



UNIVERSITÀ
DEGLI STUDI
DI TORINO



UNIVERSITÀ
DEGLI STUDI
DI GENOVA

DOCTORAL THESIS

VALICO-UD: annotating an Italian learner corpus

Author:

Elisa DI NUOVO

Supervisors:

Prof. Cristina BOSCO

Prof. Elisa CORINO

*A thesis submitted in fulfillment of the requirements
for the degree of Doctor of Philosophy in the
Dipartimento di Lingue e Letterature Straniere e Culture Moderne
and in the
Dipartimento di Informatica*

Scuola di Dottorato in Digital Humanities

Academic years: 2018/2019, 2019/2020, 2020/2021

Academic disciplines: L-LIN/01 and INF/01

Declaration of Authorship

I, Elisa DI NUOVO, declare that this thesis titled, “VALICO-UD: annotating an Italian learner corpus” and the work presented in it are my own. I confirm that:

- This work was done wholly while in candidature for a research degree at this University.
- Where any part of this thesis has previously been submitted for a degree or any other qualification at this University or any other institution, this has been clearly stated.
- Where I have consulted the published work of others, this is always clearly attributed.
- Where I have quoted from the work of others, the source is always given. With the exception of such quotations, this thesis is entirely my own work.
- I have acknowledged all main sources of help.
- Where the thesis is based on work done by myself jointly with others, I have made clear exactly what was done by others and what I have contributed myself.

Some parts of this thesis include revised versions of the following papers and talks:

- Di Nuovo, Elisa, Cristina Bosco, Alessandro Mazzei, and Manuela Sanguinetti. "Towards an Italian learner treebank in universal dependencies." In *6th Italian Conference on Computational Linguistics, CLiC-it 2019*, vol. 2481, pp. 1-6. CEUR-WS, 2019.
- Di Nuovo, Elisa, Cristina Bosco, and Elisa Corino. "How good are humans at Native Language Identification? A case study on Italian L2 writings." In *CLiC-it 2020 Italian Conference on Computational Linguistics 2020*, pp. 1-7. CEUR, 2020.

- Di Nuovo, Elisa, Manuela Sanguinetti, Alessandro Mazzei, Elisa Corino, and Cristina Bosco. "VALICO-UD: Treebanking an Italian Learner Corpus in Universal Dependencies." In *Italian Journal of Computational Linguistics*. (Accepted December 2021).
- Di Nuovo, Elisa, Cristina Bosco, and Elisa Corino. "Analisi sintattica e dell'errore in un corpus di apprendenti di italiano: il sintagma nominale." Oral presentation, 11-13 February 2021, Associazione Italiana Linguistica Applicata (AIItLA 2020).
- Di Nuovo, Elisa, Cristina Bosco, and Elisa Corino. "Etichettare gli errori in un corpus di apprendenti: problemi, strategie, soluzioni." Oral presentation, DILLE, Pisa, 2020. (Postponed to May 2022).
- Di Nuovo, Elisa, and Cristina Bosco. "VALICO-UD: Treebanking an Italian Learner Corpus." Oral presentation at Conversazioni Linguistiche, UNITN, 9 April 2020.
- Di Nuovo Elisa, Bianca Maria De Paolis, Cristina Bosco, and Elisa Corino. "Error identification, normalization and tagging: three inter-annotator agreement experiments in a picture-elicited learner corpus." Oral presentation, Learner Corpus Research, Padua, 2022.

Signed:

Date:

“Words and sentences [...] have no content of their own, but if they encounter someone who listens they become something. We are part of the data.”

Andrea Moro

I speak, therefore I am. Seventeen Thoughts About Language

Abstract

Previous work on learner language has highlighted the importance of having annotated resources to describe the development of interlanguage. Despite this, few learner resources, mainly for English L2, feature error and syntactic annotation.

This thesis describes the development of a novel parallel learner Italian treebank, VALICO-UD. Its name suggests two main points: where the data comes from—i.e. the corpus VALICO, a collection of non-native Italian texts elicited by comic strips—and what formalism is used for linguistic annotation—i.e. Universal Dependencies (UD) formalism. It is a parallel treebank because the resource provides for each learner sentence (LS) a target hypothesis (TH) (i.e., parallel corrected version written by an Italian native speaker) which is in turn annotated in UD.

We developed this treebank to be exploitable for interlanguage research and comparable with the resources employed in Natural Language Processing tasks such as Native Language Identification or Grammatical Error Identification and Correction.

VALICO-UD is composed of 237 texts written by English, French, German and Spanish native speakers, which correspond to 2,234 LSs, each associated with a single TH. While all LSs and THs were automatically annotated using UDPipe, only a portion of the treebank made of 398 LSs plus correspondent THs has been manually corrected and released in May 2021 in the UD repository. This core section features also an explicit XML-based annotation of the errors occurring in each sentence. Thus, the treebank is currently organized in two sections: the core gold standard—comprising 398 LSs and their correspondent THs—and the silver standard—consisting of 1,836 LSs and their correspondent THs.

In order to contribute to the computational investigation about the peculiar type of texts included in VALICO-UD, this thesis describes the annotation schema of the resource, provides some preliminary tests about the performance of UDPipe models on this treebank, reports on inter-annotator agreement results for both error and linguistic annotation, and suggests some possible applications.

Acknowledgements

Prima di tutto, voglio ringraziare i miei supervisori di dottorato, le Professoressa Cristina Bosco ed Elisa Corino, per tutto il loro aiuto, i loro preziosi consigli, il loro sostegno e incoraggiamento. Le ringrazio anche per la loro pazienza. Questo progetto, iniziato su loro suggerimento, non sarebbe mai arrivato a questo punto senza di loro. In particolare, voglio ringraziare la Prof. Bosco per avermi coinvolta in alcuni dei suoi progetti e per avermi iniziata al mondo di treebank, Universal Dependencies e \LaTeX ; la Prof. Corino per la linguistica dei corpora e, in particolare, dei corpora di apprendenti.

Ringrazio Manuela Sanguinetti, Alessandro Mazzei e Valerio Basile, che, nonostante non fossero i miei supervisori di tesi, sono stati sempre disponibili e gentili con me durante questi anni di dottorato e non hanno mai esitato a dedicarmi tempo, conoscenza ed esperienza.

Un sentito grazie anche alla Prof.ssa Marelli, che mi ha donato del materiale bibliografico indispensabile per questa tesi.

Inoltre, vorrei esprimere la mia gratitudine ai miei amici e colleghi (alcuni già PhD, altri lo diverranno presto, tutti ricercatori dal grande potenziale): Carola Borgia, Stefania Cicillini, Alessandra Teresa Cignarella, Davide Colla, Valentina De Iacovo, Bianca Maria De Paolis, Mirko Lai e Maria Carmela Zaccone. Sono stati un supporto non solo a livello umano, ma anche professionale, condividendo con me tempo, risate, risorse e conoscenze, indispensabili per questo progetto.

Un ringraziamento speciale va a Martin Popel, PhD, che è stato sempre disponibile nell'aiutarmi a usare UDAPI, e ai revisori anonimi, che hanno fornito un feedback sulle pubblicazioni prodotte durante questo dottorato.

Le parole sono saranno mai sufficienti per esprimere la mia riconoscenza alla mia famiglia, a cui dedico questa tesi. Ad Alberto, che nel frattempo è diventato mio marito, per essere stato paziente e di supporto durante questi anni e per avermi impedito di impazzire durante il lockdown. Ai miei genitori che mi hanno sempre amato incondizionatamente e hanno creduto in me andando oltre le mie reali capacità. Ai miei fratelli, che mi hanno sostenuto, ognuno a suo modo, sin da quando sono venuta al mondo. Alla mia cara zia, che è una seconda madre per me. Grazie a tutti voi per avermi mostrato ogni giorno cos'è l'amore.

x

Infine, anche se non leggeranno mai queste parole, voglio ringraziare Zoe, Mei e Wen, e le mie orchidee, per essere il mio promemoria quotidiano del potere della bellezza.

