ISSN 2499-4553

IJCoL

Italian Journal of Computational Linguistics

Rivista Italiana di Linguistica Computazionale

Volume 8, Number 1
june 2022

aA Accademia university press

Registrazione presso il Tribunale di Trento n. 14/16 del 6 luglio 2016

Rivista Semestrale dell'Associazione Italiana di Linguistica Computazionale (AILC)

Associazione Italiana di Linguistica Computazionale

direttore responsabile
Michele Arnese

isbn 9791255000167

Accademia University Press
via Carlo Alberto 55
I-10123 Torino
info@aAccademia.it
www.aAccademia.it/IJCoL_8_1

IJCoL

Volume 8, Number 1
june 2022

CONTENTS

VALICO-UD: Treebanking an Italian Learner Corpus in Universal Dependencies

Elisa Di Nuovo*
Università degli Studi di Torino

Manuela Sanguinetti**
Università degli Studi di Cagliari

Alessandro Mazzei†
Università degli Studi di Torino

Elisa Corino‡
Università degli Studi di Torino

Cristina Bosco§
Università degli Studi di Torino

This article describes an ongoing project for the development of a novel Italian treebank in Universal Dependencies format: VALICO-UD. It consists of texts written by Italian L2 learners of different mother tongues (German, French, Spanish and English) drawn from VALICO, an Italian learner corpus elicited by comic strips. Aiming at building a parallel treebank currently missing for Italian L2, comparable with those exploited in Natural Language Processing tasks, we associated each learner sentence with a target hypothesis (i.e. a corrected version of the learner sentence written by an Italian native speaker), which is in turn annotated in Universal Dependencies. The treebank VALICO-UD is composed of 237 texts written by non-native speakers of Italian (2,234 sentences) and the related target hypotheses, all automatically annotated using UDPipe. A portion of this resource (36 texts corresponding to 398 learner sentences and related target hypotheses)—firstly released on May 2021 in the Universal Dependencies repository—is associated with error annotation and the automatic output is fully manually checked. In this article, we focus especially on the challenges addressed in treebanking a resource composed of learner texts. In addition, we report on a preliminary data exploration that makes use of three quantitative measures for assessing the quality of the data and for better understanding the role that this resource can play in tasks lying at the intersection of Computational Linguistics and learner corpus studies.

1. Introduction

Learner corpora, also called interlanguage (Selinker 1972) or L2 corpora, are collections of data produced by foreign or second language learners (Granger 2008, for a detailed description). Most learner corpora projects were launched in the nineties and focused

* Dipartimento di Lingue e Letterature Straniere e Culture Moderne - via Giuseppe Verdi fronte 41bis, 10124 Torino, Italy. E-mail: elisa.dinuovo@unito.it
** Dipartimento di Matematica e Informatica - via Ospedale 72, 09124 Cagliari, Italy. E-mail: manuela.sanguinetti@unica.it
† Dipartimento di Informatica - corso Svizzera 185, 10149 Torino, Italy. E-mail: mazzei@di.unito.it
‡ Dipartimento di Lingue e Letterature Straniere e Culture Moderne - via Giuseppe Verdi fronte 41bis, 10124 Torino, Italy. E-mail: elisa.corino@unito.it
§ Dipartimento di Informatica - corso Svizzera 185, 10149 Torino, Italy. E-mail: bosco@di.unito.it

mainly on learner English (Tono 2003). However, in the last twenty years, learner corpus research has experienced a rapid growth and the number of publications and learner resources has meaningfully grown.[1]
Learner corpora can be used for a variety of activities—e.g. evidence-based learning, for tracing acquisition—and by a variety of users—e.g. teachers and learners, second language acquisition researchers and curriculum developers (Díaz-Negrillo, Ballier, and Thompson 2013). In addition, learner corpora have been embraced by computational linguists and computer scientists for developing language models and NLP tools, respectively.

In the last few years, several learner corpora have been compiled specifically for computational tasks such as Native Language Identification (NLI) or Grammatical Error Identification and Correction (GEI and GEC). For instance, the English L2 corpus specifically compiled for NLI, called TOEFL11 (Blanchard et al. 2013) and the English L2 corpus compiled for GEC, called NUCLE (Dahlmeier, Ng, and Wu 2013) or the dataset called Write&Improve+LOCNESS (Bryant et al. 2019). But also non-English corpora, such as German, Portuguese, and Spanish have been recently developed for computational tasks (Köhn and Köhn 2018; Del Río Gayo, Zampieri, and Malmasi 2018; Davidson et al. 2020). As far as Italian learner language is concerned, there are not corpora suitable to be used as a benchmark for NLP tasks, as better explained in Sections 2 and 4.1.

To fill this gap, in this article we present VALICO-UD, a new richly-annotated Italian learner treebank. VALICO-UD is composed of 237 texts (2,234 sentences)—written by English, French, Spanish and German native speakers—drawn from the learner corpus VALICO (Corino and Marello 2017), a collection of non-native Italian texts elicited by comic strips.[2] The linguistic annotation is based on the *de facto* standard format of Universal Dependencies (henceforth, UD), but it also includes an XML-based error annotation and the association of each learner sentence with a corresponding correct version (i.e. *Target Hypothesis*) generated by an Italian native speaker and in turn annotated in UD. In practice, the result is a parallel treebank consisting of Learner Sentences (LSs), i.e. sentences produced by Italian L2 speakers, and their corresponding Target Hypotheses (THs).

While all LSs and THs were automatically annotated using UDpipe (Straka 2018), a portion of the resource is also associated with error annotation and fully manually corrected (i.e. tokenization, lemmatization, Part-of-Speech tagging, morphology features and parsing). This portion constitutes the core gold standard of VALICO-UD: it consists of 36 texts (398 sentences) and it was released on May 2021 in the Universal Dependencies repository.[3] The remaining 201 texts (1,836 sentences) are available in a github repository as a silver standard dataset.[4]

In this article, we focus on describing the VALICO-UD design and the challenges addressed during the application of the UD format to the VALICO-UD texts, leaving

1 Learner corpus bibliography updated on a regular basis (last update: 4 June 2021): https://uclouvain.be/fr/node/12074; Learner corpora around the world list: https://uclouvain.be/en/research-institutes/ilc/cecl/learner-corpora-around-the-world.html.
2 VALICO texts are publicly available and downloadable from here: www.valico.org.
3 Universal Dependencies repository: https://universaldependencies.org; VALICO-UD downloadable here (THs in the folder named corrected): https://github.com/UniversalDependencies/UD_Italian-Valico/.
4 The CoNLL-U files automatically annotated with UDPipe are available here: https://github.com/ElisaDiNuovo/VALICO-UD_silver/.

aside the discussion on error annotation (which will be discussed in detail in another article in preparation). Furthermore, as an example of application of this resource, we report on a study of language development based on quali-quantitative features extracted from the novel parallel treebank using existing NLP tools.

The remainder of this article is organized as follows: in Section 2, we survey learner corpora, focusing on types of annotation useful for NLP tasks and introducing the related issues; in Section 3, we present an overview of VALICO-UD data; in Section 4 we describe the annotation scheme and the challenges addressed in tokenization, lemmatization, PoS tagging, morphology features and dependency annotation; in Section 5, we apply some quantitative measures and analysis to identify the characteristics of learner language included in the treebank; and finally, in Section 6, we conclude the article and present directions of future work.

2. Background and Motivation

Most learner corpora are collected during language proficiency examinations, thereby they are usually built in collaboration with language assessment centres, e.g. the Cambridge Learner Corpus (CLC). Since learner products are usually collected during language assessment tests, they mainly contain essays, e.g. International Corpus of Learner English (ICLE)[5] and CLC. Others, called peripheral learner corpora, using the term proposed by Nesselhauf (2004), are picture-elicited. Picture-elicited corpora are less authentic than free compositions or essays because learners cannot freely express themselves but are obliged to follow the comic strips. However, although this can be an issue if we consider authenticity as defined in the recommendation on corpus and text typology (EAGLES 1996, p. 7),[6] it is useful when providing a correction for non-canonical words or structures (since vocabulary and semantics are circumscribed by the pictures).

Learner corpora can be used either raw or annotated. Even though studies based on raw data are feasible (Aijmer 2002; Nesselhauf 2004), they necessarily focus on limited features that can be easily retrieved. In fact, learner corpora have a much greater potential if specific language properties have been previously identified and annotated (Díaz-Negrillo and Thompson 2013, p. 13–16). Annotation, which is essential to make explicit what is implicit in texts, usually involves tokenization, lemmatization, PoS tagging, syntactic parsing, and semantic tagging (Garside, Leech, and McEnery 1997; McEnery and Wilson 2001).

A type of annotation that has been recently associated to learner corpora are explicit THs, on which error annotation can be applied. To preserve LSs as they are, but still be able to automatically retrieve information about them, learner corpora can be built as **parallel corpora** associating to each LS the corresponding TH, keeping the LS separated from its TH—as recommended by many scholars (Lüdeling 2008; Reznicek, Lüdeling, and Hirschmann 2013; Meurers 2015). We followed this strategy in VALICO-UD.

5 Corpus webpage: `https://uclouvain.be/en/research-institutes/ilc/cecl/icle.html`.

6 Authentic data is "gathered from the genuine communications of people going about their normal business". Since learner corpora usually collect texts during language examinations, according to the recommendation, they should be considered *special corpora*, because they "involve the linguist beyond the minimum disruption required to acquire data". In fact, the linguist is the one who set the task, not only the one who acquire data. In particular, picture-elicited texts are even less authentic than free compositions or essays.

Notwithstanding the presence of explicit THs, to smoothly retrieve interlanguage features, other linguistic annotation might be needed. As Meurers and Müller (2009) proved, morpho-syntactically annotated corpora—i.e. **treebanks**—"can serve as an important component of empirically grounded syntactic research". However, the unpredictability and variation of a learner product—in terms of vocabulary, morphology and syntax—must be taken into consideration because they make the annotation a challenging task (see for example the study of Díaz-Negrillo *et al.* (2010) about PoS tagging and the studies of Astaneh and Frontini (2009) and Corino and Russo (2016) about parsing Italian L2 corpora).

As far as Italian is concerned, the only Italian learner corpus syntactically and error annotated is the MERLIN corpus (Boyd et al. 2014) a trilingual learner corpus (German, Czech and Italian as L2) containing 813 learner Italian texts, all automatically annotated using a parser trained on native language and associated to a TH correcting only part of the errors taken into account in GEI and GEC tasks (see Section 4.1). Thus, there is not a gold standard Italian learner treebank available for training tools and accomplishing tasks such as GEC, GEI or NLI, neither for training and evaluating the performance of parsing systems. For this reason, we compiled VALICO-UD, considering in the design criteria also those necessary to exploit this resource for computational tasks and not only for linguistic research.

In addition, it is important to choose a suitable annotation framework. Among the existing frameworks, we opted for the UD formalism (Nivre et al. 2020) and apply it to VALICO-UD for two reasons. First, in the last ten years, UD has emerged as a *de facto* standard for morphological and syntactic annotation, with 202 treebanks in 114 languages.[7] Second, among the published treebanks, there are also two learner treebanks, namely the English Second Language (ESL) (Berzak et al. 2016) and the Chinese Foreign Language (CFL) (Lee, Leung, and Li 2017).

The development of a treebank on top of an Italian learner corpus is especially challenging because the language addressed is non standard under several respects. Following an interlanguage approach, learner language should be annotated as a language by itself, so models should be trained on interlanguage varieties and not on native texts. However, the only available gold standard texts so far are made of standard samples, produced by Italian native speakers. This highlights the need for a gold standard of learner Italian as well, that enables the training of *ad hoc* models, in line with the interlanguage approach.

