Towards an Italian Learner Treebank in Universal Dependencies

Elisa Di Nuovo
Dipartimento di Lingue e Letterature
Straniere e Culture Moderne
University of Turin
elisa.dinuovo@unito.it

Cristina Bosco
Alessandro Mazzei
Manuela Sanguinetti
Dipartimento di Informatica
University of Turin
{bosco,mazzei,msanguin}@di.unito.it

Abstract

In this paper we describe the preliminary work on a novel treebank which includes texts written by learners of Italian drawn from the VALICO corpus. Data processing mostly involved the application of Universal Dependencies formalism and error annotation. First, we parsed the texts on UDPipe trained on the existent Italian UD treebanks, then we manually corrected them. The particular focus of this paper is on a one-hundred-sentence sample of the collection, used as a case study to define an annotation scheme for identifying the linguistic phenomena characterizing learners' interlanguage.

1 Introduction

The increasing interest in Learner Corpora (henceforth LC) is twofold motivated. On the one hand, LC are an especially valuable source of knowledge for interlanguage varieties. They allow in-depth comparisons of non-native varieties, helping to elucidate the properties of the interlanguage developed by learners with different mother tongues and learning levels. For this reason, LC are important resources enabling data-driven studies exploited within several research areas, such as Second Language Acquisition, Foreign Language Teaching, Contrastive Interlanguage Analysis, Computer-aided Error Analysis, Computer-Assisted Language Learning and L2 Lexicography (e.g. (Pravec, 2002; Granger, 2008; McEnery and Xiao, 2011)). On the other hand, LC have raised considerable computational interest, which is closely related to their usefulness in tasks such as Native Language Identification (Jarvis and Paquot, 2015; Malmasi, 2016), Grammatical-Error Detection and Correction (Leacock et al., 2015; Ng et al., 2014), and Automated Essay Scoring (Higgins et al., 2015).

In this paper we describe the development of a novel learner Italian treebank, i.e. VALICO-UD, in which Universal Dependencies (UD) formalism is tied to error annotation. The considerations of the annotation process, carried out on a set of one hundred sentences selected from a subcorpus of VALICO[1] (see Table 1) (Corino and Marello, 2017), allowed us to test a pilot scheme which pinpoints some of the features of L2 Italian.

This paper is organized as follows: in Section 2 we provide an overview of LC, focusing on Italian resources in particular; in Section 3 we present the data and the error annotation of VALICO-UD; in Section 4 we offer some examples of how we applied literal annotation to the learner sentences (LS) and, finally, in Section 6 we present conclusion and future work.

2 Related work

LC, also called interlanguage or L2 corpora, are collections of data produced by foreign or second language learners (Granger, 2008). Most LC projects were launched in the nineties and focused mainly on learner English (Tono, 2003), but recently we have witnessed an increasing interest in LC for other target languages. This has contributed to the establishment of learner corpus research (Tono, 2003).

LC can be enriched with Part of Speech (PoS) tagging, syntactic, semantic, discourse structure and error-tagging (with explicit or implicit target hypotheses[2]) annotation (Garside et al., 1997). To provide linguistic annotation, NLP tools are often used (Huang et al., 2018) and combined with

Copyright © 2019 for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

[1] http://www.valico.org/

[2] A reconstructed LS on which error identification is based (Reznicek et al., 2013).

human post-editing in order to overcome issues arising from the failures of the automatic analysis (Geertzen et al., 2013; Granger et al., 2009; Dahlmeier et al., 2013).

Among the 14 learner Italian corpora registered in the *Learner Corpora around the World* list[3], the majority are in the form of plain texts, or they only annotate PoS (COLI, LOCCLI and CAIL2[4], and VALICO), while only MERLIN (Boyd et al., 2014) annotates syntax and errors (with explicit target hypotheses).