Contents

Declaration of Authorship	iii
Abstract	vii
Acknowledgements	ix
1 Introduction	1
1.1 Research Questions	3
1.2 Overview	4
2 Literature review	5
2.1 Learner corpora	5
2.1.1 Learner corpora and NLP tasks	10
2.2 Error annotation	10
2.2.1 Defining <i>error</i>	12
2.2.2 Describing errors	13
2.2.3 Error-tagging systems	17
2.2.3.1 Target hypothesis explicitness	17
2.2.3.2 Error annotation scope	18
2.2.3.3 Error annotation technique	18
2.2.4 Inter annotator agreement	19
2.2.5 Implicit error annotation	22
2.3 Linguistic annotation	22
2.3.1 PoS tagging, lemmatization and morphological feature annotation	23
2.3.1.1 PIL2	23
2.3.1.2 NOCE	24
2.3.1.3 SALLE	25
2.3.1.4 ESL and CFL	26
2.3.2 Syntactic annotation	27
2.3.2.1 SALLE	29

2.3.2.2	ESL and CFL	30
2.3.2.3	Other approaches	31
3	VALICO-UD design	33
3.1	Data description	33
3.1.1	Gold standard	39
3.1.2	Silver standard	41
3.2	Target Hypothesis writing	43
3.2.1	Guiding principles	46
4	Error Annotation	51
4.1	Methodology	51
4.2	Tagset description	57
4.2.1	Spelling errors	60
4.2.1.1	Generic spelling errors	60
4.2.1.2	Issues with capitalization	60
4.2.1.3	Issues with double letters	61
4.2.1.4	Issues with punctuation	61
4.2.1.5	Issues with apostrophes	61
4.2.1.6	Issues with graphic accent	62
4.2.1.7	Issues with word boundaries	62
4.2.2	Derivation errors	63
4.2.3	Form errors	64
4.2.4	Inflection errors	66
4.2.4.1	Aspect	67
4.2.5	Unnecessary, missing and replace word errors	68
4.2.5.1	Unnecessary word errors	68
4.2.5.2	Missing word errors	69
4.2.5.3	Replace word or phrase errors	70
4.2.6	Word order errors	72
4.2.7	Complex errors	73
4.3	Error distribution per macro-categories and L1s	74
4.4	Inter Annotator Agreement	79
4.4.1	Methodology	80
4.4.2	Experiment 1 and 2: error identification and normal- ization	81
4.4.2.1	Quali-quantitative disagreement analysis	83
4.4.3	Experiment 3: error coding system evaluation	94

4.4.3.1	Qualitative disagreement analysis	95
5	Linguistic annotation	99
5.1	Treebanking VALICO-UD	99
5.1.1	Universal Dependencies formalism	99
5.1.2	Annotation scheme	103
5.1.2.1	Segmentation	103
5.1.2.2	Tokenization	105
5.1.2.3	Lemmatization	106
5.1.2.4	PoS Tagging	108
5.1.2.5	Morphological annotation	110
5.1.2.6	Dependency annotation	113
5.1.2.7	Multi-word Expressions	121
5.2	Inter Annotator Agreement	123
5.3	Treebank statistics	124
5.3.1	Distribution of PoS tags	124
5.3.2	Distribution of dependency relations	125
5.4	Incremental parsing evaluation	130
5.4.1	Error analysis	132
5.4.1.1	Evaluation of automatic PoS tagging	133
5.4.1.2	Evaluation of automatic parsing	135
6	Quantitative data exploration	141
6.1	Quantifying the parser performance	141
6.2	Quantifying learners' proficiency through machine translation metrics	143
6.2.1	Quantifying string distance	144
6.2.2	Quantifying tree distance	146
7	Conclusions	149
7.1	Future work	152
	Bibliography	155

List of Figures

3.1	Header provided per each VALICO text. Proper names are darkened.	34
3.2	TTM text before preprocessing.	35
3.3	TTM text after preprocessing.	36
3.4	Handwriting of <i>passava</i>	36
3.5	“Ieri al parco...” (<i>Yesterday at the park...</i>) comic strip from VALICO.	40
3.6	“Stazione” (<i>Station</i>) comic strip from VALICO.	42
4.1	Visualization of LS and TH using the edited version of Transcript’omatic.	52
4.2	<i>Err</i> field in the LS CoNLL-U file.	53
4.3	<i>Err</i> field in the TH CoNLL-U file.	53
4.4	Adding error tags in sentences with the edited version of Transcript’omatic.	55
4.5	Distribution of spelling errors and its subcategories in the core section of VALICO-UD.	62
4.6	Distribution of form errors in the core section of VALICO-UD.	65
4.7	Distribution of inflection errors in the core section of VALICO-UD.	67
4.8	Distribution of errors marked as unnecessary in the core section of VALICO-UD.	69
4.9	Distribution of missing word errors in the core section of VALICO-UD.	70
4.10	Distribution of replacement errors in the core section of VALICO-UD.	72
4.11	Distribution of word order errors in the core section of VALICO-UD.	73
4.12	Distribution of errors per macro-categories in the core section of VALICO-UD.	76
4.13	Error distribution per L1s in the core section of VALICO-UD.	77

4.14	Sources of disagreement in Experiment 1 and 2.	89
4.15	Experiments 1 and 2: distribution of categories involved in apparent disagreement.	90
6.1	Histograms of the TER values obtained for each text of the two classes, initial (Group 1) and advanced (Group 2) learners of Italian.	146
6.2	Histograms of the F1 values obtained for each text of the two classes, initial (Group 1) and advanced (Group 2) learners of Italian.	147

List of Tables

2.1	Learner corpora overview. Error-annotated corpora are in bold.	7
3.1	Summary of VALICO-UD composition. T1 and T2 stands for the two different topics eliciting the texts.	38
3.2	Summary of VALICO-UD core section.	39
3.3	Core section summary according to selection criteria (mean and standard deviation in brackets).	41
3.4	Summary of VALICO-UD silver data.	43
4.1	Summary of the letters allowed in position 1, 2 and 3 to form error tags, and their meaning.	59
4.2	Distribution of errors per macro-categories and learners' L1s in the core section of VALICO-UD.	75
4.3	Error density (ED) per error macro-category (Tag) and L1. . .	78
4.4	Summary of the two statistical tests performed. In the second column we have the grouping criteria (i.e. in the first test it is the L1, in the second test it is Year of Study (YoS)) used to distinguish the populations.	78
5.1	Agreement results on the sample set of both LSs and THs. . .	124
5.2	PoS tag distribution in the core section of VALICO-UD.	125
5.3	Classification of UD relations.	125
5.4	Dependency relation distribution in the core section of VALICO-UD.	126
5.5	Incremental evaluation of THs.	132
5.6	Incremental evaluation of LSs.	132
5.7	Evaluation of automatic UPoS tagging.	133
5.8	Evaluation of automatic parsing per dependency relation (Part 1).	136
5.9	Evaluation of automatic parsing per dependency relation (Part 2).	137

6.1	UDPIPE models' LAS when trained on ISDT and PoSTWITA and tested on LSs and THs.	142
6.2	Texts and metadata of the silver standard. Texts selected for the exploration in bold. The question mark indicates that the year of study is <i>not known</i>	145

To my family who has always had my back

Chapter 1

Introduction

Research on learner language is tackled mainly with two approaches: Second Language Acquisition (SLA) research approach and Learner Corpus Research (LCR) approach. The first one studies the acquisition of interlanguage¹ by L2 (i.e. second or foreign language) learners relying on highly controlled elicitation and experimental data, which usually are not shared with the research community. SLA research is characterized by the influence of cognitive psychology, cognitive linguistics and sociology to explain learners' developing second-language systems, relying more on longitudinal studies than cross-sectional ones, and focusing mainly on spoken language of a limited number of learners (Giacalone Ramat, 2003; Ellis, 2015). The second relies on large electronic collections of non-native natural language use data (i.e. learner corpora), consisting mainly of written texts, which are usually freely available or purchasable. LCR uses the Contrastive Interlanguage Analysis (CIA) (Granger, 1996; Granger, 2015) and the Computer-aided Error Analysis (CEA) (Dagneaux, Denness, and Granger, 1998; Díaz-Negrillo and Domínguez, 2006; Lüdeling and Hirschmann, 2015) methods. CIA consists in comparing interlanguage with native language and/or different interlanguages with each other. CEA is a method for analysing learner errors which tries to avoid the shortcomings of previous error analysis (Corder, 1967; Corder, 1971) and aims at tailoring pedagogy to learners' actual needs. Despite some of the shortcomings of previous error analysis have been overcome with CEA, error annotation is still carried out by only one annotator, thus no inter-annotator agreement is reported. Error annotation, being a highly subject task, would benefit if associated to reliability coefficients (e.g.