As a side effect, by addressing this challenge we can also collect evidences useful to investigate how much and where automatic analysis can be especially hard, and subsequently to gain a greater awareness of the usability of our resource for training and testing parsing tools.

Since a common practice in learner corpora is to automatically annotate texts using models trained on native texts (e.g. PoS tags in VALICO and PoS tags and syntactic relations in MERLIN), it would be indeed interesting also to measure the performance loss of these parsing systems. This would be useful not only to know to what point we can draw linguistic conclusions starting from automatically annotated learner texts, but also to better know the problem and find possible solutions (e.g. if a foreign language learner writes to a chatbot using its target language, what kind of problems will the chatbot encounter to analyse the learner's input?). In order to measure the performance loss of these parsing systems, it is essential to have a gold standard to evaluate the

7 Data referred to version 2.8: `https://universaldependencies.org`.

automatic annotation. To fill this gap, we describe here the annotation schema and the data collection of VALICO-UD, a novel parallel learner Italian treebank in which dependency syntactic relations are annotated, following UD formalism, both in LSs and THs.

3. VALICO-UD design

Inspired by existing resources developed for other learner languages and the extensive literature on this topic, VALICO-UD has been designed as a novel resource to investigate Italian learner language from several different perspectives. In this section, we describe the typology of texts collected in the corpus and provide basic statistics on the composition of the treebank.

We have drawn the texts from the VALICO corpus (Corino and Marello 2017) for three main reasons. First, because it is the biggest learner Italian corpus publicly available. Second, because it is a collection of non-native Italian texts elicited by comic strips,[8], hence facilitating the reconstruction of THs when non-canonical words or structures occur, because lexical choices and semantic frames are circumscribed by the comic strip (Corino and Marello 2009; Marello 2011). And third, because it collects a wide variety of metadata, hence enabling the creation of subcorpora following precise design criteria.

To build the VALICO-UD parallel treebank as a resource suitable for tasks like those cited above, we adopted two main design criteria, L1 selection and topic selection, but also the learners' year of study has been considered.
According to the first criterion, **learners' L1**, we selected texts written by German (DE), English (EN), Spanish (ES) and French (FR) native speakers. Eventually, we obtained a selection of VALICO comprising 237 texts (2,234 sentences) as shown in Table 1.
Considering the second criterion, we selected the data referred to two different comic strips, each about a different **topic**: one eliciting more descriptive texts and another mostly narrative ones.

Table 1
Summary of VALICO-UD composition.

L1	# Texts	# Sentences	# LS Tokens	# TH Tokens
DE	58	622	8,729	8,838
EN	60	662	9,834	10,029
ES	59	381	8,270	8,361
FR	60	569	8,623	8,686
EN+FR+DE+ES	237	2,234	35,456	35,914

As previously mentioned, we applied the annotation scheme to the collection in its entirety (whose composition is summarized in Table 1), and then validated to a portion of such data (see Table 2), that we define as the *core gold section* of the treebank.
The VALICO-UD core section is composed of texts elicited by one comic strip, namely the one entitled "Ieri al parco..." (*Yesterday at the park...*), shown in Figure 1.

8 Comic strips available here: http://www.valico.org/vignette.html.

Table 2
Summary of VALICO-UD core section.

L1	# Texts	# Sentences	# LS Tokens	# TH Tokens
DE	9	93	1,191	1,201
EN	9	150	2,382	2,388
ES	9	77	1,864	1,878
FR	9	78	1,347	1,365
EN+FR+DE+ES	36	398	6,784	6,832

The selected comic strip includes a series of four drawings without written words. The first drawing shows a man A reading a newspaper, which is suddenly interrupted by another man B carrying a crying woman; the second drawing shows the man A that decides to intervene; in the third one, the man A seems happy, while the man B is lying on the ground, and the woman is between astonished and worried; finally, in the last and fourth drawing, the furious woman (whose finger points downwards) seems to be arguing with the man A and over the woman's head there is a balloon with a heart.

Figure 1
"Ieri al parco..." (*Yesterday at the park...*) comic strip from VALICO.

As for the criterion regarding the **learner's year of study**, in Table 3, we report a summary of the texts sorted according to it—mean and standard deviation in brackets.

In conclusion, we collected 9 texts per each L1 elicited by the selected comic strip. As far as the year of study is concerned, for ES and FR learners we could not find exactly three texts for each year of study as we did for DE and EN. Therefore, for what concerns ES texts, we collected three texts of the first and three of the second year of study. Then, we collected one text of the third year of study, one text of the fourth year of study, and one text without explicit year of study (these are grouped in Table 3 and marked with the asterisk in Year of Study column). For what concerns FR texts, we could not find

Table 3
Core section summary according to selection criteria (mean and standard deviation in brackets).

L1	# texts	Year of Study	# Sentences	# LS Tokens
DE	3	1	33 (11±3.5)	401 (133.3±13.7)
DE	3	2	30 (10±1.7)	391 (130.3±12.7)
DE	3	3	30 (10±2.6)	399 (133±3.0)
EN	3	1	77 (25.7±13.3)	1,099 (366.3±213.6)
EN	3	2	26 (8.7±1.5)	433 (144.3±31.9)
EN	3	3	47 (16.7±17.6)	850 (283.3±290.9)
ES	3	1	31 (10.3±4.5)	673 (223.3±73.7)
ES	3	2	28 (9.3±3.5)	898 (299±233.4)
ES	3	*	18 (6±3.6)	293 (97.7±54.9)
FR	3	1	22 (7.3±1.5)	343 (114±11.1)
FR	3	3	25 (8.3±0.6)	479 (159.7±43.4)
FR	3	4	31 (10.3±3.2)	525 (174.7±63.4)

text of the second year of study, then we selected three of the first, three of the third and three of the fourth year of study. As it can be noted from the table, first and third year EN texts are those with a higher variation both in number of sentences and number of tokens, while second year ES texts vary highly only in number of tokens. In these three groups with high variation, there are three texts, one text per group, that increase the variation because learners wrote an introduction about the man A—which is usually considered by learners as the main character of the story—before narrating the story described in the comic strip.

4. Annotation scheme

In this section we describe the annotation we applied on VALICO-UD, which includes an explicit TH for each LS and a linguistic annotation in UD format on the whole data (i.e. both LSs and THs) to make VALICO-UD a parallel treebank.
For accomplishing this task, we fed the corpus to a UDPipe model (Straka 2018), trained on ISDT (Simi, Bosco, and Montemagni 2014) and PoSTWITA-UD (Sanguinetti et al. 2018) treebanks, the same used in (Cignarella et al. 2020). This was motivated by the fact that ISDT is the reference treebank for training parsers on standard Italian (Zeman et al. 2018), while PoSTWITA—which is made of Twitter data—displays features in common with learner language (e.g. spelling issues and uncontrolled syntax).

In the next subsections we first describe the challenges we addressed for writing the THs, and then the application of UD formalism to learner language, presenting examples drawn from the VALICO-UD core section.

4.1 Target Hypotheses

As stated in different studies, explicit THs can improve the replicability of the analysis applied on learner texts (Lüdeling 2008; Reznicek, Lüdeling, and Hirschmann 2013; Meurers 2015). Therefore, in VALICO-UD we decided to include the generation and parallel annotation on explicit THs, which are currently available for the whole resource

(see Table 1). In accomplishing this task, we applied some strategies to keep at the same time the THs as semantically close as possible to the corresponding LSs and away from the subjectivity of the annotator.

Even though the texts are elicited by comic strips (which often makes it easier to understand the meaning of learners' utterances), in writing the THs, it is crucial to limit the effect of subjective judgement of the annotator that generates them. We thus referred to a set of resources, comprising reference corpora, a descriptive grammar and a dictionary.

The first resource is VINCA, which is a small reference corpus specifically compiled for VALICO.[9] It includes texts elicited by the same comic strips of VALICO, but generated by Italian native speakers. In particular, from VINCA we extracted the subcorpus of 181 texts elicited by the same two comic strips used for VALICO-UD.

Moreover, in order to have a greater coverage of structures, we decided to refer to CORIS (Rossini Favretti, Tamburini, and De Santis 2002), and to the Italian treebanks available in UD (Simi, Bosco, and Montemagni 2014; Sanguinetti and Bosco 2015; Alfieri and Tamburini 2016; Sanguinetti et al. 2018; Cignarella, Bosco, and Rosso 2019) where native speakers texts are collected to represent standard Italian productions.

In addition, we referred to some general purpose resources, and mainly to the De Mauro's Dictionary[10] (De Mauro 2016) and to the Italian reference grammar *Grande Grammatica Italiana di Consultazione* (Renzi, Salvi, and Cardinaletti 2001).

All these resources were used by the Italian native speaker for writing the THs. In particular, for each LS a TH is written which differs from the LS if grammatical errors are encountered—considering as grammatical also orthographical and semantic well-formedness, and acceptability (James 1998, p. 66–74)—excluding appropriateness errors, i.e. those involving pragmatics, register, and stylistic choices (Lüdeling and Hirschmann 2015, p. 140). Once that the native speaker detects a contentious case, this must be carefully searched in the reference resources to check its validity and to avoid subjective judgments in deciding its ungrammaticality. If the contentious case results ungrammatical, a corrected version must be written, bearing in mind the intended meaning of the learner's utterance. On the other hand, dealing with picture-elicited texts does not necessarily make this task easier; as a matter of fact, meaning can be corrected in multiple ways. Thereby, to decide the corrected version of the non-canonical forms encountered, we applied the principle of *similarity*: when more than one correct alternative is admissible, the TH having a larger set of features (lexical, morphological, syntactic, and semantic) in common with the LS is selected. In Example 1, we report a case in which the intended meaning is clear, but it can correspond to at least two different THs.

(1) **LS:** Ho visto un uomo palestrato **portando** sulle spalle alla ragazza.
TH1: Ho visto un uomo palestrato **che portava** sulle spalle la ragazza.
TH2: Ho visto un uomo palestrato **che stava portando** sulle spalle la ragazza.
I saw a fit man carrying the girl on his shoulders.

According to the similarity criterion introduced above, when dealing with the error reported in Example 1, we decided to choose TH2, because the learner's *signifier* (using de Saussure's terminology) is maintained, along with other grammatical features, such as the continuous aspect of the verb using the same verb form.

9 VINCA is available here: `http://www.valico.org/vinca.html`.
10 The dictionary is accessible here: `https://dizionario.internazionale.it`.

In this respect, our THs differ from the first TH provided in MERLIN, the other error-annotated publicly-available learner Italian corpus, mentioned in Section 2. We decided to include in the THs also lexical, semantic and acceptability errors, excluded in MERLIN first TH, because they are usually considered in GEC tasks.[11]

It is worth noting that having an explicit TH for each LS, we actually develop a parallel corpus which is essential when GEC is approached as a machine translation task. In addition, by applying UD formalism to LSs and THs, written following the above-described principles, we build a parallel treebank which enables a new methodology of analysis of learner data (which differs from the Contrastive Interlanguage Analysis usually carried out in learner corpus research) and improves its replicability (Doval and Nieto 2019; Lee, Li, and Leung 2017).

4.2 UD formalism

In this subsection we introduce the UD framework. As stated in the introductory page of the project,[12] UD is a project that aims at "developing cross-linguistically consistent treebank annotation for many languages, with the goal of facilitating multilingual parser development, cross-lingual learning, and parsing research from a language typology perspective".