Although MERLIN contains 816 texts written in non-native Italian (Boyd et al., 2014), they are not balanced for learners' mother tongue and are not annotated using a standard annotation for syntax, which would allow comparisons with other resources. To fill this gap, we decided to develop VALICO-UD, a L1-balanced resource developed within the UD formalism, thus providing a greater potential for contrastive analysis. Indeed, a UD-annotated LC can be compared with other LC (therefore different interlanguages) or also with native corpora of the L1 involved. For all these reasons, we decided to develop this new learner Italian treebank within the UD formalism. References were the English and Chinese experiences, respectively the English Second Language (ESL) (Berzak et al., 2016) and the Chinese Foreign Language (CFL) (Lee et al., 2017) treebanks.

The scholars involved in the annotation of the ESL and CFL treebanks decided to follow a well-established line of work, for which learner language analysis is centered upon morpho-syntactic surface evidence. This is motivated by various studies, e.g. (Díaz-Negrillo et al., 2010; Ragheb and Dickinson, 2012), in which the difference between morphological and distributional PoS is stressed. We decided to follow this line of research annotating discrepancies between morphological and distributional PoS, as described in the next sections. However, in lieu of carrying out manual annotation from scratch, such as in the ESL, we combined automatic annotation and manual post-editing (as shown in the next section).

3 Data and annotation

The data of VALICO-UD are drawn from the VALICO corpus (Corino and Marello, 2017), a collection of non-native Italian texts elicited by comic strips proposed to the learners. It consists of a selection of narrative and descriptive texts providing a large variety of structures beyond simple presentative/existential constructions.

The portion of VALICO that we selected for the treebank is made up of 237 texts (2,261 LS) organized in four sections as shown in Table 1.

L1	# Texts	# LS Tokens
English (EN)	60	8,285
French (FR)	59	7,301
German (DE)	58	7,417
Spanish (ES)	60	7,365
EN+FR+DE+ES	237	30,368

Table 1: VALICO-UD in figures – LS section.

Although the unpredictability and variation of a learner product, in terms of vocabulary, morphology and syntax, makes parsing a LC an especially challenging task (Corino and Russo, 2016; Díaz-Negrillo et al., 2010), it is highly recommendable for smoothly retrieving interlanguage features. Due to this peculiarity of interlanguage, keeping separated the LS from its specifically built target hypothesis (TH) is highly recommended (Lüdeling et al., 2005).

Our annotation scheme for learner Italian uses the inventory of the Italian UD PoS tags and dependency relations (Bosco et al., 2013; Bosco et al., 2014) and the related guidelines. In addition, we tried to follow as much as possible the ESL treebank to have comparable resources.

First, we trained UDPipe (Straka et al., 2016) on the Italian UD corpora, which include standard texts, ISDT (Bosco et al., 2014), and Twitter posts, POSTWITA-UD (Sanguinetti et al., 2018). Second, we automatically parsed VALICO-UD. Third, we manually corrected the treebank. This step is currently ongoing and we envision the treebank to be released in the UD repository in a few months.

For each sentence in VALICO-UD we provide two distinct versions both annotated in UD and tied to an error encoding system (see Section 3.1): one version for the LS and the other for its TH. The latter will differ from the former only when some errors occur. As a trial for this scheme, we selected one hundred sentences (i.e. sample set) containing each at least one error to be annotated.

[3] https://uclouvain.be/en/research-institutes/ilc/cecl/learner-corpora-around-the-world.html.

[4] COLI, LOCCLI and CAIL2 are developed at Università per Stranieri di Perugia and coordinated by Stefania Spina.

sent_id = NameSurname00135LS
text = Può essere un rubadore perche ha la cara chiusa e minacciata.
err = Può essere un ⟨RN⟩⟨i⟩rubadore⟨/i⟩⟨c⟩rubatore⟨/c⟩⟨/RN⟩ ⟨MI⟩⟨i⟩perche⟨/i⟩⟨c⟩perché⟨/c⟩⟨/MI⟩ ha la ⟨FNL⟩⟨i⟩cara⟨/i⟩ ⟨c⟩faccia⟨/c⟩⟨/FNL⟩ chiusa e ⟨DJ⟩⟨i⟩minacciata⟨/i⟩ ⟨c⟩minacciosa⟨/c⟩⟨/DJ⟩.
segment =
typo = 8 ADJ, 11 VERB
foreign = 8 NOUN
context = 4 NOUN

