, in Figure 3.2a, we show a sentence in which a noun is used as an indirect object, while in in the TH it is replaced by an oblique (Figure 3.2b).

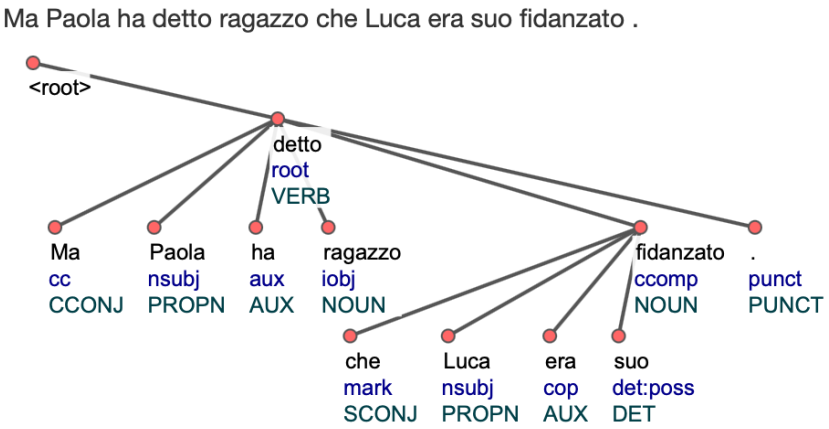

Figure 3.2a: Dependency tree *iobj*.

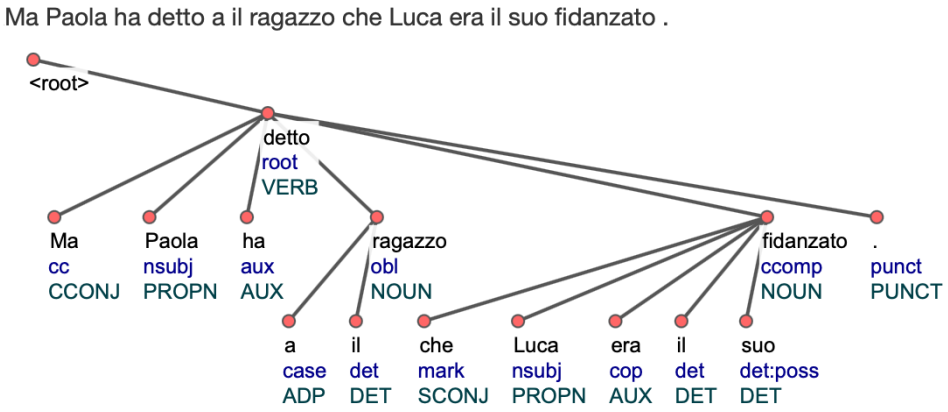

Figure 3.2b: TH version of dependency tree 3.2a.

In Figure 3.3a-b, we show an example of an oblique which corresponds to a direct object (TH).

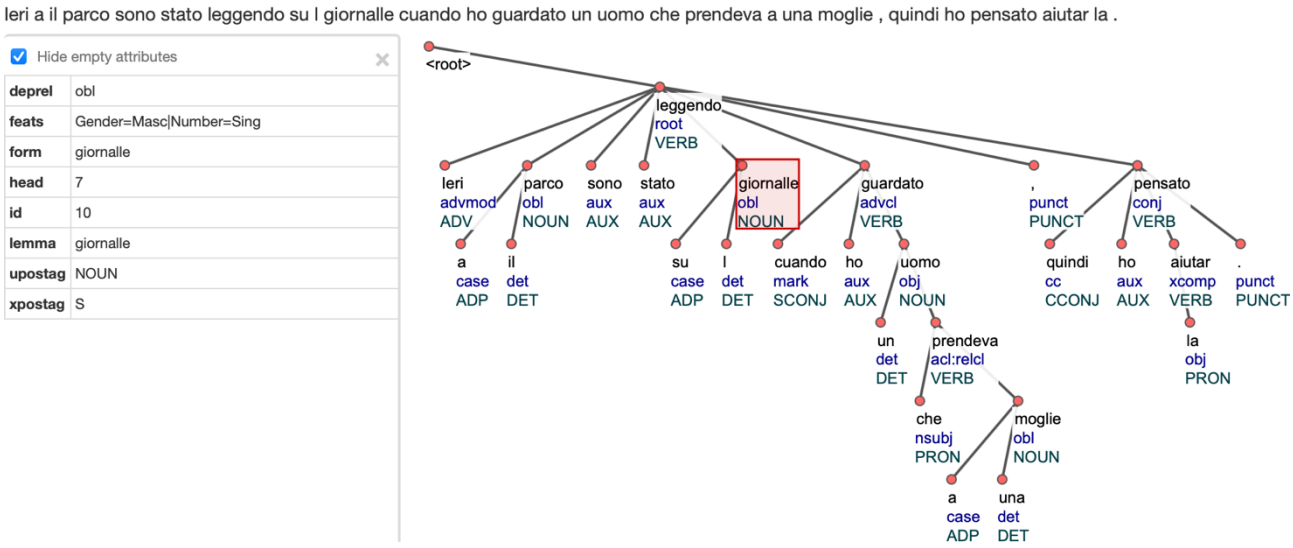

Figure 3.3a: Dependency tree *obl*.