The UD format, usually shown in CoNLL-U encoding, starts with metadata lines (e.g. sentence identification and sentence raw text), blank lines indicating sentence boundaries, and word lines containing morphological and syntactical information about each word/token annotated in ten columns separated by a single tab. Thereby, a sentence consists of one or more word lines, and word lines are composed of the following columns:

1. **ID** contains an integer number identifying the token. The identifier of the first token of each new sentence is 1. It may be a range for multi-word tokens (see first column in Example 2).
2. **FORM** contains the word/token form (i.e. signifier) or punctuation symbol (see second column in Example 2).
3. **LEMMA** contains the lemma of the word form (see third column in Example 2).
4. **UPOS** contains the PoS tag (see fourth column in Example 2).[13]
5. **XPOS** contains the language-specific PoS tag (see fifth column in Example 2).
6. **FEATS** contains a pipe separated list of morphological features from the universal feature inventory or from a defined language-specific extension (features are not reported in the sixth column of Example 2 for space reasons).[14]

11 We plan to create a second TH covering also pragmatics, register and stylistic errors.
12 Available here: `https://universaldependencies.org/introduction.html`.
13 The complete list of Universal PoS tags is available here: `https://universaldependencies.org/u/pos/index.html`.
14 Universal features are listed on this page: `https://universaldependencies.org/u/feat/index.html`; Features allowed for each language are indicated here: `https://universaldependencies.org/ext-feat-index.html`.

7. **HEAD** contains the ID of the current word/token's governor. It is `0` if the token is the root (see seventh column in Example 2).
8. **DEPREL** contains the universal dependency relation to the HEAD (see eighth column in Example 2).
9. **DEPS** contains the enhanced dependency graph in the form of a list of HEAD-DEPREL pairs. In VALICO-UD, this column is not used like in other resources where enhanced dependency relations are not annotated (for space reasons it is deleted in Example 2).
10. **MISC** contains any other annotation, including information about the absence of spaces after the token (see ninth column in Example 2).

To summarize, morphological annotation is included in four columns (i.e. LEMMA, UPOS, XPOS, FEATS), syntactic annotations in three (HEAD, DEPREL, DEPS). In Example 2, a sample of CoNLL-U is presented (where the ninth column is deleted and features in the sixth are substituted with `[...]` for space reasons). Underscores indicate unspecified values.[15] In case of multi-word tokens (e.g. the article-preposition contraction `sulla` in sent_id 1), the ID column contains a range (in the example, `5-6`), while all the other columns except FORM are left empty (i.e. contain an underscore). The tokens composing the multi-word token are then separately analyzed in other word lines (i.e. *su* and *la*). The `SpaceAfter=No` attribute is finally used to mark the absence of a space between words when they do not form a multi-word token.

(2)

sent_id = 5-06_fr-3
text = Strappava **Marco,** toccava **sulla** sua testa e Marco **cadeva.**
[*He tore Marco, touched his head and Marco fell.*]

1	Strappava	strappare	VERB	V	[...]	0	root	_
2	**Marco**	Marco	PROPN	SP	_	1	obj	**SpaceAfter=No**
3	,	,	PUNCT	FF	_	4	punct	_
4	toccava	toccare	VERB	V	[...]	1	conj	_
5-6	**sulla**	_	_	_	_	_	_	_
5	**su**	su	ADP	E	_	8	case	_
6	**la**	la	DET	RD	[...]	8	det	_
7	sua	suo	DET	AP	[...]	8	det:poss	_
8	testa	testa	NOUN	S	[...]	4	obl	_
9	e	e	CCONJ	CC	_	11	cc	_
10	Marco	Marco	PROPN	SP	_	11	nsubj	_
11	**cadeva**	cadere	VERB	V	[...]	1	conj	**SpaceAfter=No**
12	.	.	PUNCT	FS	_	1	punct	**SpacesAfter=\n**

This introduction to the UD format is functional for the understanding of the annotation choices described in what follows.

15 All details about UD format are available at `https://universaldependencies.org/guidelines.html`.

4.3 Treebanking VALICO in UD

In this section we describe in detail the annotation challenges and consequent annotation choices made to adapt the UD format to Italian L2, benefiting from the experience gained from the annotation of the core section of VALICO-UD.

Tokenization – Errors involving tokenization can be encoded in the text or due to the tokenizer (i.e. the annotation pipeline component that is in charge of the tokenization). The latter can occur in presence of multi-word tokens, which in Italian are mainly represented by preposition-article and verb-clitic contractions. The former can occur with any word and can be the direct consequence of typing issues or of insufficient knowledge of the language. Typing errors are defined as performance errors—those called in the literature also as mistakes (Corder 1967, 1971)—and can also occur in native language. The other ones are defined as competence errors (making use of Chomsky's distinction) and mostly occur in learner language (L1 or L2 learners). Both performance and competence errors can produce two types of tokenization errors: hypersegmentation (i.e. wrongly split words) and hyposegmentation (i.e. wrongly merged words) (Sparrow 2014).

The presence of hyposegmented and hypersegmented words have a significant impact on the results produced by a parsing system, because tokenization is the starting point for all other annotations. Therefore, a preliminary check was carried out on the output from UDPipe, focusing first on tokenization issues and their correction. The UD format provides some recommendations to deal with both types of tokenization errors;[16] those same principles were also adopted in VALICO-UD. As reported in Example 3, wrongly merged words were split, adding in the MISC column of the first word involved (`nessuno` in the example) the `SpaceAfter=No` attribute accompanied by `CorrectSpaceAfter=Yes`.

(3)

sent_id = 36-02_es-3
text = Nel parco non c'era **nessunosolo** io
[*In the park there was **no one, only** me.*]

1-2	Nel	_	_	_	_	_	_	_
1	In	in	ADP	E	_	3	case	_
2	il	il	DET	RD	[...]	3	det	_
3	parco	parco	NOUN	S	[...]	6	obl	_
4	non	non	ADV	BN	_	6	advmod	_
5	c'	ci	PRON	PC	[...]	6	expl	SpaceAfter=No
6	era	essere	VERB	V	[...]	0	root	_
7	nessuno	nessuno	PRON	PI	[...]	6	nsubj	**SpaceAfter=No|CorrectSpaceAfter=Yes**
8	solo	solo	ADV	B	_	9	advmod	_
9	io	io	PRON	PE	[...]	6	orphan	SpaceAfter=No
10	.	.	PUNCT	FS	_	6	punct	SpacesAfter=\n

Conversely, in case of wrongly split words, as shown in Example 4, we promoted the first part of the word (`co`, word line 10) to syntactic head, which thus bears the lemma, PoS tag and morphology of the entire word; the remaining token(s) (`si`, word

16 Available here: `https://universaldependencies.org/u/overview/typos.html`.

line 11) were attached to the head with the `goeswith` relation and PoS-tagged with the generic `X` tag, thus leaving lemma and morphology unspecified.

(4)

sent_id = 26-10_de-1
text = Ma quando la ragazza ha visto il suo raggazzo **co si**, era desperata.
[*But when the girl saw her boyfriend **like this**, she was desperate.*]
[...]

5	ha	avere	AUX	VA	[...]	6	aux	_
6	visto	vedere	VERB	V	[...]	14	advcl	_
7	il	il	DET	RD	[...]	9	det	_
8	suo	suo	DET	AP	[...]	9	det:poss	_
9	raggazzo	raggazzo	NOUN	S	[...]	6	obj	_
10	**co**	**cosi**	**ADV**	**B**	_	**6**	**advmod**	_
11	**si**	_	**X**	_	_	**10**	**goeswith**	SpaceAfter=No
11	,	,	PUNCT	V	[...]	1	conj	_

[...]

Lemmatization – Apart from the tool's lemmatization errors—which usually involve open-class words, e.g. nouns, verbs, adjectives—in learner language corpora, lemmatization problems arise also because not all the words belong to the target language—i.e. Italian in VALICO-UD—nor to other known languages. In literature, different strategies are reported. They include not annotating lemmas (e.g. ESL treebank in UD), or rather annotating them using the lemma of the target form, in the presence of false friends or spelling errors (e.g. CFL treebank in UD).

In VALICO-UD we applied standard lemmatization rules for all tokens, also for tokens that are not reported in Italian dictionaries because they are borrowed from other languages or because they contain spelling or other errors. In this way, we maintain the form actually written by the learner. This allows us to treat uniformly all types of errors, also borderline ones.[17] Thus, in VALICO-UD, misspelled words have their own lemma, according to the PoS assigned, as shown in Examples 5 and 6.

(5) **LS**: Lui Era in **colera**, Lei era **terrozzata** [...]
Lemma: [...] colera [...] terrozzato [...]
PoS: [...] NOUN [...] ADJ [...]
TH: Lui era in **collera**, lei era **terrorizzata** [...]
Lemma: [...] collera [...] terrorizzato [...]
PoS: [...] NOUN [...] ADJ [...]
*He was **furious**, she was **terrified*** [...]

(6) **LS**: La **dona** ringraziava suo salvatore [...]
Lemma: [...] dona [...]
PoS: [...] NOUN [...]
TH: La **donna** ringraziava il suo salvatore [...]
Lemma: [...] donna [...]
PoS: [...] NOUN [...]
*The **woman** thanked her saviour* [...]

17 Borderline errors are those in which it is not trivial to assign an error type because more than one could fit; as an example, spelling errors resulting in actual words could be categorized also as replacement errors.

In Example 5, *colera* is a spelling error (intended signifier *collera*, 'anger') resulting in an existent word meaning 'cholera'. From the context it is clear that the learner meant to say *collera*, however, we lemmatized it as *colera*—keeping the spelling error—following the Italian lemmatization rule that applies to nouns. In turn, *terrozzata* (non-existent word likely used instead of *terrorizzata*, 'terrified') was lemmatized using the masculine singular form, as it is envisaged for adjectives. The word *dona* in Example 6—another spelling error resulting in a real word—was annotated considering its distributional and not the morphological marking, thereby it was treated as a noun (TH: *donna*) and not as a verb (third person singular indicative present of the verb *donare*, 'to give') as its form suggests. Thus, the lemma annotated is *dona* and not *donare* (nor its correct version *donna*).

When (non-)adapted loanwords occur, if they are in a plausible semantic context and they are borrowed from one of the learners' L1s, we lemmatized them following the lemma of the donor language, even retaining any spelling errors, as shown in Examples 7 and 8.

(7) **LS**: [...] ma Io può **derribar**le salvare a la donna.
Lemma: [...] derribar [...]
PoS: [...] VERB [...]
TH: [...] io potevo **batter**lo e salvare la donna.
Lemma: [...] battere [...]
PoS: [...] VERB [...]
[...] *but I can beat him and save the woman.*

(8) **LS**: [...] perchè non l'aveva fatto il **discaount** del 10% [...]
Lemma: [...] discaount [...]
PoS: [...] NOUN [...]
TH: [...] perché non le aveva fatto lo **sconto** del 10% [...]
Lemma: [...] sconto [...]
PoS: [...] NOUN [...]
[...] *because she had not given her a 10% discount* [...]

Due to our lemmatization choice, when an irregular verb is inflected using a wrong but existent inflectional variant, the corresponding lemma—following the standard lemmatization rules—remains the correct one. In Example 9, the irregular verb *volere* is conjugated by extending the stem of the first person (i.e. *vogli-*) to the third person singular (*vogli-e* instead of the correct *vuol-e*). Conversely, in Example 10, there is a spelling error which does not result in an existent inflectional variant of the correspondent verb (i.e. *partire*), hence the lemma reflects the learner's signifier.