1	Può	potere	AUX	VM	_	4	aux
2	essere	essere	AUX	V	_	4	cop
3	un	uno	DET	RI	_	4	det
4	rubadore	rubadore	NOUN	S	_	0	root
5	perche	perché	SCONJ	CS	_	6	mark
6	ha	avere	VERB	V	_	4	advcl
7	la	il	DET	RD	_	8	det
8	cara	caro	NOUN	S	_	6	obj
9	chiusa	chiuso	ADJ	A	_	8	amod
10	e	e	CCONJ	CC	_	11	cc
11	minacciata	minacciato	ADJ	A	_	9	conj
12	.	.	PUNCT	FS	_	4	punct

sent_id = NameSurname00135TH
text = Può essere un rubatore perché ha la faccia chiusa e minacciosa.
err = Può essere un ⟨RN⟩⟨i⟩rubadore⟨/i⟩⟨c⟩rubatore⟨/c⟩⟨/RN⟩ ⟨MI⟩⟨i⟩perche⟨/i⟩⟨c⟩perché⟨/c⟩⟨/MI⟩ ha la ⟨FNL⟩⟨i⟩cara⟨/i⟩ ⟨c⟩faccia⟨/c⟩⟨/FNL⟩ chiusa e ⟨DJ⟩⟨i⟩minacciata⟨/i⟩ ⟨c⟩minacciosa⟨/c⟩⟨/DJ⟩.
segment =
typo = 8 ADJ, 11 VERB
foreign = 8 NOUN
context = 4 NOUN

1	Può	potere	AUX	VM	_	4	aux
2	essere	essere	AUX	V	_	4	cop
3	un	uno	DET	RI	_	4	det
4	rubatore	rubatore	NOUN	S	_	0	root
5	perché	perché	SCONJ	CS	_	6	mark
6	ha	avere	VERB	V	_	4	advcl
7	la	il	DET	RD	_	8	det
8	faccia	faccia	NOUN	S	_	6	obj
9	chiusa	chiuso	ADJ	A	_	8	amod
10	e	e	CCONJ	CC	_	11	cc
11	minacciosa	minaccioso	ADJ	A	_	9	conj
12	.	.	PUNCT	FS	_	4	punct

Figure 1: Example of two CoNLL-U trees of the LS (left) and TH (right) number #35: *He-can to-be a thief because he-has the face closed and threaten_PP*.

3.1 Error Annotation

In writing the TH we decided to adhere as much as possible to the LS and to focus on linguistic correctness (e.g. grammaticality) rather than linguistic appropriateness (e.g. register) (Reznicek et al., 2013)[5]. For this reason, sometimes we sacrificed naturalness for the sake of adherence to the LS. This principle was applied also to lexical errors requiring replacement. For instance, in Figure 1, the term "rubadore" in the LS was replaced with "rubatore" and not with its more common synonym "ladro", *thief*.[6] With this principle in mind, we decided to correct words if they are not present neither in the VINCA corpus[7] (the reference corpus specifically compiled for VALICO and containing texts based on the same comic strips but written by Italian native speakers) nor in our reference dictionary, *Il Nuovo Vocabolario di Base della Lingua Italiana* (De Mauro, 2016). In fact, the VINCA corpus is quite small and the language used sounds quite unnatural though being produced by speakers whose mother tongue is namely Italian (see Corino and Marello (2017, p. 12)).

Once the target hypotheses are written, we applied to them a coding system based on Nicholls (2003), which was used also in the NUCLE (Dahlmeier et al., 2013) and FCE (Yannakoudakis et al., 2011) corpora. Our system follows Nicholls's same principle: "the first letter represents the *general type of error* (e.g. wrong form, omission), while the second letter identifies the *word class of the required word*".

To provide a finer-grained description of errors, we used a large variety of letters in the first and second position (e.g. I: inflection, X: auxiliary) and a third letter which encodes information about some grammatical features (e.g. T: tense, M: mood, G: gender) (Simone, 2008, pp. 303–346) and other phenomena involved (e.g. capitalization, language transfer and government). Finally, Nicholls included a catch-all code (CE: complex error) to cover complex, multiple errors. In our sample set, we did not use it because we managed to describe all errors encountered using nested XML tags. However, we do not exclude that, applying the error codes to the whole corpus, we might find particularly complex errors which need to be marked using this code.