¹Interlanguage is the term introduced by Selinker, 1972 for indicating the "mental grammar that a learner construct and reconstruct" (Ellis, 2015, p. 20) giving learner language the status of language by itself.

Inter-Annotator Agreement, IAA). In fact, recently, a few studies have reported IAA (Köhn and Köhn, 2018; Rosen et al., 2014; Del Río Gayo and Mendes, 2018b; Larsson, Paquot, and Plonsky, 2020). The results reported in these studies have highlighted different degrees of agreement depending on the error type, but also on the text type (higher agreement in texts elicited in more controlled conditions, e.g. picture-elicited texts).

Whilst the first learner corpora, which emerged in the nineties of the last century, were all focused on learner English (Tono, 2003), recently, we have witnessed the emergence of learner corpora in languages different than learner English (Siemen, Lüdeling, and Müller, 2006; Tenfjord, Meurer, and Hofland, 2006; Lozano, 2009; Hana et al., 2012; Mendes et al., 2016; Ruzaitė et al., 2020). In addition, a more recent trend is to exploit learner corpora for Natural Language Processing (NLP) tasks, and to compile them purposely with NLP tasks at mind (e.g. Blanchard et al., 2013; Köhn and Köhn, 2018; Bryant et al., 2019; Davidson et al., 2020). NLP refers to automatically analyse natural language (i.e. human language). The analysis spans from relatively simple tasks (e.g. splitting sequences of characters into words or sentences, i.e. tokenization and segmentation, respectively) to more complex tasks (e.g. annotating syntactic features). There are also learner-specific NLP tasks, such as Native Language Identification or Grammatical Error Identification and Correction. The first task consists in automatically detecting the native language of a learner based only on their (written or spoken) text in their target language. The second task refers to two tasks that can be tackled together or separately, i.e. identifying grammatical errors in a learner text and propose a correction. NLP tools, thus, can be used to analyse a wide range of linguistic phenomena at scale on learner texts. However, since the majority of these tools are trained using native texts, it is important to know their strengths and limitations. In fact, not all the linguistic features can be annotated with a reasonably high degree of accuracy (Lu, 2010).

As far as learner Italian is concerned, there is no learner corpus featuring error annotation nor manual linguistic annotation available publicly, thus there are no studies reporting IAA nor measuring the performance loss of NLP tools applied to learner language.

To fill this gap, this dissertation describes the development of a novel L2 Italian publicly available resource, richly annotated and compiled to fulfill different research goals that can be of interest for SLA, LCR and NLP.

The novel resource, VALICO-UD, consists of 237 texts, drawn from the well-known L2 Italian corpus VALICO² (Corino and Marello, 2017), elicited by two comic strips³ written by 237 German (DE), English (EN), Spanish (ES) and French (FR) native speakers. In selecting this VALICO subcorpus, we tried to create a balanced selection taking L1s and topics into account. This language grouping is of interest because it could shed light on the influence of L1s and previously known L2s in the acquisition of the target language (which is Italian in this case), focusing in particular on the impact of language typology similarity. A subcorpus of VALICO-UD, made of 36 texts, L1 balanced and all elicited by a single comic strip, has already been released in the Universal Dependencies (UD) repository. This subcorpus, is completely manually-checked and features also error annotation.

1.1 Research Questions

The research questions that have driven my research are:

1. Is error identification more reliable if linguistic and extra-linguistic context is given?
2. When different annotators agree on the presence of an error, do they agree also on its normalization?
3. Is error annotation more reliable with explicit target hypotheses provided?
4. Is Universal Dependencies (UD) formalism adaptable to L2 Italian? Considering that UD has been successfully applied on standard varieties of Italian (e.g. legal texts, newspapers and Wikipedia) and on social media texts (i.e. Twitter), we want to see if the format can address also the challenges of interlanguage, by testing its repertoire of labels against VALICO data.
5. What is the performance loss of a parser trained on standard texts when applied to learner language? That is, how much interlanguage features

²VALICO is the biggest publicly available learner Italian corpus with a rich collection of metadata, about not only the writer but also the typology of writing. The corpus is available here: <http://www.valico.org>.

³Amore e Stazione, available here: <http://www.valico.org/vignette/amore.pdf> and <http://www.valico.org/vignette/stazione.pdf>.

affect parser performance? Systematic differences between native and learner language can emerge when we apply on the latter automatic annotation tools usually developed for the former.

6. Can the similarity between LSs and THs—expressed using quantitative machine translation evaluation metrics such as Translation Error Rate (TER)—be exploited as an indicator of language development/proficiency?

1.2 Overview

The remainder of this thesis is organized as follows:

- In Chapter 2 we provide literature review on learner corpora, error annotation and linguistic annotation.
- In Chapter 3 we describe VALICO-UD design, describing selection criteria and the principles used to create the parallel normalized version of learner data.
- In Chapter 4 we describe in detail the methodology and error taxonomy used in VALICO-UD. In addition, we provide error statistics and report on three inter annotator agreement experiments.
- In Chapter 5 we describe in detail the linguistic annotation applied to VALICO-UD, reporting on an inter annotator agreement experiment and providing statistics of the manually-corrected section of the treebank. We conclude the chapter reporting on an incremental evaluation of the model used to obtain the first automatically-annotated draft of the resource.
- In Chapter 6 we report on three quantitative metrics used to explore the treebank for assessing the quality of the data and for better understanding the role that this resource can play in the future in the context of computational linguistics.
- We conclude with Chapter 7 where we recap the main points of this thesis, comment on the limitations and discuss future applications of the treebank.

Chapter 2

Literature review

This literature review is divided into three sections. The first presents learner corpora. The second section is devoted to error annotation. Finally, the third section focuses on linguistic annotation.

2.1 Learner corpora

Learner corpora, also called interlanguage (Selinker, 1972) or L2 corpora, are collections of data, produced by foreign or second language learners.¹ Granger, 2002, p. 5, defined them as “electronic collections of authentic FL/SL [Foreign Language/Second Language] textual data according to explicit design criteria for a particular SLA/FLT purpose”. In this definition, two fields making use of learner data are mentioned, Second Language Acquisition (SLA) and Foreign Language Teaching (FLT). Both fields have benefited from learner data and used it for different purposes: in the former, the main objective is to study the process of language acquisition; in the latter, the main purpose is to exploit *authentic* learner data for improving the learning and teaching of foreign languages.

However, the concept of **authenticity** is problematic when associated to language learners. In fact, the recommendation on corpus and text typology (EAGLES, 1996, p. 7), states that 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

¹Although foreign and second language learners are distinguished by linguists and are usually linked to the acquisition/learning dichotomy, in this dissertation we will use L2 as umbrella term encompassing both foreign (i.e. learners of a foreign language receiving formal instruction) and second language (i.e. users of a language which is not their first language but it is acquired in linguistic ‘immersion’ and can involve both formal instruction than subconscious learning) learners, as usually happens in the NLP community.

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 sets the task, not only the one who acquires data.