(9) **LS**: [...] gente che soltanto **voglie** chiamare un pò l'atenzione.
Lemma: [...] volere [...]
PoS: [...] AUX [...]
TH: [...] gente che soltanto **vuole** richiamare un po' l'attenzione.
Lemma: [...] volere [...]
PoS: [...] AUX [...]
[...] *people who just want to call some attention to themselves.*

(10) **LS**: [...] la ragazza voleva **pertire** ma il ragazzo la teneva.
Lemma: [...] pertire [...]
PoS: [...] VERB [...]

TH: [...] la ragazza voleva **andare via** ma il ragazzo la teneva.
Lemma: [...] andare via [...]
PoS: [...] VERB ADV [...]
[...] *the girl wanted to leave but the boy was holding her*

In addition, morphologically speaking and not considering the cotext (i.e. the linguistic context in which the word occurs, as defined in (Lennon 1991)), *voglie* is a noun (meaning 'cravings'), but distributionally a modal verb, so we tagged it accordingly and lemmatized it as *volere*, since it is a case of overextension error, a good indicator of learning development.

Part of Speech tagging – Previous studies on the annotation of PoS tags in learner data have discussed the necessity of annotating more than one tag per each word in which discrepancies with the target language occur (Díaz-Negrillo et al. 2010; Ragheb and Dickinson 2012, 2014a). In particular, Díaz-Negrillo, *et al.* (2010) proposed the annotation of PoS tags taking into account three sources of evidence which can display discrepancies in learner language: distribution (i.e. the token position in the sentence), morphological marking (i.e. affixes attached to a word stem), and lexical stem lookup (i.e. lexical properties of a word). However, annotating separately these discrepancies would result in manageability and annotation-time issues.
For this reason, in presence of erroneous words, we annotate only one PoS per word/token. As mentioned above, two complementary criteria are followed, that is those of distributional and literal annotation.
We mainly apply literal annotation in all those cases in which following the grammar rules of Italian we coherently describe what the learner wrote. When non-words or existent words in inappropriate context appear, we apply distributional annotation, with only one exception. This is the case of words belonging to closed-class PoS with PoS inconsistent with the context, as exemplified in Example 11, where a preposition (*Durante*) is used instead of a multi-word expression functioning as an adverb (*Nel mentre*, meaning 'meanwhile').

(11) **LS**: **Durante** un ragazzo è passato.
Lemma: [...] durante [...]
PoS: [...] ADP [...]
TH: **Nel mentre** un ragazzo è passato..
Lemma: [...] nel mentre [...]
PoS: [...] ADP DET SCONJ [...]
***Meanwhile**, a boy passed by.*

(12) **LS**: [...] per la esatteza del **relato** devo descrivere quello che ho visto: Un uomo portava una donna sulle spale e questa quiedeva aiuto.
Lemma: [...] relato [...]
PoS: [...] NOUN [...]
TH: [...] per la esattezza del **racconto** devo descrivere quello che ho visto: un uomo portava una donna sulle spalle e questa chiedeva aiuto.
Lemma: [...] racconto [...]
PoS: [...] NOUN [...]
[...] *for the accuracy of the **story** I have to describe what I saw: a man was carrying a woman on his shoulders and she was asking for help.*

(13) **LS**: Sono **cerca** della città, ci sono due ragazzi e una ragazza.
Lemma: [...] cerca [...]
PoS: [...] ADP [...]
TH: Sono **vicino** alla città, ci sono due ragazzi e una ragazza.
Lemma: [...] vicino [...]
PoS: [...] ADP [...]
*They are **next** to the city, there are two boys and a girl.*

(14) **LS**: [...] ha pensato il fratello è stato un **rapinato** e ha salvato la ragazza.
Lemma: [...] rapinato [...]
PoS: [...] NOUN [...]
TH: [...] ha pensato che il fratello fosse un **sequestratore** e ha salvato la ragazza
Lemma: [...] rapinatore [...]
PoS: [...] NOUN [...]
*he thought his brother was a **kidnapper** and rescued the girl.*

Distributional annotation, in turn, is applied to words featuring spelling errors, adapted and non-adapted loanwords, and existent words (except for closed-class words) used in an unusual context for the original PoS. In particular, when dealing with spelling errors, even those resulting in real-word errors,[18] we let distributional properties prevail on lexical features, as shown in Example 6, in which *dona* is annotated as `NOUN` instead of `VERB`. When dealing with foreign adapted or non-adapted words, we annotate following the distributional annotation, even if these borrowed words exist in Italian with another PoS and/or meaning, as in Examples 12 and 13. In the former, *relato* is likely borrowed from Spanish with the meaning of 'story', and it is not the unusual Italian adjective meaning 'related',[19] thus we annotated it as `NOUN` and not `ADJ`. In the latter, the learner borrowed a Spanish preposition (*cerca de* meaning 'next to') which in Italian results in a verb (lemma *cercare*, meaning 'to search') followed by a preposition-article contraction (*della*, meaning 'to the'). Finally, other existent words (except for closed-class words), used in contexts that are unusual for the original PoS, are annotated distributionally, as in Example 14, in which the learner used a past participle (*rapinato*, meaning 'robbed', functioning also as adjective) in the distributional context of a noun.

Morphological features – Morphological information encoded in the FEATS column contains lexical (e.g. `Foreign`, which indicates that a word is a foreign word) and inflectional features (e.g. `Gender` or `Number`) of the word. These features, however, can be annotated only in conjunction with specific UPOS and based on the specific morphological traits of a given word, such as verb form and tense (in Italian verbs, for instance, gender is available only for past participles). As it can also be seen in the CoNLL-U examples 2–4, proper nouns, prepositions, conjunctions, punctuation and adverbs have no morphological features in Italian. These format specifications do have an impact on the treebank data at hand, and its relative annotation. Let us consider the word *contra* in Example 15. *Contra* is a Latin preposition meaning 'against' which is used in French, Spanish and German maintaining meaning and function. In Italian it corresponds to

18 Real-word errors are spelling errors resulting in actual words, and its identification can only be made by looking at the context (Dickinson, Brew, and Meurers 2013, p. 37), e.g. *I will have the chocolate **desert** instead of **dessert***.

19 This meaning is probably unknown to most native speakers together with the Latin loanword *de relato*; the phrase, used in legal settings to refer to an indirect testimony, might be reported in a legal dictionary but not in our reference dictionary.

contro. Naively, one could think that *contra* displays misleading morphological features with respect to the UPOS to which is assigned—i.e. morphological marking of feminine singular but PoS `ADP`. In similar cases, we do not mark these features for two reasons. First, it is not allowed in UD formalism. Second, considering the final '-a' in *contra* as a morphological marking of feminine singular is highly debatable.

(15) **LS**: Stà furiosa **contra** il ragazzo che non comprende quello che pasò.
Lemma: [...] contra [...]
PoS: [...] ADP [...]
TH: È furiosa **contro** il ragazzo che non comprende quello che è successo.
Lemma: [...] contro [...]
PoS: [...] ADP [...]
*She is furious **with** the boy who does not understand what happened.*

In cases in which the second reason does not stand, we avoided the constraint posed by the UD framework adding these features in the MISC column of the CoNLL-U file, when their annotation is of interest. This is the case of foreign words, for example. To avoid format issues arising from the presence of foreign words with not allowed UPOS, we annotated the `Foreign` feature in the MISC column. Hence, differing from the other UD treebanks in which the foreign feature is annotated in the FEATS column, in VALICO-UD this information is always annotated in the MISC column.[20]

In learner texts, also inflectional features are really useful because their annotation can help in the detection of discrepancies with the target language (e.g. agreement errors). For this reason, contrarily to the other Italian treebanks in which gender and number information of elided pronouns and determiners (e.g. *l'uomo*) is never annotated, in VALICO-UD we added this information to recover possible discrepancies.
If the referent cannot be traced from the cotext, and gender and number of the determiner or pronoun cannot be derived from its form, we do not mark this information, as in Example 16, in which we annotate the person (3rd) of the pronoun, but not gender and number of *gle* (orthographically correct *glie*) because the referent cannot be identified with certainty.

(16) **LS**: Lui era un ragazzo buono e ardito: si è alzato, e li è seguito; quando se li ruscito, ha detto: "Lascia**gle**la"
Lemma: [...] lasciare gle lo [...]
PoS: [...] VERB PRON PRON [...]
TH: Lui era un ragazzo buono e ardito: si è alzato e li ha seguiti; quando li ha raggiunti, ha detto: "Lasciala".
Lemma: [...] lasciare lo [...]
PoS: [...] VERB PRON PRON [...]
*He was a good and bold boy: he got up and followed them; when he reached them, he said: "leave **her**".*

If the syntactic relation between the words is clear, we annotate distributionally also morphological features. In Example 17, the adjective *forte* (*strong*, singular and gender invariant in Italian), modifies the noun *parole* (*words*, feminine plural). We treated this adjective as a case of over-extension of *-e*, thereby we added the morphological features

20 An anonymous reviewer criticized this decision as it does not fully comply with UD specifications. However, we decided to follow a distributional approach also in this case, to be consistent with other similar contexts.

for feminine (`Gender=Fem`) and plural (`Number=Plur`). We made this decision also because in the text there are not agreement errors, so we thought that it was likely used as feminine plural (producing the correct agreement) and not masculine singular.

(17) **LS**: Avevo sentito delle **parole forte**, una donna sta gridando et un uomo la portava con lui brutalmente.
Lemma: [...] parola forte [...]
PoS: [...] NOUN ADJ [...]
TH: Avevo sentito delle **urla**, una donna stava gridando e un uomo la portava con sé brutalmente.
Lemma: [...] urla [...]
PoS: [...] NOUN [...]
*I heard **shouting**, a woman was screaming and a man was brutally carrying her.*

Other learner-related phenomena were also marked in the MISC column. One of these is the presence of evaluative suffixes, which were marked with the attribute `EvalMorph`. Currently, the only value used is `Dim` and indicates the presence of diminutives in both LSs and THs. Thanks to this information, it is possible to retrieve examples such as the one reported in 18 in which the learner used *canino* instead of *cagnolino*, producing a word having a different meaning ('canine tooth' instead of 'doggy').[21]

(18)

sent_id = 34-07_en-3
text = [...] tale come camminare il canino al parco.
[*such as walking the **doggy** in the park.*]

9	tale	tale	ADJ	A	[...]	11	mark	_
10	come	come	SCONJ	CS	_	9	fixed	_
11	camminare	camminare	VERB	V	[...]	4	acl	_
12	il	il	DET	RD	[...]	13	det	_
13	**canino**	cane	NOUN	S	[...]	11	obj	**EvalMorph=Dim**
14-15	al	_	_	_	_	_	_	_
14	a	a	ADP	E	_	16	case	_
15	il	il	DET	RD	[...]	16	det	_
16	parco	parco	NOUN	S	[...]	11	obl	SpaceAfter=No
17	.	.	PUNCT	FS	_	2	punct	SpacesAfter=\n

sent_id = 34-07_en-3
TH-text = [...] come passeggiare con il cagnolino al parco.
[*such as walking the **doggy** in the park.*]