Figure 1 shows an annotation example of a LS along with its corresponding TH in the typical CoNLL-U format and with the resource-specific fields used to encode the error information. The **sent_id** field contains the identification code of the sentence: in the example, NameSurname001 (anonymized here) indicates the unique identifier of the text and refers to the transcribers name and surname; the following two-digit number, 35 in the example, indicates the position of the sentence in the text; finally, LS or TH indicates learner sentence and target hypothesis, respectively. The **text** field contains the uncoded sentence (which can be the learner sentence or the target hypothesis). The **err** field contains the error annotation based on

[5] In the future we plan to provide a second TH, focusing on linguistic appropriateness.

[6] Although "rubadore" is reported and marked as obsolete in the Italian Dictionary Olivetti, "rubatore" is the variant reported in De Mauro (2016), our reference dictionary.

[7] http://www.valico.org/vinca.html

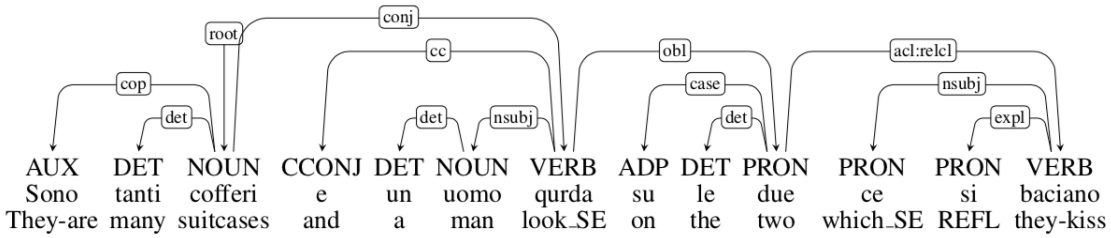

Figure 2: LS #10.

⟨SVS#⟩⟨MAX⟩⟨i⟩ Sono ⟨/i⟩⟨c⟩ Ci Sono ⟨/c⟩⟨/MAX⟩⟨i⟩ Ci Sono ⟨/i⟩⟨c⟩ Ci sono ⟨/c⟩ ⟨/SVS#⟩ ⟨IDG#⟩⟨FNL⟩⟨i⟩ tanti cofferi ⟨/i⟩⟨c⟩ tanti valigie ⟨/c⟩ ⟨/FNL⟩⟨i⟩ tanti valigie ⟨/i⟩⟨c⟩ tante valigie ⟨/c⟩⟨/IDG#⟩ e un uomo ⟨MAR⟩⟨i⟩⟨/i⟩⟨c⟩ che ⟨/c⟩⟨/MAR⟩ ⟨SV⟩⟨i⟩ qurda ⟨/i⟩⟨c⟩ guarda ⟨/c⟩⟨/SV⟩ ⟨UT⟩ ⟨IDG⟩ ⟨i⟩ sulle ⟨/i⟩⟨c⟩ sui ⟨/c⟩⟨/IDG⟩ ⟨i⟩ sui ⟨/i⟩⟨c⟩ i ⟨/c⟩⟨/UT⟩ due ⟨SAR⟩⟨i⟩ ce ⟨/i⟩⟨c⟩ che ⟨/c⟩⟨/SAR⟩ si baciano.

Figure 3: Error-annotated sentence #10.

the coding scheme introduced above. The **foreign** field includes the index and the PoS of the words which are considered errors due to language transfer. The **context** field contains the index and the PoS of the words which need replacement due to wrong context-bound lexical choices[8]. Finally, in line with the ESL, we used the **segment** field when a sentence was wrongly divided and the **typo** field to indicate PoS distributional-morphological discrepancies.