Corpus-based studies making use of learner data have contributed to the emergence of a new field of study, Learner Corpus Research (LCR). In the last twenty years, LCR has experienced a rapid growth, as can be seen from the number of publications in the field and the growing number of learner resources.² Although the aim of LCR is that of joining SLA and corpus linguistics, this is still not a reality. In fact, to date studies making use of learner corpora have focused more on describing differences between native and non-native language and have been used for pedagogical purposes, more than for addressing SLA research questions (Lozano and Mendikoetxea, 2013, pp. 66–72). According to Tono, 2003, p. 806, this is due to the fact that “SLA researchers typically know little about what corpora can do for them” and thus learner corpora have been mainly used by corpus linguists who “do not know enough about the theoretical background of SLA research”. This has led to LCR studies focused more on describing differences than addressing SLA research topics.

The first projects based on the development and analysis of learner corpora were launched in the nineties and focused mainly on learner English (Tono, 2003). Granger and colleagues of the *Université Catholique de Louvain* are considered the pioneers of learner corpora and LCR, thanks to the development of the first large corpus of learner English, the International Corpus of Learner English (ICLE) (Granger et al., 2002). Then also other projects devoted to the development of large learner corpora have been launched, such as the Cambridge Learner Corpus (CLC)³, used by Cambridge University Press to develop courses and materials (dictionaries included) for learners of English. However, these large learner corpora (million words) still remain an exception.

A recent trend is the emergence of learner corpora projects focused on languages different than learner English, such as learner German FALKO

²Learner 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>.

³The uncoded version is publicly available here: <https://www.sketchengine.eu/cambridge-learner-corpus/>. To know more about it, go to: <https://www.cambridge.org/it/cambridgeenglish>.

(Siemen, Lüdeling, and Müller, 2006), learner Norwegian ASK (Tenfjord, Meurer, and Hofland, 2006), learner Spanish CEDEL2 (Lozano, 2009), learner Czech CzeSL (Hana et al., 2012), learner Portuguese COPLE2 (Mendes et al., 2016), learner Lithuanian LLC (Ruzaitė et al., 2020). In Table 2.1 we provide an overview of learner corpora that we cite in this chapter. Error-annotated corpora are in bold. The information reported in the table, when not available in the cited articles, has been retrieved from the online platforms from which it is possible to query the corpora.⁴

Learner Corpus	Approximate Size (w t)	L2 Level	L1	Mode	Target Language
ICLEv3	5,500,000 w	B2–C2	Various	Written	English
CLC	29,000,000 w	A1–C2	Various	Written	English
NUCLE	1,200,000 t	N/A	N/A	Written	English
FALKOv2	280,000 t	B2–C2	Various	Written	German
MERLIN	340,000 t	A1–C2	Various	Written	Czech German Italian
VALICO	380,000 t	A1–C2	Various	Written	Italian
ASK	1,200,000 t	A2–C1	Various	Written	Norwegian
CEDEL2	280,000 w	A1–C2	English	Written	Spanish
CzeSL	200,000 w	A1–C2	Various	Written	Czech
COPLE2	180,000 t	A1–C1	Various	Written & Spoken	Portuguese
COMIGS	18,000 t	A2–C1	Various	Written	German
COWS-L2H	900,000 t	A1–C1	Various	Written	Spanish
ESL treebank	97,000 t	B2	Various	Written	English
CFL treebank	7,000 t	N/A	N/A	Written	Chinese
SALLE treebank	600 t	A1–C2	Various	Written	English
NOCE	300,000 w	B2–C2	Spanish	Written	English
PIL2	120,000 t	N/A	Various	Written & Spoken	Italian
TOEFL11	4,000,000 t	A1–C2	Various	Written	English
Write&Improve	750,000 t	A1–C2	Various	Written	English
LLC	300,000 t	A1–B2	Various	Written & Spoken	Lithuanian

TABLE 2.1: Learner corpora overview. Error-annotated corpora are in bold.

⁴ASK: <https://clarino.uib.no/korpuskel/overview?session-id=251861814774133>;
CLC: <https://www.latl.leeds.ac.uk/wp-content/uploads/sites/49/2019/07/CLC-demo-final.pdf>; MERLIN: <https://clarin.eurac.edu/repository/xmlui/handle/20.500.12124/6>; COMIGS: <https://nats.gitlab.io/comigs/>.

Learner corpora, in order to be adequate to be used for L2 research, need to be collected observing good design practices. According to their design criteria, learner corpora can be used for a variety of activities and by a variety of users. The main design criteria concerns: learners' sociolinguistic information (e.g. age, sex, proficiency level⁵), task information (e.g. graded examination, non-graded homework), medium (e.g. oral, written, both), genre (e.g. essay, mail, picture-elicited), languages involved (both learners' mother tongue than target languages), time (being synchronic or diachronic, also called cross-sectional or longitudinal, respectively). However, as pointed out by Granger, 2004, p. 126, "one must admit that [...] there are so many variables that influence learner output that one cannot realistically expect ready-made learner corpora to contain all the variables for which one want to control".

Nevertheless, 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 Natural Language Processing (NLP) tools, respectively. This topic is discussed in Section 2.1.1.

Although learner data can be exploited as they are, studies based on raw data must be focused on phenomena that can be easily retrieved (Aijmer, 2002; Nesselhauf, 2004). In fact, learner corpora have more prospects if they are annotated (Díaz-Negrillo and Thompson, 2013, pp. 13–16).

Annotation, which is essential to make explicit what is implicit in texts, usually involves tokenization, lemmatization, Part of Speech (PoS) tagging, syntactic parsing, and semantic tagging (Garside, Leech, and McEnery, 1997; McEnery and Wilson, 2001). In addition, learner corpora can feature also error annotation (discussed in Section 2.2), which allows the use of the methodological approach—that has had an influence in three major areas associated with learner corpus (SLA, FLT and computational linguistics)—called by Granger, 2002, computer-aided error analysis.

Annotation in corpora guides the topics that can be investigated. For instance, since the first version of ICLE was only tokenized and lemmatized, most of the studies based on this version of the corpus focused on lexical

⁵Proficiency level is usually expressed using the Common European Framework of Reference for languages (CEFR) (Little, 2006).

aspects. The addition of morpho-syntactical annotation (i.e. PoS tagging) in the second version (Granger et al., 2009) allowed for studies focused beyond lexis (e.g. Aarts and Granger, 2014). Annotated learner corpora, in fact, can be used for providing a data-based understanding of interlanguage (for second language acquisition researchers), to develop pedagogical tools using real examples and targeting the actual issues of language learners (for foreign language teaching), or to target some NLP tasks (for computational linguistics researchers).

Annotation is only one part of corpus design. In fact, apart from annotated vs. unannotated corpora, we can distinguish cross-sectional vs. longitudinal corpora, monolingual vs. multilingual (for both target and native language) corpora, oral vs. written corpora, corpora with data from more vs. less controlled conditions (Díaz-Negrillo and Thompson, 2013, p. 10).

Most learner corpora are collected during language proficiency examinations, thereby they are usually built in collaboration with language assessment centres, e.g. CLC. Learner corpora can be classified according to the typology of learner products they collect. The majority of them—especially those built in collaboration with assessment centres—collect essays (e.g. ICLE and CLC). Those called peripheral learner corpora, using the term proposed by Nesselhauf, 2004, are elicited by pictures. Being picture-elicited, these corpora contain products that are subject to greater control and can therefore be considered more similar to the data used by SLA researchers (e.g. Ellis, 1995, pp. 103–105). For this reason, they might be a valid resource for testing SLA research questions. In addition, although picture-elicited texts are less authentic (because learners must follow what is drawn in the picture and are not totally free to express themselves), they are more reliable for writing target hypotheses (THs). THs are the normalized versions of Learner Sentences (LSs) and have been considered essential by different scholars in order to have an explicit reference on which to base error annotation (Lüdeling, 2008; Reznicek, Lüdeling, and Hirschmann, 2013; Rosen et al., 2014; Meurers, 2015). However, it is acknowledged that THs are difficult to formulate as it is not always possible to bring incorrect forms into correct ones that are also objective. Nonetheless, thanks to the controlled context imposed by pictures, it is possible to implicitly circumscribe the vocabulary and semantics of the learner sentences (Corino and Mareello, 2009; Mareello, 2011; Köhn and Köhn, 2018) and reconstruct more reliable THs.