10	come	come	ADV	B	_	11	advmod	_
11	passeggiare	passeggiare	VERB	V [...]	4	acl	_	
12	con	con	ADP	E	_	14	case	_
13	il	il	DET	RD	[...]	14	det	_
14	cagnolino	cane	NOUN	S	[...]	11	obl	**EvalMorph=Dim**
15-16	al	_	_	_	_	_	_	_
15	a	a	ADP	E	_	17	case	_
16	il	il	DET	RD	[...]	17	det	_
17	parco	parco	NOUN	S	[...]	11	obl	SpaceAfter=No
18	.	.	PUNCT	FS	_	2	punct	SpacesAfter=\n

21 Note that in Italian masculine diminutives are formed adding *-ino* to the word stem. *Cane* makes an exception.

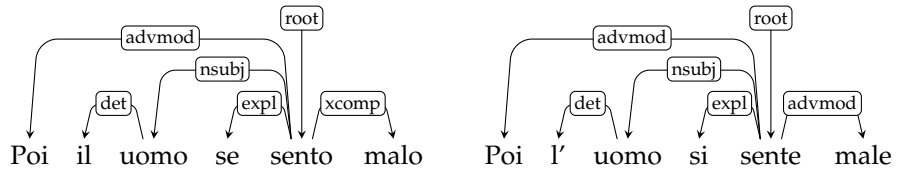
In addition, we marked also the presence of multi-word expressions that are not usually treated as such in the other available UD treebanks. In this way, we annotated trees following the same annotation rules adopted in the other treebanks, but we added in the MISC column the attribute `LOC` followed by the lowercase letter of the UPOS indicating the function of the multi-word expression. In particular, we used `adv` for adverbial and `adj` for adjectival. Other multi-word expressions, instead, are annotated using the `fixed` DEPREL as it is the case in the other UD treebanks (e.g. *tale come* in Example 18 is further explained in the next paragraph, *come se*, *fino a*). Thanks to this annotation, it is possible to retrieve occurrences of creative multi-word expressions which would be inevitably missed without this annotation, as the one reported in Example 19 in which the learner invented a new multi-word expression (*al invece*) functioning as an adverb.

(19) **LS**: Ma lei **al invece** s'era arrarbi contro l'uomo carino dicendo che lui, aveva fatto male al suo amore.
TH: Ma lei **invece** s'era arrabbiata con l'uomo carino dicendo che lui aveva fatto male al suo amore.
***On the contrary**, she got angry with the nice man saying that he had hurt her love.*

Syntactic annotation – Following the same two principles (i.e. distributional and literal annotation) applied for dealing with the annotation of non-canonical forms at the different levels of annotation discussed above, here we describe how we dealt with erroneous syntactic structures. Following the other projects that syntactically annotated learner language as it is (Dickinson and Ragheb 2009; Berzak et al. 2016; Lee, Leung, and Li 2017), we annotated dependencies taking morphological and distributional evidences into account, rather than the speaker's intended meaning (Ragheb and Dickinson 2014b, pp. 137–138). In this way, we reduce the subjectivity of annotators, since they do not have to interpret the learners' intended meaning, but must rely on formal grammatical and distributional evidences.

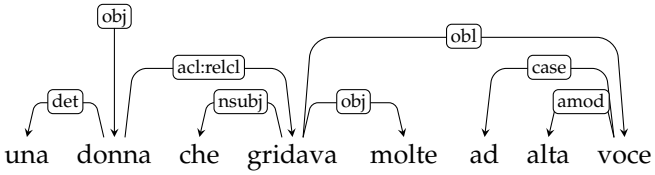
For example, in the LS in 20, the word *malo* (which does not exist in Italian) is annotated as an adjective because of the morphological features of masculine singular, while in the TH it is substituted by the adverb *male*. It follows that in the LS tree the relation connecting *malo* to its governor *sento* is `xcomp`, a relation also used in constructions that are known as *secondary predicates* or *predicatives*.

(20)



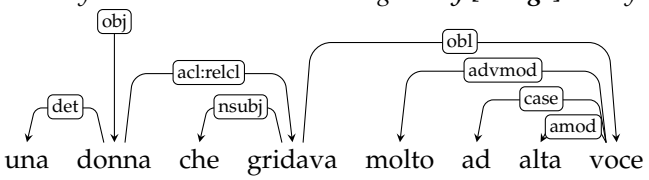
LS: Poi il uomo se **sento malo**. **TH**: Poi l'uomo si **sente male**.
- *Then the man **feels ill**.*

A similar case is the one reported in Example 21, where *molte* is annotated as indefinite pronoun rather than as adverb (like in the TH), due to the ending *-e*, normally used for feminine plural. As a result, the LS tree is different from the TH tree not only for the dependency relations (`obj` → `advmod`), but also for the nodes' governors (*gridava* → *voce*).



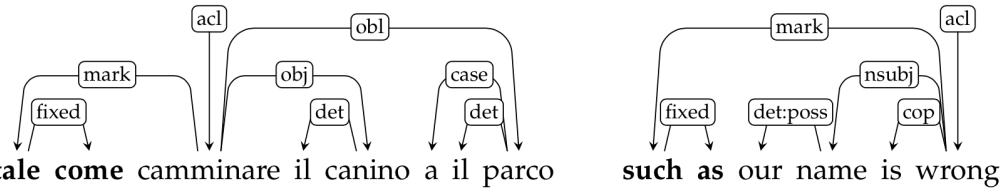
(21)

LS: All'improviso ha sentito una donna che gridava **molte** ad alta voce.
*Suddenly he heard a woman shouting **many [things]** loudly.*



TH: All'improvviso ha sentito una donna che gridava **molto** ad alta voce.
*Suddenly he heard a woman shouting **very** loudly.*

When annotating LSs as they are following the L2 grammar, problems arise when learners' structures do not correspond to the L2 grammar. Let us consider the example reported in 18, in which the subordinate clause is a word-for-word translation of the English structure (reported in the figure between square brackets), a syntactic calque. In this case, we annotated *tale come* as a fixed expression with the function of conjunction, although it does not exist in Italian, and *il canino* as direct object of *camminare* even though this verb is intransitive. In this way, the resulting annotation is not only comparable to other Italian treebanks, but also to English sentences, highlighting the similarities; in Example 22, we show two comparable structures retrieved from VALICO-UD and the English Web Treebank (EWT) (Silveira et al. 2014): they are both (non-)finite clauses modifying a nominal (`acl`), introduced by *such as* (literally, *tale come*).

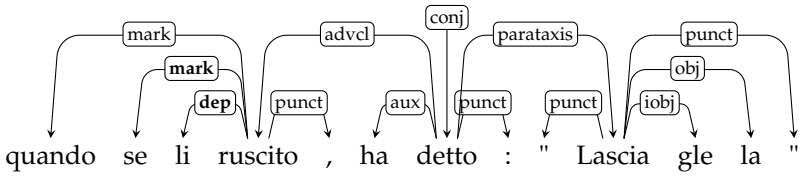


(22) **LS:** Ho provato molte strategie per attrarre le ragazze tale come camminare il canino al parco.
I have tried many strategies to attract girls such as walking the doggy in the park.
EWT:[22] I looked at the UEComm Master and had some comments–such as our name is wrong, [...]

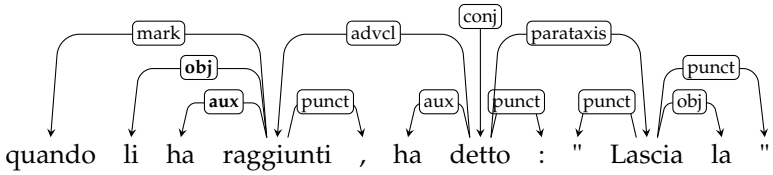
22 Found looking for *such as* here: `http://match.grew.fr/?corpus=UD_English-EWT@2.8`.

As *extrema ratio*, when learners' structures do not coincide with the L2 grammar, and it is not possible to infer the syntactic function of one or more words, we resorted to the general dependency *dep*, as shown in Example 23.

(23)



LS: Lui era un ragazzo buono e ardito: si è alzato, e li è seguito; quando **se li ruscito**, ha detto: "Lascia**glela**"
*He was a good and bold guy: he got up and followed them; when **if he did**, he said: "Leave **it to her**".*



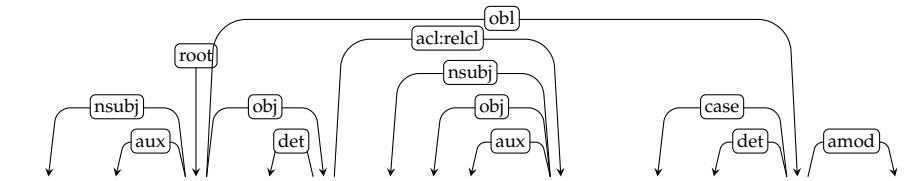
TH: Lui era un ragazzo buono e ardito: si è alzato e li ha seguiti; quando li ha raggiunti, ha detto: "Lascia**la**".
*He was a good and bold guy: he got up and followed them; when he reached them, he said: "Leave **her**".*

We annotated the subordinate clause as if starting with two conjunctions—even though *se* could also be annotated as a pronoun—and then, since it was not possible to understand the syntactic function of *li*, we used the general dependency *dep*. Perhaps, the verb in the subordinate adverbial clause is used as if it was a pronominal verb, thus in this case *se*, together with *li*, should have been annotated as a pronoun and with dependency relation `expl`. Nevertheless, since this would be a highly subjective choice, we labeled them with a general dependency relation.

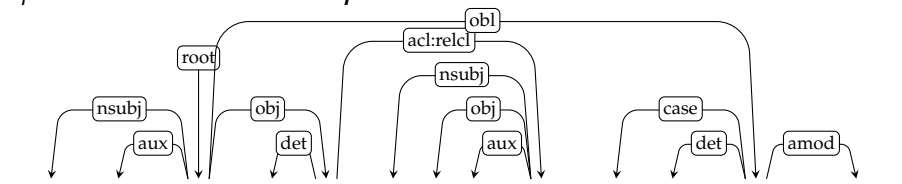
As it might be easily predictable, semantics errors do not pose problems in syntactically annotating VALICO-UD. In Example 23, indeed, we annotated *lasciaglela* literally, following the L2 grammar, giving the sentence the meaning of *leave her/it to her/him/them*, thereby ignoring the intended meaning of the learner (i.e. *leave her*). Rendering the meaning is in turn addressed in the TH, in which *Lasciaglela* (orthographically correct *Lasciagliela*) is corrected in *Lasciala*, deleting the learner's indirect object (*gle*), to render syntactically the meaning of the sentence.

Another example that perhaps better illustrates the concept is the one reported in Example 24. Any Italian speaker reading this sentence can syntactically annotate it, even though the sentence makes no sense at all. The only ambiguity might be about the governor of *sguardo*, which could be also *conservata*, although we believe that human annotators would perceive the semantic affinity of *guardato* ('looked') and *sguardo* ('look') and resolve the ambiguity. This is even more plausible because it is unlikely that someone (i.e. the man) can *conservare* ('keep') someone else (i.e. Sophia) with a medium look.