In the error-annotated sentence (the "err" field mentioned above), we report the wrong form(s) inside the ⟨i⟩_⟨/i⟩ tag and the corrected form(s) inside the ⟨c⟩_⟨/c⟩ tag. Figure 3 shows three examples of nested tag and two examples of *cascade* errors (i.e. an error which is due to the correction of another token) (Andorno and Rastelli, 2009, p. 52). The ⟨MAX⟩_⟨/MAX⟩ tag at the beginning of the sentence, for example, indicates a missing existential-construction pronoun, i.e. "Sono" (*are*) instead of "Ci sono" (*there are*). After the insertion of the missing pronoun "Ci", the capital "S" in "Sono" needs to be changed into a lowercase "s": this is a case in which we have a cascade capitalization error and we mark it adding a hashtag after the normal error code, as in ⟨SVS#⟩_⟨/SVS#⟩. Another cascade error is found in the next nested tag: we have an Inflection Determiner Gender error which is caused by the correction of the expression "tanti cofferi", involving a determiner and a noun ("cofferi" is a German word adapted to Italian and meaning luggages); thus, we have a cascade ⟨IDG#⟩_⟨/IDG#⟩ tag which embeds a ⟨FNL⟩_⟨/FNL⟩ tag (Form Noun Language_transfer). The next three tags, ⟨MAR⟩_⟨/MAR⟩, ⟨SAR⟩_⟨/SAR⟩ and ⟨SV⟩_⟨/SV⟩, indicate Missing pronoun (A) Relative ("che", *that*), Spelling pronoun Relative ("ce" instead of "che") and Spelling Verb errors ("qurda" instead of "guarda", *look*), respectively. There is, finally, another example of nested tag involving an Inflection Determiner Gender and an Unneccessary preposiTion errors; this has been used to indicate the multiple-step shift from the LS "sulle" (*on the_Fem_Pl*) to its TH counterpart "i" (*the_Masc_Pl*): the shift involved a change in the gender of the article (from feminine to masculine) and the drop of the preposition "su" (*on*), mistakenly used in the LS.

In order to ensure consistency across different annotators, the error annotation guidelines provide a hierarchical order to be applied when dealing with nested tags. We organized the errors in a pyramid with at the bottom mechanical errors (i.e. tokenization, capitalization, spelling and punctuation) and, proceeding towards the apex, morphological (derivation and inflection), lexical (form and replace), and syntactic (missing, unnecessary and word order) errors. For example, following this hierarchical order, mechanical errors should be corrected before a syntactic error. However, cascade errors make an exception and change the correction order, as we seen in Figure 3 in which we have a cascade capitalization error (SVS#) caused by a missing pronoun error (MAX)

[8]Only those choices in which there is no mismatch between distributional and morphological PoS are registered in this field.

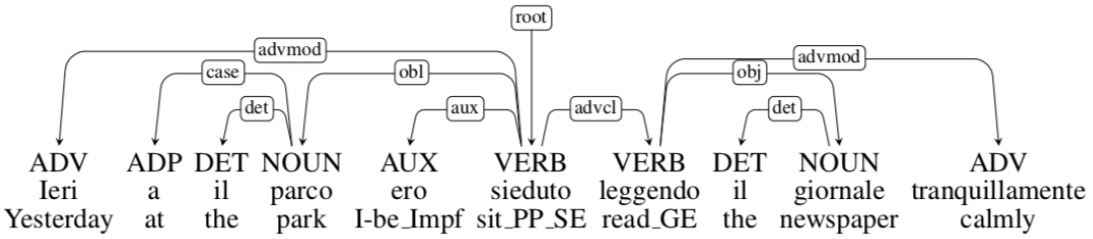

Figure 4: LS #88.

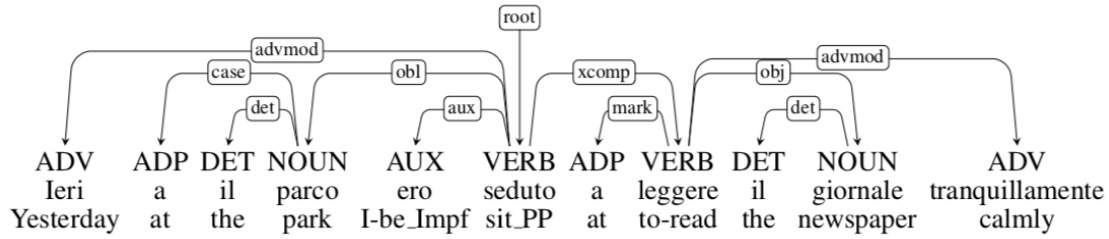

Figure 5: TH #88.

and a cascade inflection error (IDG#) due to a lexical error (FNL).