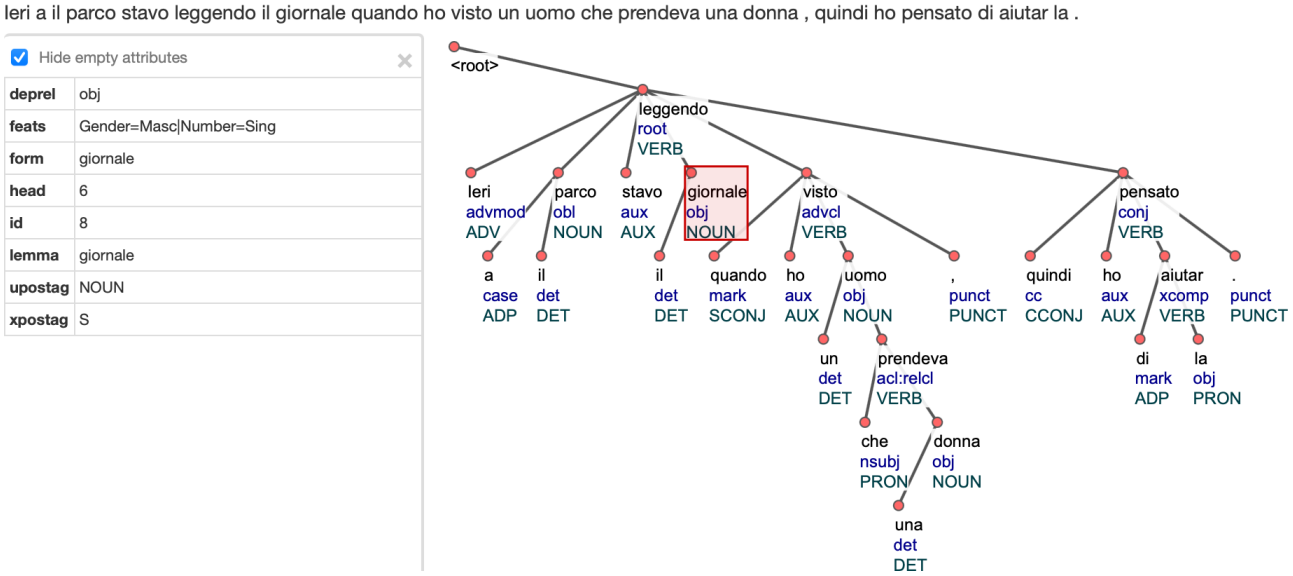

Figure 3.3b: TH version of dependency tree 3.3a.

We always try to annotate the sentence as we can read it literally without interpreting. So, in Figure 3.4a-b we show an example in which a different final vowel changes the resulting tree. It is worth noticing that adverbs are invariant, so *molte* in 3.4a is annotated as a pronoun, despite it is substituted by an adverb in the TH (3.4b).

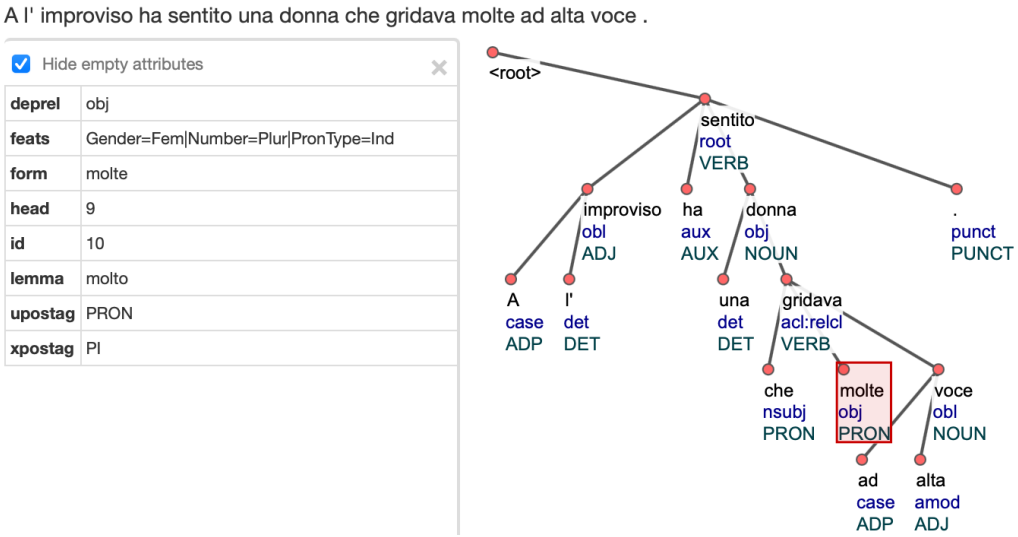

Figure 3.4a: Dependency tree of *molte*.