2.1.1 Learner corpora and NLP tasks

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 dataset compiled for GEC, called W & I (Write & Improve) + LOCNESS (Bryant et al., 2019) which joins two corpora (one of natives, i.e. LOCNESS, and the other of learners, i.e. Write & Improve). But also non-English corpora, such as German, Portuguese, and Spanish have been recently developed with computational tasks in mind (Köhn and Köhn, 2018; Del Río Gayo, Zampieri, and Malmasi, 2018; Davidson et al., 2020, to name a few). As far as Italian learner corpora are concerned, the only learner corpus that has been used as dataset for NLI is VALICO (Corino and Marello, 2017), the biggest corpus of this kind freely available online.⁶

2.2 Error annotation

Error annotation is a peculiar type of annotation which can be associated to learner corpora, as introduced in the previous section. Error annotation, although questioned by SLA researchers because of its inadequacy for wholly explaining language acquisition, is a valuable resource to provide insights into proficiency stages—as shown by Abe and Tono, 2005 and Tono, 2013—and, in FLT, it can be a real pedagogical tool by itself (Granger, 2009, p. 24). Reusing the words stated by Cook, 1993, p. 22, error analysis is a “methodology for dealing with data, rather than a theory of acquisition”.

As previously introduced, error annotation is exploited via the methodology called Computer-aided Error Analysis (CEA). This methodology is the evolution of the Error Analysis (EA) of the 60s—at that time considered as an acceptable alternative to the Contrastive Analysis (CA) of the 50s (called by James, 1998, as Behaviourism-tainted)—and that was severely criticised in the 70s, due to the diffusion of concepts as *idiosyncratic dialect* (Corder, 1971)—which is the evolution of *transitional competence* (Corder, 1967)—, *interlanguage* (Selinker, 1972), and *approximative system* (Nemser, 1971).

According to Corder, 1971, pp. 152–153, the use of *error*, *deviant*, *ill-formed* or *ungrammatical* are all terms that collide with the status of idiosyncratic

⁶VALICO texts are available here: www.valico.org.

dialect assigned to learners' language (i.e. having a grammar in itself). If the reason for studying learner's language is "to discover why it is as it is", then, for him, these terms are not admissible because they imply explanations before providing a description.

These are not the only criticisms raised against the EA of that time. Bell, 1974, p. 35, called EA as a *pseudoprocedure* and attacked it for lack of predictiveness, high subjectivity and poor statistical inference. Others have also criticised the practice of extracting only the erroneous structures, without considering the correct ones. All these criticisms culminated with the Bley-Vroman's *comparative fallacy*. The comparative fallacy is "the mistake of studying the systematic character of one language by comparing it to another" (Bley-Vroman, 1983, p. 15).

However, nor EA or CA did stop after those criticisms. Without comparisons it would be impossible to perceive "the difference between the learner's internalized description of his [/her] L2 and the internalized descriptions that native speakers have" (Meara, 1984, p. 231). Thus, on the one hand, it is important to keep in mind Bley-Vroman's directive in order to focus on objective and non-derivative descriptions of interlanguage; on the other hand, having a reference still remains necessary, as proved by the studies based on error annotated corpora (Lüdeling et al., 2005; Wisniewski et al., 2013; Köhn and Köhn, 2018; Del Río Gayo and Mendes, 2018b; Davidson et al., 2020) and the number of error-annotated learner corpora, in boldface in Table 2.1.

As Meara, 1984, James, 1998, Granger, 2002 and other scholars have pointed out, comparisons can highlight a range of features that allow us to better understand interlanguage—which is the aim of SLA—and to help learners to improve their proficiency, bringing it closer to target language norms—which is the aim FLT.

Error analysis, in particular, has overcome some of the weaknesses pointed out in the 70s. Within the CEA methodology, indeed, erroneous occurrences of linguistic items can be visualized alongside correct ones, and can be investigated together with the context of use and the co-text (i.e. linguistic context). In addition, in order to deal with subjectivity, error categories are defined and fully documented in guidelines, which are usually provided with the error-annotated corpus (Granger, 2002).

To day, however, error annotation is a practice with a very low interoperability. In fact, despite various attempts of creating a gold standard, all the error-annotated learner corpora have their own error-tagging scheme and

their annotation tools.

2.2.1 Defining *error*

In this dissertation, we endorse Lennon's definition of error: "a linguistic form or combination of forms which, in the same context and under similar conditions of production, would, in all likelihood, not be produced by the speakers' native speaker counterparts" (Lennon, 1991, p. 182).

As far as causes are concerned, errors can be due to grammaticality and correctness, or to acceptability, strangeness and infelicity (James, 1998, pp. 64–76). On the one hand, there is a distinction made on grammatical/ ungrammatical and correct/ incorrect that can be attributed to what it is written in the grammar and what is "influenced by prescriptive notions inculcated at school" (Radford, 1988, p. 12). On the other hand, the distinction is made by usability and probabilistic rules that can be attributed to users, context of discourse and authentic data collections (i.e. corpora). In particular for what concerns the context of discourse, we can talk about covert and overt errors (Corder, 1973, p. 272). Covert errors are those errors which result in well-formed forms but do not match intentions. For example, *la figlia si rivolge al suo salvatore* ('the daughter turns to her savior') in which the learner used *figlia* instead of *donna*, and we can know their intention only thanks to the comic strip that elicited the text and the co-text. Overt errors are those that can be identifiable without the context (e.g. spelling errors resulting in non-words).

Still depending on learners' intention, errors (i.e. the term used in the literature as umbrella term to indicate erroneous non-canonical forms) can be distinguished into **deviations**, **errors** and **mistakes**. Deviations are intentional errors produced by advanced users having specific purposes (e.g. hilarity, poetry, advertisement). By contrast, errors and mistakes are both unintentional errors that can be produced both by advanced and initial users. The difference between errors and mistakes relies in the self-correction. The former are *not* self-correctible. The latter are self-correctible. Corder was the first who introduced this error-mistake dichotomy to the modern debate (Corder, 1967; Corder, 1971), associating it to the Chomsky's competence-performance dichotomy (Chomsky, 1965).

Depending on the extent to which the erroneous forms deviate from native speakers' forms, we can distinguish different levels of **error gravity**. According to James, 1998, pp. 206–234, error gravity should be based on linguistic criteria, and in particular on grammaticality. Excluding the irritation factor and noticeability—which in our opinion are too much in the eye of the beholder—the other criteria discussed by James are: rule infringement (the higher the feature violated, the more serious the error), rule generality (larger the number of instances the rule applies to, the more serious the error), frequency (the higher the frequency, the more serious the error), comprehensibility (the more unintelligible it is, the more serious the error). The studies measuring error gravity perception usually use a 5-point scale and deal with differences between native and non-native teachers' judgments in rating different error types (Schmitt, 1977; Lalande, 1981; Vann, Meyer, and Lorenz, 1984; Lennon, 1991; Schmitt, 1993; Salem, 2007, to mention a few). Our hierarchy follows in part the results reported in Vann, Meyer, and Lorenz, 1984 in which spelling errors resulted to be the most tolerated whilst word order the least tolerated by the academic community. The hierarchy followed in VALICO-UD is described in depth in Section 3.2.

2.2.2 Describing errors

In order to describe errors, it is necessary to identify them. The task of identifying an error consists of only be aware that an error is present.

Experiments about error identification carried out in the literature of applied linguistics (at sentence level) underline how this task is not straightforward as one might think, both for (non-)native teachers of English than native non-teachers (Hughes and Lascaratou, 1982). Recently, error identification has been carried out at token level (Dahlmeier, Ng, and Wu, 2013; Rosen et al., 2014; Köhn and Köhn, 2018; Boyd, 2018; Del Río Gayo and Mendes, 2018b), as in Grammatical Error Identification task.