In the LS sample set, containing 1,860 tokens, we marked 496 errors (which represent 26,66% of the LS sample set tokens) distributed as shown in Table 2.

Error category	Tag	# occ	% tot
Derivation	D	24	4.84%
Form	F	71	14.31%
Inflection	I	72	14.51%
Spelling	S	92	18.55%
Word segmentation	T	16	3.22%
Word order	W	15	3.02%
Missing word	M	76	15.32%
Unnecessary word	U	55	11.09%
Replace word	R	75	15.12%
Total	–	496	–

Table 2: Error categories as encoded in the first letter (general error type) and their distribution in the sample set.

4 From VALICO to VALICO-UD

In this Section we describe how we applied literal annotation to the (morpho-)syntactic structure of the LS in particular, relying on the Universal Dependencies scheme.

Literal Annotation

We annotated UD PoS and relations sticking as much as possible to the literal reading of the learner sentence, thereby creating a treebank in line with the two existing learner treebanks in the UD framework (ESL and CFL).

Argument Structure: When some extraneous or unnecessary prepositions occur, we annotate the dependencies accordingly. Figure 2 shows a LS in which the verb "guardare", *look*, is used as an intransitive verb, thus we annotate its direct object as an oblique[9].

Missing or Unnecessary Words: We annotate literally when there are missing or unnecessary words. In the example in Figure 2 the clitic pronoun "ci" is missing , thus we treated "sono" as a copular verb. There are other cases in which the clitic pronoun "ci" is mistakenly combined with the verb *to be* forming an existential clause, and consequently causing a distributional mismatch (e.g. LS: "[...] non *ci era pericoloso* o *violento*", TH: "[...] non *era* pericoloso o violento"[10]). In these cases we mark in the "typo" field the morphological PoS and in the PoS column the distributional PoS, cf. Figure 1.

Extraneous Word Forms: When the learner misuses existent word forms, we annotate them literally. In Figure 4, the learner used a gerund, "leggendo" (*reading*), instead of the infinitive " a

[9]In all the examples SE stands for spelling error, REFL for reflexive pronoun, PP for past participle, GE for gerund and Impf for imperfect tense.

[10]LS: "[...] not *there it-be_Impf dargerous* or *violent*", TH: "[...] not *it-be_Impf* dangerous or violent".

leggere" (*to read*). We then labeled it as an adverbial clause in the LS (Figure 4) and as an open clausal complement in the TH (Figure 5).

Exceptions to Literal Annotation

Spelling: Some examples of spelling errors are presented in Figure 2. We lemmatize and PoS-tag them referring to their correct versions, similarly to Andorno and Rastelli (2009, p. 58). Thus, "ce" was treated as "che", *which*,[11], and "qurda" as "guarda" *look*.

Word Formation: We do not treat literally valid words that are contextually implausible. We consider them differently depending on the PoS of the intended word: if the intended word has the same PoS we signal it in the "context" field (e.g. LS: "[...] salvando una ragazza *indefessa*", TH: "[...] salvando una ragazza *indifesa*"[12]), if it is different in the "typo" field (cf. Figure 1).

Nonexistent Words: In cases in which the learner wrote a word which does not exist in Italian and it is arguably a foreign word, we signal it in the "foreign" field[13]. In the example in Figure 1 the word "cara" (i.e. an adjective translatable into *beloved*) is arguably a transfer from the Spanish noun meaning *face*. In this case we lemmatize it with the correct lemma of "cara". In addition, in the "typo" field we mark the occurring mismatch between distributional and morphological PoS.

Word Tokenization: If one word is mistakenly segmented into two, we use the "goeswith" relation, as germane to UD annotation guidelines[14]. If two words are mistakenly segmented into one, we use X as PoS and decide the relation on a case-by-case basis. For example in LS: "[...] butta tutto *perterra*", TH: "[...] butta tutto *per terra*"[15] we assigned to "perterra" PoS 'X' and dependency relation 'obl'.