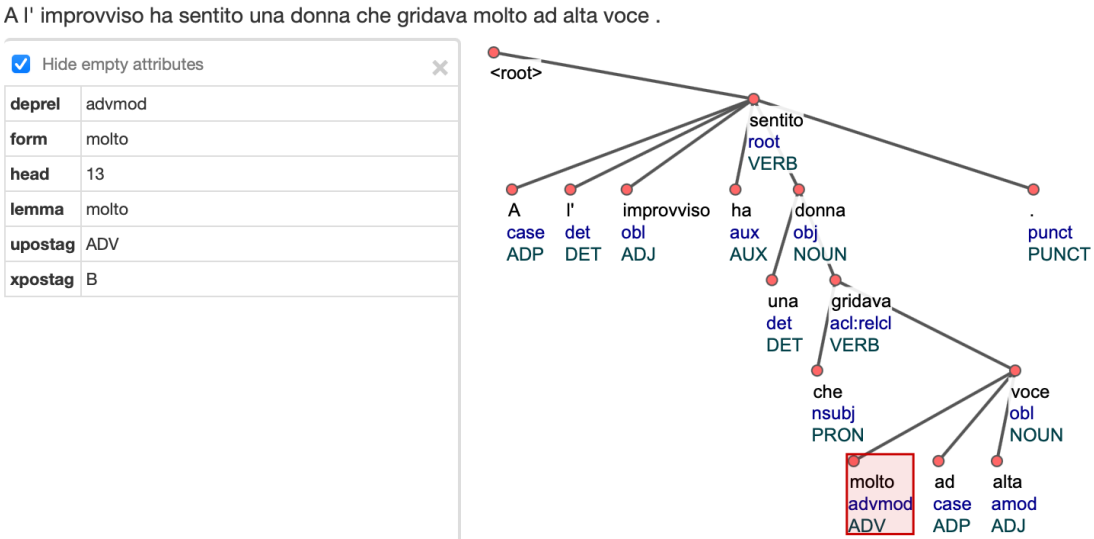

Figure 3.4b: TH version of dependency tree 3.4a.

If it is not clear the (logical) syntactical function of an element, we annotate it with the generic dependency relation label *dep*, as shown in Figure 3.5a-b.

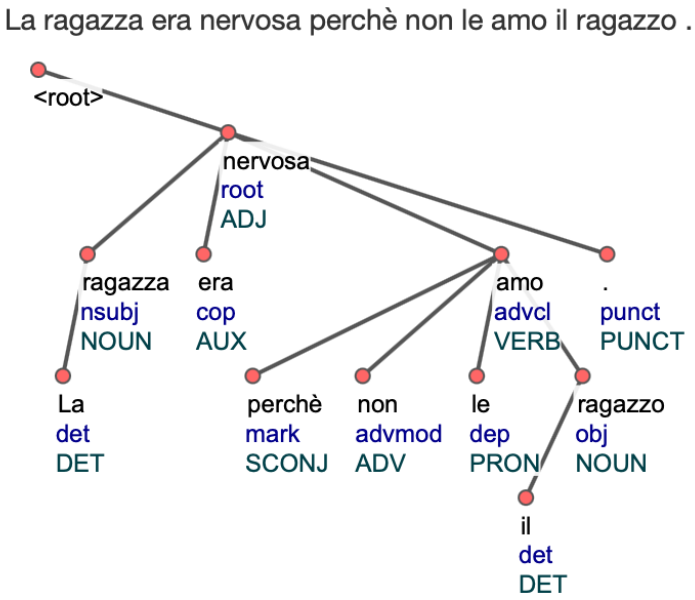

Figure 3.5a: Dependency tree of *dep.*

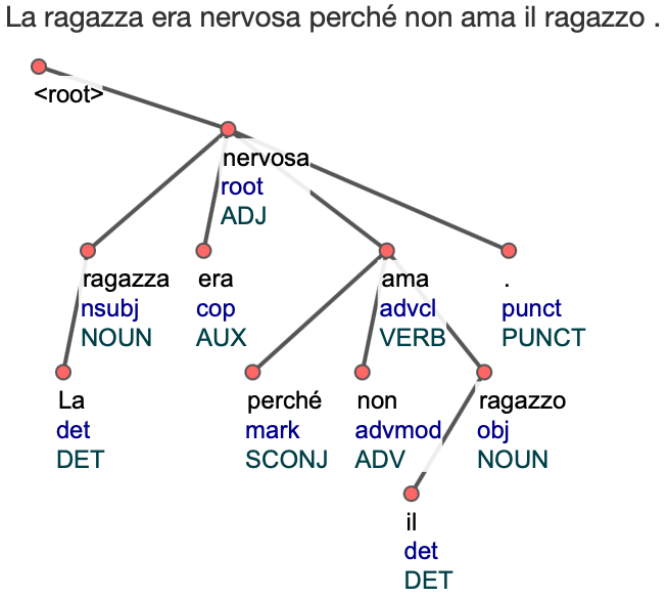

Figure 3.5b: TH version of dependency tree 3.5a.