(24)

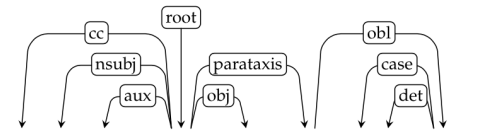


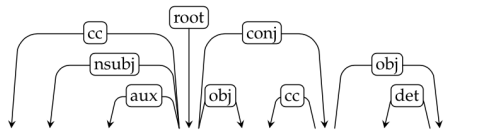
LS: Sophia ha guardato l'uomo che la ha **conservata** con uno sguardo **medio**.
*Sophia looked at the man who **kept** her with a **mean** look.*



TH: Sophia ha guardato l'uomo che la ha **salvata** con uno sguardo **cattivo**.
*Sophia looked at the man who **rescued** her with a **mean** look.*

Another challenge—although less problematic that the presence of words/structures which do not belong to any known language, as the case of *se li riuscito*—is how to syntactically annotate sentences in which foreign words occur. We literally annotate loanwords belonging to one of the four considered learners' L1s (i.e. DE, EN, ES, FR) when they are in a plausible syntactic and semantic context, as shown in Example 25, in which the verb *derribar* and the clitic pronoun *le*, meaning 'to take him down', is borrowed from Spanish and inserted in a plausible semantic and syntactic context (with *le* referring to *uomo*). However, since *le* is also an Italian pronoun, we decided to annotate it following the L2 grammar and not the language from which is borrowed, thus avoiding the creation of a new rule which would annotate *le* as a direct object referring to a masculine singular antecedent. Since *le* in Italian can be a pronominal direct object referring to a feminine plural antecedent or a pronominal indirect object referring to a feminine singular antecedent, we decided to annotate it as the former, thereby maintaining the relation but losing the morphology information (prioritizing syntax over morphological features).





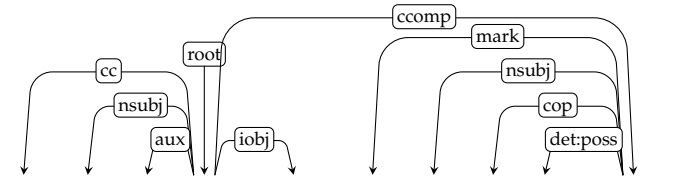
(25) **LS**: Il uomo era alto, forte e molto musculuso, ma Io può **derribarle** salvare a la donna.
TH: L'uomo era alto, forte e molto muscoloso, ma io potevo **batterlo** e salvare la donna.
*The man was tall, strong and very muscular, but I could **beat him** and save the woman.*

Thanks to our annotation choices, comparing the trees in Example 25 with the correspondent TH, we can obtain the interpretation of the learner' errors. In Example 25, the

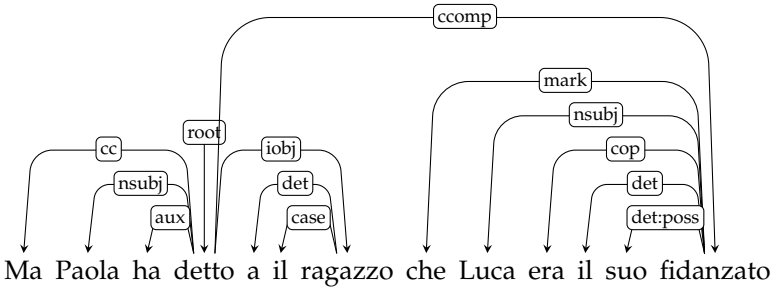
syntactical changes consist in the insertion of a coordinate conjunct (*e salvare la donna*) instead of the paratactical structure (juxtaposition of the two clauses without conjunction), and the deletion of the preposition in the direct object (*salvare **a** la donna*). The morphological changes concern *può*, which changes from third person to first person, and *le*, changing from the feminine plural to the masculine singular.

Following the other principle, the distributional annotation of LSs, we considered the verb as a guide for the annotation. In Example 26, since the verb *dire* 'to say' has a valency of three, we saturated its valency annotating *ragazzo* as indirect object and not as a direct object, which could be the case if we do not consider neither semantics not the cotext, annotating it as if the sentence ends at *ragazzo*.

(26)



LS: Ma Paola **ha detto ragazzo** che Luca era suo fidanzato
*But Paola **told the boy** that Luca was her boyfriend*



TH: Ma Paola **ha detto a il ragazzo** che Luca era il suo fidanzato
*But Paola **told to the boy** that Luca was her boyfriend*

It is worth noticing that having a parallel treebank is useful not only for syntactic information, but for the morphological one. In fact, some decisions made in the annotation, such as the choice of maintaining learners' signifiers when lemmatizing, can be useful only if compared with the TH. Since the aim of lemmatization is to retrieve all the inflections of a word, a word which is lemmatized maintaining spelling errors or lexical errors could be seen as a problem if we aim at retrieving all the contexts in which a learner wrote that word and its inflection. However, having a parallel treebank allow us to be able to carry out these queries without forcing the annotation of the interlanguage using the intended form.

Inter annotator agreement – Once the syntactic annotation scheme was defined, with the aim of assessing the annotation quality of the treebank as well as the quality of the annotation guidelines and their applicability, two independent annotators annotated independently a 200-sentence sample of VALICO-UD (100 LSs and the 100 corresponding THs) as described in (Di Nuovo et al. 2019).

The inter annotator agreement was computed considering two measures in particular: UAS (Unlabeled Attachment Score) and LAS (Labeled Attachment Score) for the assignment of both parent node and dependency relation, and the Cohen's kappa

coefficient (Cohen 1960) for dependency relations only (similarly to Lynn (2016)). UAS and LAS were computed with the script provided in the second CoNLL shared task on multilingual parsing (Zeman et al. 2018).

The results are reported in Table 4, and though showing slightly higher results for the TH set, overall they are very close across the sets. Especially as regards the LS section, this is evidence of guidelines clarity and of annotators' consistency, even when dealing with non-canonical syntactic structures.

Table 4
Agreement results on the sample set of both LSs and THs.

Set	UAS	LAS	*kappa*
LS	92.11%	88.63%	0.8988
TH	92.47%	88.88%	0.9068

Parser evaluation – In order to quantify the manual effort required by human annotators to correct the output obtained with the UDPIPE model trained on PoSTWITA and ISDT (see Section 4), in Table 5 we report the average F1 achieved in both LSs and THs with respect to parts of speech (UPoS) and UAS and LAS. F1 is obtained using the official evaluation script provided for the second CoNLL shared task.

Table 5
Average F1 on LSs and THs of automatic annotation of parts of speech (UPoS) and syntactic parsing (UAS and LAS).

Set	UPoS	UAS	LAS
LS F1	94.22%	84.75%	79.61%
TH F1	96.19%	88.34%	84.69%

To summarize, in this section we described the challenges faced in treebanking VALICO-UD, motivating the choices made and highlighting the benefits of annotating linguistic information. In the next section, we present a preliminary exploration of the treebank making use of quantitative measures.

5. Treebank exploration: preliminary evaluation via quantitative measures

In this section we explore the treebank using three quantitative measures for assessing data quality and for better understanding the role that this resource can play in the future in tasks lying at the intersection of parsing and learner corpus studies. Firstly, we evaluate the parser performance on out-of-domain texts (i.e. using the gold standard, LSs and THs) using LAS (Labelled Attachment Score). Then, we evaluate it on LSs and THs separately. Secondly, we validate the hypothesis that a distance exists between two groups of learners at different stage of learning, by applying a string-based measure, called TER (i.e. Translation Error Rate) (Snover et al. 2006), between LSs and THs texts. Finally, we apply also a new tree-similarity measure (i.e. UDAPI's F1 LAS) (Popel, Zabokrtský, and Vojtek 2017) between LS and TH trees to further confirm this validation. These three evaluations and related measures are described in the next three subsections.

5.1 Evaluation based on parsing performance

The first evaluation exercise is aimed to assess how hard parsing VALICO-UD—and learner language in general—can be.

As reported in Section 4, the resource has been built by feeding a subcorpus of VALICO to a UDPipe model trained on ISDT and PoSTWITA UD treebanks.

Part of the annotated output, then, has been manually corrected to obtain the core gold section of the resource, while the rest is released as silver standard. The availability of the manually-checked core section of VALICO-UD (detailed in Section 3) has been crucial for the evaluation exercises reported in this section: we used it as gold standard against which to compare the parser output produced on the same data, and subsequently to evaluate the quality of the automatic result.

The first evaluation we carried out is on the totality of the parallel treebank (LSs and THs together) and it allows us to measure the parsing performance on an out-of-domain test set. As usual, when training a model on a text domain and testing it on another domain, a loss in performance is expected. As a second step, we evaluated the parser results separately for LSs and THs to quantitatively measure how much interlanguage affects the performance. On the one hand, in the evaluation based on THs, even though they are written by an Italian native speaker, we expect a loss in performance, since THs are out of domain with respect to ISDT and PoSTWITA. On the other hand, we expect a bigger loss in performance when the same model is tested on LSs, because not only LSs are out of domain, but also they may contain errors introduced by learners at all levels of linguistic analysis.

To be comparable with state-of-the-art parsers, we evaluated two UDPIPE models trained separately on ISDT and PoSTWITA. As evaluation metric, we observed the F1 on LAS computed with the official evaluation script released for the CoNLL 2018 Shared Task (Zeman et al. 2018). Results are reported in Table 6.

Table 6
Parser performance LAS and UAS on VALICO-UD compared with State of the Art (SotA) in-domain LAS results.

Data	Trained on ISDT	Trained on PoSTWITA	SotA on ISDT	SotA on PoSTWITA
	LAS ⁞ UAS	LAS ⁞ UAS	LAS	LAS
LSs + THs	86.79 ⁞ 89.95	84.69 ⁞ 87.45	92.00	79.39
LSs	85.34 ⁞ 89.12	83.93 ⁞ 87.30	–	–
THs	88.25 ⁞ 90.77	85.46 ⁞ 87.60	–	–

The model trained on ISDT achieved $F1 = 86.79$, the model trained on PoSTWITA $F1 = 84.70$. As expected these results are lower than the state of the art for in-domain standard Italian parsing ($F1 = 92.00$), that is the result achieved by the best performing parser, HIT-SCIR (Che et al. 2018) trained and tested on ISDT, at the CoNLL 2018 Shared Task. However, both are higher than the best result achieved by the same parser trained and tested on PoSTWITA ($F1 = 79.39$). Separately on LSs and THs, as expected, both models better performed on THs. However, the model trained on PoSTWITA has a smaller gap between LSs and THs than ISDT (i.e. 1.53 points in PoSTWITA and 2.91 in ISDT). For both models LSs are not significantly affecting out-of-domain performance. Both models better performed on THs, achieving a lower score with respect to the cited state of the art on standard Italian (i.e. ISDT), but, on average, 6 points higher than the

state of the art achieved on social media (i.e. PoSTWITA). This could be explained by different factors: more noise in PoSTWITA data (e.g. wrongly split words and wrongly merge words), and segmentation issues attaining tweets (usually formed by more than one sentence, but analysed as one).

In the next sections we provide some evaluation exercises based on two metrics, the first one based on the difference and the other on the similarity occurring among LS and TH strings and trees, respectively. Both are correlated with learners' year of study of Italian, grouping them in two classes, i.e. initial and advanced learners.

5.2 Evaluation based on string distance between LSs and THs

To measure the distance between LSs and THs at string level we exploited a tool called TER COM (Snover et al. 2006). TER COM is a software, available in Java and Perl, that computes a distance metric called TER (Translation Error Rate) that is used in machine translation to measure the number of edits required to change a system output (i.e. LSs) into one of the references (i.e. THs). Its value goes from 0, meaning that the two compared sentences are the same, to 1, meaning that the two compared sentences are completely different. In brief, the lower the score, the better. In Example 27, we show a LS-TH sentence pair with a TER value of 0.375.

(27) **LS:** Ieri al parco è successuto qualcosa stana.
TH: Ieri al parco è successo qualcosa di strano.
Yesterday in the park something strange happened.

Once we computed this metric on the LS-TH text pairs, we compared the results obtained for the two groups of learners.