When moving from sentence to single tokens, we move from mere identification (or *detection* using the term by James, 1998) to location of errors. As a result, further difficulties are added. One concerns those errors called *global* errors (Burt and Kiparsky, 1972), that is errors diffused throughout the sentence or located at textual level. In addition, locating errors can be problematic for, at least, other three aspects:

1. Do we consider as error what the learner wrote or what they should have written?
2. Do we adopt an explanatory or a descriptive approach in analysing errors?
3. What can be inserted in the error tag? Should errors caused by the correction of other errors be counted as learner's errors?

To answer to question 1, let us confront the two most used and known learner English corpora, ICLE and CLC. Depending on the error-annotated corpus, the error is encoded following the wrong word class or the corrected one. In fact, the first principle reported in the ICLE Error tagging manual (Dagneaux et al., 1996, p. 5) states: "Do not tag on the basis of the corrected/targeted word/phrase, but on the basis of the incorrect word/phrase" with the exception of the Saxon genitive. On the contrary, the CLC error system annotates the "*word class of the required word*" (Nicholls, 2003, p. 573), hence exactly the opposite of ICLE. In Example 1 we report an example used in the ICLE Error tagging manual to explain their first annotation principle, whilst in 2 we annotate the same example using the CLC annotation rules.

- (1) **ICLE:** The main feature of a campus like Louvain-la-Neuve is (GA) the \$its\$ conviviality.
- (2) **CLC:** The main feature of a campus like Louvain-la-Neuve is <#RP>the|its</#RP> conviviality.

In Example 1 the substitution of the article with a personal pronoun is labeled as a grammatical error involving an article (i.e. the wrong form), whilst in Example 2, the same error is tagged as a replacement error affecting a pronoun (i.e. the correct form).⁷ The examples reported above, hence, show differences not only in what should be considered as error but also how the errors are marked (using round brackets and dollar signs in 1, and angle brackets and pipes in 2).

Question 2 is closely related to the type of **taxonomy** used. A taxonomy classifies errors according to specific constitutive criteria. Depending on the criteria used in the classification, we can distinguish four types of

⁷Note that we are using P to refer to pronouns only for convenience and clarity. The actual tag for pronouns in the CLC is A.

taxonomies: linguistic categories, surface strategy (also called target modification), comparative analysis, and communicative effect taxonomies (Dulay, Burt, and Krashen, 1982, pp. 146–197). A fifth type of taxonomy, was first suggested by James, 1998, p. 114, and consists in a combination of the linguistic category and surface strategy taxonomies. The five above mentioned taxonomies can be divided in two types of taxonomies: descriptive (i.e. linguistic category, surface strategy taxonomies, and their combination) and explanatory taxonomies (i.e. comparative and communicative effect taxonomies). As the name suggests, a descriptive taxonomy aims at describing the errors grouping together those sharing some features (which can be linguistic categories or target alterations). Explanatory taxonomies aim at finding causes (comparative taxonomy) or a hierarchy of gravity (communicative effect taxonomy). In this dissertation we only talk about descriptive taxonomies. On the one hand, linguistic category taxonomies classify errors according to the linguistic level they affect, e.g. phonology/orthography, grammar, lexis, text or discourse. They are then identified by the word class and can be further specified by the grammatical category involved (Simone, 2008, pp. 303–346). One corpus that adopted this type of taxonomy in its error annotation scheme is ICLE (Dagneaux et al., 1996). On the other hand, surface strategy taxonomies describe errors according to the ways surface structures are altered (omission, addition, misordering, misformation). The CLC error system is an example (Nicholls, 2003).⁸

However, it is worth noticing that choosing one taxonomy instead of another not only puts the attention on different constitutive criteria that define errors—returning a different image of interlanguage—, but also can influence error span and error counting. Let us consider one example drawn from the ICLE annotation guidelines (Dagneaux et al., 1996, p. 18), reported in Example 3, and its annotation applying the error-tagging system of CLC, as shown in Example 4.

- (3) **ICLE:** Students have the
(XNCO) possibility to leave \$possibility of leaving\$

⁸Note that both ICLE and CLC error systems are not completely linguistic category and surface strategy taxonomies, respectively. In fact, the first describes the errors using seven main categories, and one of the categories contains three tags that follow the surface strategy (i.e. word redundant, word missing and word order). The second, mainly follow the surface strategy but tags are always made of a first letter indicating the alteration and a second letter indicating the word class of the correct word. In addition, it features tags that refer to specific grammatical errors, e.g. CN for Countability of Noun error. Furthermore, both error taxonomies employed a tag to indicate false friends, which is more an explanatory than a descriptive tag.

(4) CLC: Students have the possibility

<#FV><#RT>to | of</#RT>leave | leaving</#FV>

In Example 3 it is marked one error (i.e. XNCO). In Example 4, in the same sentence, two errors are marked (i.e. RT and FV). Thus, depending on the error-tagging system considered, the same learner sentence can be tagged with one or more errors, if we consider different error taxonomies. In addition, while this sentence is marked in ICLE with a tag indicating erroneous complementation of nouns (i.e. XNCO), the CLC highlights the differences in the learner forms and target hypothesis, resulting in the replacement of a preposition (i.e. RT) and the selection of a wrong verb form (i.e. FV). Finally, the error span in ICLE involves three tokens (i.e. *possibility to leave*) but in the tag only the noun part of speech is marked with N; differently, in CLC it involves two tokens (i.e. *to leave*), and both parts of speech are marked in the tags (i.e. preposition T and verb V). These examples therefore represent well how the use of different taxonomies can lead to different descriptions of interlanguage.

Question 3 can only be answered if two concepts are defined: **scope** and **substance**. As defined in Quirk et al., 1985, pp. 85–86, “SCOPE [...] describes the semantic ‘influence’ which such words have on neighbouring parts of a sentence”. Dobrić and Sigott, 2014, pp. 114–117, collocate scope in the context of error analysis and defying it as “the amount of textual or extratextual context that is required for recognising the presence of an error” and accompany it with the concept of substance that “refers to the smallest constituent in the learner production that needs to be modified so that the error will disappear”. Some error systems insert in the tag the error scope (e.g. in Example 3 *possibility* does not need edits but is part of the error scope), others only the error substance (e.g. in Example 4 only *to* and *leave* are marked as errors). As a consequence, a follow-up question might be asked: are *cascade* errors learner’s errors? Cascade errors—called *errori a cascata* by Andorno and Rastelli, 2009, p. 52, and being in a way similar to the *snowballing effect* by Stemberger, 1982, p. 325—are errors within errors or errors that are caused by the correction of another error. Although cascade errors are discussed in the literature, to our knowledge, no error-tagging system has annotated them.

2.2.3 Error-tagging systems

As mentioned, each project usually has its type of error annotation and its tool. Error annotation can vary along three major axes: target hypothesis explicitness, annotation scope and annotation technique. The first differentiates annotation with explicit target hypotheses from those without them. The second spans from general error annotation to error-specific annotation (i.e. all error types are annotated or only a specific error type is marked, respectively). The third depends on the tools used and concerns the space in which annotations are inserted: standoff annotation (i.e. annotations are inserted in additional tiers aligned with the text), e.g. FALKO, or inline annotation (i.e. usually making use of XML-like tags) e.g. ICLE or CLC.

2.2.3.1 Target hypothesis explicitness

Most error-annotated learner corpora mark only the deviant form(s) with an error tag without providing a correct version (i.e. TH) of it (i.e. leaving the TH implicit). Explicit THs can be distinguished in partial or full THs. Partial THs are those that consist only of correct forms, which are usually inserted in a specific part of the tag (e.g. between dollar signs in Example 3). Or they can be full THs, such as those usually reported in standoff annotations, in which THs have a tier of their own (as shown in Example 5 taken from Reznicek et al., 2010, p. 40). In this way, LSs are separated from THs, as recommended by many scholars (Lüdeling, 2008; Reznicek, Lüdeling, and Hirschmann, 2013; Meurers, 2015).