5 Inter-Annotator Agreement

As stated above, the complete manual revision of the treebank is still in progress; however, with the aim of assessing the annotation quality of this preliminary sample set, as well as the quality of the annotation guidelines (especially the ones concerning the LS section) both LS and TH sections were annotated by two independent annotators. The inter-annotator agreement was then 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)[16]. The results are reported in Table 3, 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 the guidelines clarity and of the annotators' consistency, even when dealing with non-canonical syntactic structures.

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

Table 3: Agreement results on the sample set of both LS and TH.

6 Conclusion and future work

In this paper we introduced VALICO-UD and proposed an annotation scheme suitable for texts of learner Italian encompassing both UD and error annotation. Our scheme follows the principle of "literal annotation" and takes PoS and dependency morphological-distributional mismatches into account. Our error tag set seems adequate to bookmark errors, providing also a fine-grained description of some of them.

There are a number of possible applications for the monolingual parallel treebank proposed in this paper. In the near future, we plan to apply the tree edit distance to LS and TH to measure linguistic competence. Recently, the tree edit distance has been applied to various tasks (Emms, 2008; Tsarfaty et al., 2011; Plank et al., 2015), and a study has formalized the notion of *syntactic anisomorphism* (Ponti et al., 2018). We aim to explore a correlation between these notions and the linguistic competence to describe the achievements of foreign language learners.

[11] When "ce" is used instead of "c'è", *there is*, we treat it as a single token and mark it as root, in line with what we would have done if it were "c'è".

[12] LS: "[...] saving a *untiring* girl", TH: "[...] saving a *vulnerable* girl".

[13] The lemma will be its Italian (quasi-)equivalent.

[14] https://universaldependencies.org/u/overview/typos.html

[15] [...] he-throw everything *on the ground*.

[16] http://universaldependencies.org/conll18/evaluation.html

References

Cecilia Maria Andorno and Stefano Rastelli. 2009. Un'annotazione orientata alla ricerca acquisizionale. In Cecilia Maria Andorno and Stefano Rastelli, editors, *Corpora di italiano L2: tecnologie, metodi, spunti teorici*, pages 49–70. Guerra.

Yevgeni Berzak, 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.

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

Cristina Bosco, Felice Dell'Orletta, Simonetta Montemagni, Manuela Sanguinetti, and Maria Simi. 2014. The EVALITA 2014 Dependency Parsing Task. In *Proceedings of EVALITA 2014 Evaluation of NLP and Speech Tools for Italian*, pages 1–8.

Adriane Boyd, 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.

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

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

Elisa Corino 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.

Daniel Dahlmeier, 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.

Tullio De Mauro. 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.

Ana Díaz-Negrillo, 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.

Martin Emms. 2008. Tree Distance and Some Other Variants of Evalb. In *Proceedings of the Sixth International Conference on Language Resources and Evaluation*, pages 1373–1379.

Roger Garside, Geoffrey N. Leech, and Tony McEnery. 1997. *Corpus Annotation: Linguistic Information from Computer Text Corpora*. Taylor & Francis.

Jeroen Geertzen, Theodora Alexopoulou, and Anna Korhonen. 2013. Automatic Linguistic Annotation of Large Scale L2 Databases: The EF-Cambridge Open Language Database (EFCAMDAT). In *Proceedings of the 31st Second Language Research Forum*, pages 240–254.

Sylviane Granger, Estelle Dagneaux, Fanny Meunier, and Magali Paquot. 2009. *International Corpus of Learner English*. Louvain University Press.

Sylviane Granger. 2008. Learner Corpora. In Anke Lüdeling and Merja Kytö, editors, *Corpus Linguistics*, volume 1, pages 259–275. Walter de Gruyter.

Derrick Higgins, Chaitanya Ramineni, and Klaus Zechner. 2015. Learner Corpora and Automated Scoring. In Sylviane Granger, Gaëtanelle Gilquin, and Fanny Meunier, editors, *The Cambridge Handbook of Learner Corpus Research*, pages 587–604. Cambridge University Press.