As introduced above, the hypothesis we want to test is that the difference between LSs and THs should be larger in texts belonging to the group of initial learners, and smaller in texts produced by more advanced learners.

The data we used are the 402 texts (i.e. 201 LS texts and their 201 TH texts) composing the silver standard of the treebank. The available texts and their metadata about learners' L1 and year of study of Italian are shown in Table 7. The texts used for the data exploration are in bold.

Table 7
Texts and metadata of the silver standard. Texts selected for the exploration are in bold. The question mark indicates that the year of study is *not known*.

Learners' L1	**# texts per year of study**					
	1	**2**	**3**	**4**	**>4**	**?**
DE	**8**	**2**	10	11	**14**	4
EN	**7**	**21**	3	**13**	**3**	4
ES	22	3	0	2	0	23
FR	**7**	**13**	5	4	**20**	2

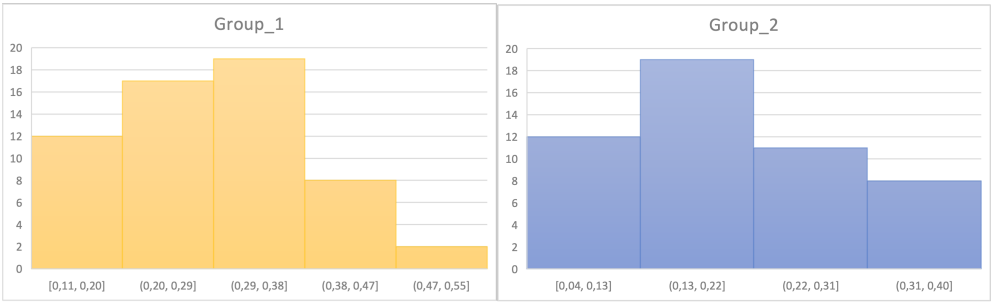
In particular, we selected 58 texts for the group of initial learners—i.e. all texts produced by DE, EN and FR learners at their first or second year of study of Italian—, and 50 texts for the group of advanced learners—i.e. all texts produced by DE, EN and

FR learners being at least at their fourth year of study of Italian—and their corresponding THs. Hence, we obtained two groups of texts that we used in order to verify our hypothesis.

For the group of initial learners, comparing LSs and THs, the mean obtained is 0.29 (with a standard deviation $\sigma = 0.10$). For the group of advanced learners the mean obtained is 0.20 ($\sigma = 0.10$).

These results have confirmed our initial hypothesis that the difference between LSs and THs is larger for for initial learners than for advance learners.

Then we wanted to test if the difference given by the TER values of the two populations is statistically significant. To assess if the unpaired t-test is reasonable for our data, we visualized in two histograms the TER values obtained on each text of the two groups, as shown in Figure 2. In the histogram, the y axis indicates the number of texts (frequency) having a TER value included in the range indicated in the x axis. For example, in the group of initial learners (Group 1), there are 12 texts with TER value from 0.11 to 0.20.



(a) Group 1: initial learners.

(b) Group 2: advanced learners.

Figure 2
Histograms of the TER values obtained for each text of the two groups, initial (Group 1) and advanced (Group 2) learners of Italian. The y axis indicates the frequency of texts having a TER value included in the range indicated in the x axis.

Since the data, as shown by the histograms, have more or less a Gaussian distribution, with no outliers, and the standard deviations are the same in the two groups, the idea of carrying out an unpaired t-test with equal variances seems reasonable. The obtained two-tailed P value is less than 0.0001, which is considered to be extremely statistically significant.

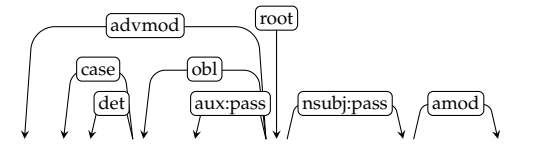
5.3 Evaluation based on tree distance between LSs and THs

To assess if the two groups are different also when comparing the syntactic trees, we exploited UDAPI (Popel, Zabokrtský, and Vojtek 2017), an API and framework, available for Python, Perl and Java, for processing UD data that can be used for a wide range of use cases (e.g. tree viewer, format conversion, querying, automatic parsing, evaluation).[23] In particular, we used the function *F1* of the block *eval* that is used for computing the similarity between two different UD trees using the F1 metric. Its value

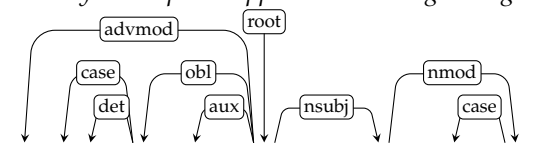
23 Documentation available here: `https://udapi.readthedocs.io/en/latest/udapi.block.eval.html\#module-udapi.block.eval.f1`.

goes from 0 to 100, with 100 meaning that the two compared sentences are the same. Contrarily to TER, the higher the score, the better.
Note that in Section 5.1 LAS is also expressed in F1, but the two compared sentences are identical at string level. In contrast, in this section we are considering LSs as system trees and THs as gold trees, hence the two sentences can be different and are automatically aligned.
In Example 28, we show the trees of the two sentences reported in Example 27. Note that we are using the automatically parsed trees. The obtained F1 LAS is 73.68.

(28)



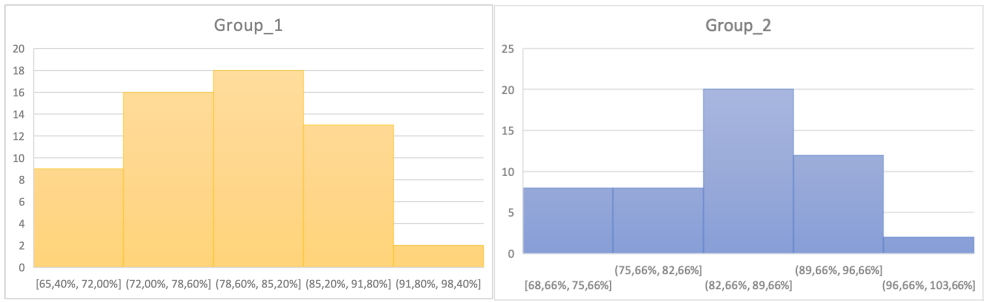
LS: Ieri al parco è successuto qualcosa stana
Yesterday in the park happened something strange



TH: Ieri al parco è successo qualcosa di strano
Yesterday in the park happened something strange

For the group of initial learners the mean obtained is 80 ($\sigma = 7$). For the group of advanced learners the mean obtained is 86 ($\sigma = 7$).

Again, to test the difference between the two groups statistically, we visualized in two histograms the F1 values obtained on each text of the two groups, as shown in Figure 3.



(a) Group 1: initial learners. (b) Group 2: advanced learners.

Figure 3
Histograms of the F1 values obtained for each text of the two groups, initial (Group 1) and advanced (Group 2) learners of Italian. The y axis indicates the frequency of texts having a F1 LAS value included in the range indicated in the x axis.

As happened when comparing the obtained TER values, also in this case the idea of carrying out an unpaired t-test with equal variances seems reasonable. Again, the

obtained two-tailed P value is less than 0.0001, which is considered to be extremely statistically significant.

Although the difference between the two groups is not large at both string and tree level, it is statistically significant for distinguishing two different populations of learners (i.e. initial and advanced learners). This is explained by the TH-writing design principles. Indeed, one of the principles is to normalize choosing the nearest correct version of what the learner wrote. For this reason, the distance between LS and TH strings is not so big as one might think. As far as the distance between the two LS and TH trees is concerned, this low difference can be explained by the consistent output provided by the parser in both sets. In fact, looking at performance on dependency relations individually, it performs comparably in the two sets with few exceptions.

The results obtained using TER COM and UDAPI's eval.F1 are promising for two reasons. First, they quantitatively confirm that two different population of learners exist (initial and advanced learners). Second, they indirectly assess and validate the consistency of the normalization of LSs (i.e. THs writing).

6. Conclusion

In this article, we have introduced the new learner Italian treebank VALICO-UD, which has been recently made available in the UD repository. We reported the challenges addressed in annotating it at different linguistic levels of analysis, showing that the UD framework proved to be flexible enough to be applied to interlanguage. The parsing evaluation applied on the resource confirms the quality of the resource, which has been in part released as a manually-checked gold standard and in part as silver standard.

The parallel nature of the treebank, where each Learner Sentence is indeed paired with a corresponding Target Hypothesis, moreover allowed us to statistically confirm the hypothesis that at least two distinct populations of learners exist, initial and advanced learners, and that they can be identified according to two different quantitative metrics.

The extension of the resource and the inclusion of a further level of annotation devoted to the representation of learners' errors will be useful for improving the validity of the results achieved until now.

Acknowledgments

We want to thank Dr. Martin Popel for helping us use UDAPI, and for being always ready and quick to fix bugs.

References

Aijmer, Karin. 2002. Modality in advanced Swedish learners' written interlanguage. *Computer learner corpora, second language acquisition and foreign language teaching*, pages 55–76.

Alfieri, Linda and Fabio Tamburini. 2016. (Almost) Automatic Conversion of the Venice Italian Treebank into the Merged Italian Dependency Treebank Format. In *Proceedings of the Third Italian Conference on Computational Linguistics (CLiC-IT)*, pages 19–23, Naples, Italy, December.

Astaneh, Sadegh and Francesca Frontini. 2009. L'adattamento di un parser di italiano L1: problemi e prospettive. *Corpora di italiano L2: tecnologie, metodi, spunti teorici*, 2:199–216.

Berzak, Yevgeni, Jessica Kenney, Carolyn Spadine, Jing Xian Wang, Lucia Lam, Keiko Sophie Mori, Sebastian Garza, and Boris Katz. 2016. Universal Dependencies for Learner English. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics*, pages 737–746, Berlin, Germany, August.

Blanchard, Daniel, Joel Tetreault, Derrick Higgins, Aoife Cahill, and Martin Chodorow. 2013. TOEFL11: A corpus of non-native English. *ETS Research Report Series*, 2013(2):i–15.

Boyd, Adriane, Jirka Hana, Lionel Nicolas, Detmar Meurers, Katrin Wisniewski, Andrea Abel, Karin Schöne, Barbora Stindlová, and Chiara Vettori. 2014. The MERLIN corpus: Learner Language and the CEFR. In *Conference on Language Resources and Evaluation*, pages 1281–1288, Reykjavik, Iceland, May.

Bryant, Christopher, Mariano Felice, Øistein E. Andersen, and Ted Briscoe. 2019. The BEA-2019 shared task on grammatical error correction. In *Proceedings of the Fourteenth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 52–75, Florence, Italy, August. Association for Computational Linguistics.

Che, Wanxiang, Yijia Liu, Yuxuan Wang, Bo Zheng, and Ting Liu. 2018. Towards better UD parsing: Deep contextualized word embeddings, ensemble, and treebank concatenation. In *Proceedings of the CoNLL 2018 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies*, pages 55–64, Brussels, Belgium, October. Association for Computational Linguistics.

Cignarella, Alessandra Teresa, Valerio Basile, Manuela Sanguinetti, Cristina Bosco, Paolo Rosso, and Farah Benamara. 2020. Multilingual irony detection with dependency syntax and neural models. *arXiv preprint arXiv:2011.05706*.

Cignarella, Alessandra Teresa, Cristina Bosco, and Paolo Rosso. 2019. Presenting TWITTIRÒ-UD: An Italian Twitter treebank in Universal Dependencies. In *Proceedings of the Fifth International Conference on Dependency Linguistics (Depling, SyntaxFest 2019)*, pages 190–197, Paris, France, August. Association for Computational Linguistics.