(5)

ctok	Man	hat		ihr	es		geglaubt	.
ZH1	Man	hat	es	ihr			geglaubt	.

THs can vary also depending on what types of errors are corrected. In this regard, form-based and meaning-based THs can be distinguished. In the former only grammatical errors are corrected, ignoring the context and the learner's intended meaning. In the latter also the context and the intended meaning are taken into account, exploiting information from the task. Indeed, meaning-based THs depend more on annotators' interpretation. Therefore, the more the context of learner texts is circumscribed, the higher the agreement between annotators writing the THs (Meurers, 2015, pp. 538–543). Reznicek, Lüdeling, and Hirschmann, 2013—FALKO's authors—suggest that

two THs per sentence should be annotated: one form-based and another meaning-based correcting also writing style issues. To date, only COMIGS (Köhn and Köhn, 2018) and MERLIN (Boyd et al., 2014) have followed the suggestion annotating also a second TH. As far as Italian is concerned, in MERLIN this second TH is available only for 19% of the texts. Furthermore, correcting lexical and semantic errors only in a second TH, in our opinion, is not efficient to be used in NLP tasks like GEC, in which these types of errors are usually corrected. Please see Section 3.2 for the choices made in writing the THs in VALICO-UD.

2.2.3.2 Error annotation scope

Error annotated learner corpora can feature general error annotation (e.g. ICLE, CLC) or specific error annotation (e.g. Tetreault and Chodorow, 2008a; Davidson et al., 2020). Error annotation is typically a costly labour (in terms of time and money), so sometimes, in order to have large-scale corpus data annotated, only some errors are annotated throughout the corpus. For example in the corpus of L2 Spanish COWS-L2H (Davidson et al., 2020), consisting in 892,023 tokens, only three errors have been annotated throughout the whole corpus: number and gender agreement and prepositional accusative.

2.2.3.3 Error annotation technique

With error annotation technique we refer to its implementation. In the last two decades, error annotation implementation has evolved considerably. From simple pasting or typing error tags in the learners' sentences to XML-based and multi-layered standoff formats. However, it is worth noticing that inline error tags are not bounded only to first attempts, but simply at the beginning of error annotation more sophisticated XML and multi-layer formats were not implemented. In fact, ICLE and a more recent corpus COWS-L2H share a similar error-annotation scheme in the sense that they both are inline tags pasted or typed directly on the learners' texts. In COWS-L2H, however, round brackets and dollar signs are not used, instead square brackets contain the erroneous form(s), curly brackets the TH and angle brackets the error tag. In Example 6 we report an example taken from Davidson et al., 2020, p. 7240, in which the COWS-L2H error-tagging system is exemplified.

(6) **LS:** Yo vivo en el ciudad.

I live in the city.

Error-tagged: Yo vivo en [e1]{1a}<ga:fm:art>ciudad.

As stated in the survey on error tagging systems by Díaz-Negrillo and Domínguez, 2006, p. 87, error-tagged corpora not always make accessible information about their error systems, for this reason they surveyed only the best documented and representative systems. For a comprehensive review about error-tagging systems and their tools please see also Lüdeling and Hirschmann, 2015.

Error annotation in the core section of VALICO-UD consists in a XML-based general error annotation performed by a native speaker using a in-house tool, originally developed for transcribing VALICO texts and metadata. In VALICO-UD, we adapted the error-tagging scheme of CLC as reported in Nicholls, 2003. We selected this kind of error-tagging system because it is, to our opinion, the most adaptable to be used for languages different to the one for which it was developed. The adaptation consisted in adding a third level that specifies the grammatical category involved (Simone, 2008, pp. 303–346), hybridising even more the surface edit taxonomy with the linguistic category one. The annotation scheme is detailed in Chapter 4.

2.2.4 Inter annotator agreement

Inter-annotator agreement (IAA) is a common practice in Computational Linguistics and NLP for comparing if two (or more) annotators make the same decision annotating the same product (e.g. text, audio). The reasons behind this practice are manifold. IAA can be used to validating and improving annotation schemes and guidelines. IAA can help in identifying ambiguities, difficulties or bias. These issues can be due to annotators, product or the task itself. For instance, annotating learner texts, a challenge is indeed how to deal with non-standardized forms. Annotation performed by multiple annotators can show the range of valid interpretations of the same non-canonical forms, but also highlight unforeseen (or confirm hypotheses about) problematic areas of the task, or reveal annotator bias (see Artstein, 2017 to know more about inter-annotator agreement for linguistic annotation and Hovy and Prabhumoye, 2021 to know more about five different sources of bias in NLP).

The first error-annotated learner corpora were usually tagged by one coder and revised by another (e.g. CLC), thus they did not report IAA studies. This issue was first raised by Meurers and Müller, 2009 who accounted for an almost total lack of studies reporting IAA on error annotation. Since then, a

number of studies have started to pay attention to this issue (Rozovskaya and Roth, 2010; Boyd, 2012; Lee, Dickinson, and Israel, 2012; Dahlmeier, Ng, and Wu, 2013; Rosen et al., 2014; Köhn and Köhn, 2018; Boyd, 2018; Del Río Gayo and Mendes, 2018b).

One of the issues in reporting IAA concerns the decision of the best-suited measure for the particular kind of task. This issue is reported in almost all the above mentioned studies. A thorough survey of methods by Artstein and Poesio, 2008 suggests the use of Krippendorff’s α when dealing with tasks in which category labels are not equally distinct from one another, such as hierarchical tagsets and set-value interpretations. They also attest the use of Cohen’s κ and Krippendorff’s α in the vast majority of studies they reported and restate their appropriateness, as they abstract away from the bias of specific annotators. However, they suggest that to avoid annotator bias, increasing the number of annotators is the best strategy. To date, Cohen’s κ and Krippendorff’s α are the most used measures, and κ in particular is the most used for IAA in learner corpus field.

In this section we report on the IAA experiments by Dahlmeier, Ng, and Wu, 2013; Köhn and Köhn, 2018; Boyd, 2018 and Del Río Gayo and Mendes, 2018b because are somewhat comparable to the experiments carried out for this dissertation reported in Section 4.4.

Dahlmeier, Ng, and Wu, 2013 in their study, making use of the NUCLE corpus (see Table 2.1), measure IAA under three different conditions in sequence: **identification** of the error, **tag choice**, and **exact match**. Tag choice is measured only when the annotators agree on the identification, whilst exact match considers tag choice and correction. They selected 96 essays, not included in the final version of NUCLE, and had three annotators to code them in a way that each essay was annotated by two annotators. Since they did not instruct their annotators about how to deal with missing errors (i.e. using web-based tools, for example, it is not possible to select a white space, thus these errors must be encoded with the previous or with the following token), or on deciding the minimal portion of text that should be considered when selecting a text span as error (i.e. annotators had the chance to select also characters not only at token level, thus an error involving a wrong tense, e.g. *use* corrected into *used* can be corrected at token level or selecting only a part of it, such as *e* corrected into *ed*), they had to perform text processing before comparing the two annotations (Dahlmeier, Ng, and Wu, 2013, pp. 25–26). They report a κ of 0.39 for identification, 0.55 for tag choice and 0.48 for

exact match, that in Landis and Koch terms (Landis and Koch, 1977) can be considered fair (identification) and moderate (tag choice and exact match) agreement.

Köhn and Köhn, 2018 used a picture-elicited corpus of learner German in which two THs are annotated (as in Reznicek, Lüdeling, and Hirschmann, 2013, see Section 2.2.3.1). In their paper, they reported a *kappa* of 0.79 on error identification and of 0.64 when considering the correction.⁹

Using a learner German corpus too, in particular a reading comprehension corpus, Boyd, 2018 reported on the meaning-based TH error identification task a *kappa* of 0.68. In cases in which the annotators agreed in the identification, 70% of the time annotators agreed also about the correction.