Yan Huang, Akira Murakami, Theodora Alexopoulou, and Anna Korhonen. 2018. Dependency Parsing of Learner English. *International Journal of Corpus Linguistics*, 23(1):28–54.

Scott Jarvis and Magali Paquot. 2015. Learner Corpora and Native Language Identification. In Sylviane Granger, Gaëtanelle Gilquin, and Fanny Meunier, editors, *The Cambridge Handbook of Learner Corpus Research*, pages 605–628. Cambridge University Press.

Claudia Leacock, Martin Chodorow, and Joel Tetrault. 2015. Automatic Grammar- and Spell-Checking for Language Learners. In Sylviane Granger, Gaëtanelle Gilquin, and Fanny Meunier, editors, *The Cambridge Handbook of Learner Corpus Research*, pages 567–586. Cambridge University Press.

John Lee, 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.

Anke Lüdeling, Maik Walter, Emil Kroymann, and Peter Adolphs. 2005. Multi-level error annotation in learner corpora. In *Proceedings of Corpus Linguistics 2005*, volume 1, pages 14–17.

Teresa Lynn. 2016. *Irish Dependency Treebanking and Parsing*. Ph.D. thesis, Dublin City University, Ireland and Macquarie University, Sydney, Australia.

Shervin Malmasi. 2016. *Native Language Identification: explorations and applications*. Ph.D. thesis, Macquarie University, Sydney, Australia.

Tony McEnery and Richard Xiao. 2011. What corpora can offer in language teaching and learning. In Eli Hinkel, editor, *Handbook of Research in Second Language Teaching and Learning*, volume 2, pages 364–380. Routledge.

Hwee Tou Ng, Siew Mei Wu, Ted Briscoe, Christian Hadiwinoto, Raymond Hendy Susanto, and Christopher Bryant. 2014. The CoNLL-2014 shared task on grammatical error correction. In *Proceedings of the Eighteenth Conference on Computational Natural Language Learning: Shared Task*, pages 1–14.

Diane Nicholls. 2003. The Cambridge Learner Corpus: Error coding and analysis for lexicography and ELT. In *Proceedings of the Corpus Linguistics 2003 Conference*, volume 16, pages 572–581.

Barbara Plank, Héctor Martínez Alonso, Željko Agić, Danijela Merkler, and Anders Søgaard. 2015. Do dependency parsing metrics correlate with human judgments? In *Proceedings of the Nineteenth Conference on Computational Natural Language Learning*, pages 315–320.

Edoardo Maria Ponti, Roi Reichart, Anna Korhonen, and Ivan Vulić. 2018. Isomorphic transfer of syntactic structures in cross-lingual NLP. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics*, pages 1531–1542.

Norma A. Pravec. 2002. Survey of Learner Corpora. *ICAME journal*, 26(1):8–14.

Marwa Ragheb and Markus Dickinson. 2012. Defining syntax for learner language annotation. In *Proceedings of COLING 2012: Posters*, pages 965–974.

Marc Reznicek, Anke Lüdeling, and Hagen Hirschmann. 2013. Competing target hypotheses in the falko corpus. In Ana Díaz-Negrillo, Nicolas Ballier, and Paul Thompson, editors, *Automatic treatment and analysis of learner corpus data*, volume 59, pages 101–123. John Benjamins Publishing Company.

Manuela Sanguinetti, 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.

Raffaele Simone. 2008. *Fondamenti di linguistica*. Laterza.

Milan Straka, Jan Hajic, and Jana Straková. 2016. Udpipe: Trainable pipeline for processing conll-u files performing tokenization, morphological analysis, pos tagging and parsing. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation*, pages 88–99.

Yukio Tono. 2003. Learner corpora: design, development and applications. In *Proceedings of the Corpus Linguistics 2003 conference*, pages 800–809.

Reut Tsarfaty, Joakim Nivre, and Evelina Andersson. 2011. Evaluating dependency parsing: Robust and heuristics-free cross-annotation evaluation. In *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*, pages 385–396.

Helen Yannakoudakis, Ted Briscoe, and Ben Medlock. 2011. A new dataset and method for automatically grading esol texts. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 180–189.

Daniel Zeman, 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 1–19.