Cohen, Jacob. 1960. A Coefficient of Agreement for Nominal Scales. *Educational and Psychological Measurement*, 20(1):37–46.

Corder, Stephen Pit. 1967. The significance of learner's errors. *International Review of Applied Linguistics*, 5(4):161–170.

Corder, Stephen Pit. 1971. Idiosyncratic dialects and error analysis. *International Review of Applied Linguistics*, 9(2):147–160.

Corino, Elisa and Carla Marello. 2009. Elicitare scritti a partire da storie disegnate: il corpus di apprendenti VALICO. In Cecilia Andorno and Stefano Rastelli, editors, *Corpora di italiano L2: tecnologie, metodi, spunti teorici*. Guerra.

Corino, Elisa and Carla Marello. 2017. *Italiano di stranieri. I corpora VALICO e VINCA*. Guerra.

Corino, Elisa and Claudio Russo. 2016. Parsing di Corpora di Apprendenti di Italiano: un Primo Studio su VALICO. In *Proceedings of the 3rd Italian Conference on Computational Linguistics, CLiC-it 2016*, pages 105–110, Naples, Italy, December.

Dahlmeier, Daniel, Hwee Tou Ng, and Siew Mei Wu. 2013. Building a large annotated corpus of learner English: The NUS corpus of learner English. In *Proceedings of the Eighth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 22–31, Atlanta, Georgia, June.

Davidson, Sam, Aaron Yamada, Paloma Fernandez Mira, Agustina Carando, Claudia H. Sanchez Gutierrez, and Kenji Sagae. 2020. Developing NLP tools with a new corpus of learner Spanish. In *Proceedings of The 12th Language Resources and Evaluation Conference*, pages 7238–7243, Marseille, France, May.

De Mauro, Tullio. 2016. *Il Nuovo Vocabolario di Base della Lingua Italiana*. Internazionale, http://www.internazionale.it/opinione/tullio-de-mauro/2016/12/23/il-nuovo-vocabolario-di-base-della-lingua-italiana.

Del Río Gayo, Iria, Marcos Zampieri, and Shervin Malmasi. 2018. A Portuguese native language identification dataset. In *Proceedings of the thirteenth workshop on innovative use of NLP for building educational applications*, pages 291–296, New Orleans, USA, June.

Di Nuovo, Elisa, Cristina Bosco, Alessandro Mazzei, and Manuela Sanguinetti. 2019. Towards an Italian learner treebank in universal dependencies. In *6th Italian Conference on Computational Linguistics, CLiC-it 2019*, volume 2481, pages 1–6, Bari, Italy, November. CEUR-WS.

Díaz-Negrillo, Ana, Nicolas Ballier, and Paul Thompson. 2013. *Automatic treatment and analysis of learner corpus data*, volume 59. John Benjamins Publishing Company.

Díaz-Negrillo, Ana, Detmar Meurers, Salvador Valera, and Holger Wunsch. 2010. Towards Interlanguage POS Annotation for Effective Learner Corpora in SLA and FLT. *Language Forum*, 36(1-2):139–154.

Díaz-Negrillo, Ana and Paul Thompson. 2013. Learner corpora. *Automatic treatment and analysis of learner corpus data*, 59.

Dickinson, Markus, Chris Brew, and Detmar Meurers. 2013. *Language and computers*. Wiley-Blackwell.

Dickinson, Markus and Marwa Ragheb. 2009. Dependency annotation for learner corpora. In *Proceedings of the Eighth Workshop on Treebanks and Linguistic Theories (TLT-8)*, pages 59–70, Milan, Italy, December.
Doval, Irene and M. Teresa Sánchez Nieto. 2019. *Parallel corpora for contrastive and translation studies: New resources and applications*, volume 90. John Benjamins Publishing Company.
EAGLES. 1996. *Preliminary recommendations on Corpus Typology*. Expert Advisory Group on Language Engineering Standards.
Garside, Roger, Geoffrey N. Leech, and Tony McEnery. 1997. *Corpus Annotation: Linguistic Information from Computer Text Corpora*. Taylor & Francis.
Granger, Sylviane. 2008. Learner Corpora. In Anke Lüdeling and Merja Kytö, editors, *Corpus Linguistics*, volume 1. Walter de Gruyter, pages 259–275.
James, Carl. 1998. *Errors in language learning and use*. Pearson Educational Limited.
Köhn, Christine and Arne Köhn. 2018. An annotated corpus of picture stories retold by language learners. In *Proceedings of the Joint Workshop on Linguistic Annotation, Multiword Expressions and Constructions (LAW-MWE-CxG-2018)*, pages 121–132, Santa Fe, New Mexico, USA, August.
Lee, John, Herman Leung, and Keying Li. 2017. Towards Universal Dependencies for Learner Chinese. In *Proceedings of the NoDaLiDa 2017 Workshop on Universal Dependencies (UDW 2017)*, pages 67–71, Gothenburg, Sweden, May.
Lee, John, Keying Li, and Herman Leung. 2017. L1-l2 parallel dependency treebank as learner corpus. In *Proceedings of the 15th International Conference on Parsing Technologies*, pages 44–49, Pisa, Italy, September.
Lennon, Paul. 1991. Error: Some problems of definition, identification, and distinction. *Applied linguistics*, 12(2):180–196.
Lüdeling, Anke. 2008. Mehrdeutigkeiten und Kategorisierung: Probleme bei der Annotation von Lernerkorpora. *Fortgeschrittene Lernervarietäten*, pages 119–140.
Lüdeling, Anke and Hagen Hirschmann. 2015. Error annotation systems. In S. Granger, G. Gilquin, and F. Meunier, editors, *The Cambridge handbook of learner corpus research*. Cambridge University Press, Cambridge, pages 135–157.
Lynn, Teresa. 2016. *Irish Dependency Treebanking and Parsing*. Ph.D. thesis, Dublin City University, Ireland and Macquarie University, Sydney, Australia.
Marello, Carla. 2011. Interpretare testi scritti composti a partire da storie disegnate. In Klaus Hölker and Carla Marello, editors, *Dimensionen der Analyse von Texten und Diskursen. Dimensioni dell'analisi di testi e discorsi.*, volume 1. LIT Berlin, pages 283–304.
McEnery, Tony and Andrew Wilson. 2001. *Corpus Linguistics: An Introduction (Second Edition)*. Edimburgh University Press, Edimburgh.
Meurers, Detmar. 2015. Learner corpora and natural language processing. *The Cambridge handbook of learner corpus research*, pages 537–566.
Meurers, Walt Detmar and Stefan Müller. 2009. Corpora and syntax (article 42). In Anke Lüdeling and Merja Kytö, editors, *Corpus linguistics*, volume 2. Mouton de Gruyter, Berlin, pages 920–933.
Nesselhauf, Nadja. 2004. Learner corpora and their potential for language teaching. *How to use corpora in language teaching*, 12:125–156.
Nivre, Joakim, Marie-Catherine de Marneffe, Filip Ginter, Jan Hajic, Christopher D. Manning, Sampo Pyysalo, Sebastian Schuster, Francis M. Tyers, and Daniel Zeman. 2020. Universal dependencies v2: An evergrowing multilingual treebank collection. *CoRR*, abs/2004.10643.
Popel, Martin, Zdenek Zabokrtský, and Martin Vojtek. 2017. Udapi: Universal API for Universal Dependencies. In Marie-Catherine de Marneffe, Joakim Nivre, and Sebastian Schuster, editors, *Proceedings of the NoDaLiDa Workshop on Universal Dependencies, UDW@NoDaLiDa 2017*, pages 96–101, Gothenburg, Sweden, May. Association for Computational Linguistics.
Ragheb, Marwa and Markus Dickinson. 2012. Defining syntax for learner language annotation. In *Proceedings of COLING 2012: Posters*, pages 965–974, Mumbai, India, December.
Ragheb, Marwa and Markus Dickinson. 2014a. Developing a corpus of syntactically-annotated learner language for English. In *Proceedings of the 13th International Workshop on Treebanks and Linguistic Theories (TLT13)*, pages 137–148, Tübingen, Germany, December.
Ragheb, Marwa and Markus Dickinson. 2014b. The effect of annotation scheme decisions on parsing learner data. *CLARIN-D*, pages 137–148.
Renzi, Lorenzo, Giampaolo Salvi, and Anna Cardinaletti. 2001. *Grande grammatica italiana di consultazione*, volume 1–3. Il mulino.

Reznicek, Marc, Anke Lüdeling, and Hagen Hirschmann. 2013. Competing target hypotheses in the falko corpus. *Automatic treatment and analysis of learner corpus data*, 59:101–123.
Rossini Favretti, Rema, Fabio Tamburini, and Cristiana De Santis. 2002. CORIS/CODIS: A corpus of written Italian based on a defined and a dynamic model. *A rainbow of corpora: Corpus linguistics and the languages of the world*, pages 27–38.
Sanguinetti, Manuela and Cristina Bosco. 2015. ParTUT: the Turin University Parallel Treebank. In *Harmonization and development of resources and tools for Italian natural language processing within the PARLI project*. Springer, pages 51–69.
Sanguinetti, Manuela, Cristina Bosco, Alberto Lavelli, Alessandro Mazzei, Oronzo Antonelli, and Fabio Tamburini. 2018. PoSTWITA-UD: an Italian Twitter Treebank in Universal Dependencies. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation*, pages 1768–1775, Miyazaki, Japan, May.
Selinker, Larry. 1972. Interlanguage. *International Review of Applied Linguistics*, 10(3):209–231.
Silveira, Natalia, Timothy Dozat, Marie Catherine De Marneffe, Samuel R. Bowman, Miriam Connor, John Bauer, and Christopher D. Manning. 2014. A Gold Standard Dependency Corpus for English. In *Proceedings of the 9th International Conference on Language Resources and Evaluation, LREC 2014*, pages 2897–2904, Reykjavik, Iceland, May. ELRA.
Simi, Maria, Cristina Bosco, and Simonetta Montemagni. 2014. Less is more? Towards a reduced inventory of categories for training a parser for the Italian Stanford Dependencies. In *Language Resources and Evaluation 2014*, pages 83–90, Reykjavik, Iceland, May. European Language Resources Association (ELRA).
Snover, Matthew, Bonnie Dorr, Richard Schwartz, Linnea Micciulla, and John Makhoul. 2006. A study of translation edit rate with targeted human annotation. In *Proceedings of the 7th Conference of the Association for Machine Translation in the Americas: Technical Papers*, pages 223–231, Cambridge, Massachusetts, USA, August.
Sparrow, Wendy. 2014. Unconventional word segmentation in emerging bilingual students' writing: A longitudinal analysis. *Applied linguistics*, 35(3):263–282.
Straka, Milan. 2018. UDPipe 2.0 prototype at CoNLL 2018 UD shared task. In *Proceedings of the CoNLL 2018 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies*, pages 197–207, Brussels, Belgium, October.
Tono, Yukio. 2003. Learner corpora: design, development and applications. In *Proceedings of the Corpus Linguistics 2003 conference*, pages 800–809, Lancaster, United Kingdom, March.
Zeman, Daniel, Jan Hajič, Martin Popel, Martin Potthast, Milan Straka, Filip Ginter, Joakim Nivre, and Slav Petrov. 2018. CoNLL 2018 shared task: Multilingual parsing from raw text to Universal Dependencies. In *Proceedings of the CoNLL 2018 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies*, pages 1–21, Brussels, Belgium, October. Association for Computational Linguistics.