Del Río Gayo and Mendes, 2018b measured the IAA obtained in the annotation of errors in a learner Portuguese corpus (COPLE2). They evaluated two tag sets in two different samples. In the first, token-based, only errors affecting single tokens have to be corrected and classified as orthographical, grammatical or lexical. In the second, a fine-grained tag set is tested and the correction is also requested. The achieved IAA on the token-based is *kappa* = 0.86 (*kappa* = 0.85 if considering also the correction) and that on the fine-grained tag set a *kappa* of 0.85 (0.84 with the correction).

As far as error identification and error types are concerned Previous studies reporting disagreement on error identification and error types reported a higher agreement for orthographic and grammatical errors (Del Río Gayo and Mendes, 2018b 0.96, 0.93 respectively), lower for lexical errors (Del Río Gayo and Mendes, 2018b 0.70). Also Rosen et al., 2014 reported low agreement for lexical and usage errors. They reported higher agreement for incorrect morphology, improper word boundaries and foreign expressions ($\kappa > 0.80$, $\kappa > 0.60$, and $\kappa > 0.40$, respectively). Lower agreement involved categories for which a target hypothesis was difficult to establish. A fair agreement was achieved for agreement errors, and syntactic dependency errors. For some other errors identifiable by formal linguistic criteria, they reported very low IAA and attributed this to unclear guidelines.

⁹Note that the *kappa* values reported here are the mean of the results obtained in their two THs, in order to be comparable to our single TH (see Section 3.2). This is motivated by the fact that lexical errors—which usually trigger disagreement (Rosen et al., 2014; Del Río Gayo and Mendes, 2018b)—in their annotation scheme are corrected in the second TH. However, it is worth noticing that our TH do not correct stylistic errors that are instead considered in their second TH.

2.2.5 Implicit error annotation

Here, we call implicit error annotation those projects that explicitly do not mark errors, but they implicitly do it marking divergent behaviour at different levels of evidence. For example, PIL2, SALLE and parallel treebanks allow the retrieval of information about learners' errors without explicitly marking them as such. That is to say that, when PIL2 annotates information about PoS at *source* level that conflicts at *tendential* level (see Section 2.3.1.1), this divergence can be considered as an implicit version of error annotation. The same can be argued for SALLE, not only at PoS annotation (see Section 2.3.1.3), but also at syntactic annotation (see Section 2.3.2.1).

Furthermore, THs can also form a parallel corpus together with the corpus composed of LSs—and being both annotated syntactically—can enable the retrieval of specific learner errors, such as word order errors (Lee, Li, and Leung, 2017) or overused and underused syntactic structures (Li and Lee, 2018).

Notwithstanding the error annotation types, to smoothly retrieve interlanguage features, error-tagged learner corpora may not be sufficient and other linguistic annotation might be needed. In the next paragraph, we will talk about linguistic annotation.

2.3 Linguistic annotation

With linguistic annotation we refer to different levels of linguistic analysis, i.e. lemmatization, PoS tagging, morphological feature annotation, and syntactic annotation. In this section we survey how these analyses have been associated to learner corpora. Although segmentation and tokenization, two processes usually carried out in the preprocessing phase in order to perform linguistic annotation, are not usually discussed,¹⁰ in Section 5.1.2 we describe their challenges too. Here we review the annotation choices made in other learner annotation projects about PoS tagging, lemmatization and the annotation of morphological features. We conclude the chapter surveying syntactic annotation focusing on learner corpora.

¹⁰Few studies dealing with learner language (L1 or L2) mention the issues of word or sentence segmentation, e.g. Brunato and Dell'Orletta, 2016 and Berzak et al., 2016.

2.3.1 PoS tagging, lemmatization and morphological feature annotation

The second most commonly annotated feature on learner corpora is PoS tagging (error annotation being the first). When PoS tagging a corpus, also lemmatization and other morphology features have to be taken into account. The PoS tags used vary from learner corpus to learner corpus.

However, this is not the only difference that we can find in PoS annotated corpora. In fact, this is only the most superficial difference. An important distinction between PoS-annotated corpora concerns the automatizing of the process. On the one hand, it is possible to run pre-trained taggers on learner texts, i.e. treating the task as domain transfer. Domain transfer indicates the running of automatic taggers trained on particular text genres (i.e. native language of a particular type) on corpora of texts from different genres (i.e. learner language of a different type than the training). This strategy to automatically annotate learner corpora has been adopted in e.g. ICLE and MERLIN (Boyd et al., 2014). On the other hand, manually PoS-annotated corpora, although they use the target language's PoS tags, are characterized by attempts of developing multi-facet *ad hoc* annotation. This is the case of PIL2 (Andorno and Rastelli, 2009), NOCE corpus (Díaz-Negrillo et al., 2010), SALLE (Ragheb and Dickinson, 2014a), ESL (Berzak et al., 2016), and CFL (Lee, Leung, and Li, 2017), to cite a few.

The strategies adopted in the above mentioned projects for manually annotating PoS tags can be divided in two groups: those that annotate more than one PoS tag in order to highlight the non-canonical structures of learner language, and those who annotate one PoS tag per token. PIL2, NOCE and SALLE belong to the former, ESL and CFL to the latter.

2.3.1.1 PIL2

The annotation proposed by Andorno and Rastelli, 2009 is based on the annotation of two layers: one called *source* and the other called *tendenziale* ('tendential'). The former refers to the features of the *forms* actually used by the learner, whilst the latter indicates the features of the *lexemes* to which the forms can be attributed. As explained by the authors, they called *tendential* the second annotation layer because this annotation points towards lexical features pertaining to the target language (Andorno and Rastelli, 2009, p. 60), that are not related to the learner's interlanguage. Hence, even though the

authors affirm that their double annotation is not comparable to studies in which the double annotation is based on the interlanguage form and on the target language features, what they actually do when annotating the tendential properties is to annotate the features of the target language. The authors try to annotate both source than tendential features considering only the target language features of the decontextualized word.¹¹ The narrow context, however, is used in cases of ambiguity. When dealing with non-existent words, they use the most likely target word, which is chosen using as selecting criterion the similarity to the learner's form. If the non-existent word is not straightforwardly attributable to one target word, underspecified labels are selected instead. They exemplify it through the examples *dormiscono* (wrongly inflected verb of *dormire* maintaining the information of third person plural in the present tense of the indicative mood) that they attribute to *dormono* ('they-sleep') and *acomisati*, labeled as masculine plural past participle of a not known verb lemma (Andorno and Rastelli, 2009, pp. 57–58). In VALICO-UD too we followed the similarity principle to redirect unknown forms to TH forms. Differently from Andorno and Rastelli, 2009, we annotated the lemma also for words similar to *acomisati*, which in VALICO-UD would have had the lemma *acomisare*, a non existing verb of the first conjugation that follows the morphology features displayed in *acomisati* (see Section 5.1.2.3).

2.3.1.2 NOCE

Whilst the annotation of PIL2 is bipartite, the one based on NOCE corpus is tripartite. They use three types of evidence: lexical information, morphological information and distribution to disambiguate and assign a PoS tag. The three types of evidence interact: if a token can be unambiguously referable to a PoS looking at a lexicon, then that PoS is selected (the information selected in the *tendential* annotation of PIL2). If non-existent words occur, but morphological clues can provide unambiguous PoS information, these are used to assign the PoS (the information called *source* in PIL2). Finally, distributional information is used to annotate cases in which a word commonly used with a specific PoS is instead used with another PoS (used in PIL2 only for ambiguous forms). The three types of evidence are proposed to

¹¹Note that tendential features can be assigned only to content word that have defined features in the target language.

be annotated in three different layers so to describe the mismatches typical of interlanguage.

SALLE, ESL and CFL not only annotate PoS tags but also dependency relations. In what follows we describe their annotation strategy for PoS tags. Please refer to Section 2.3.2 for the discussion about syntactic annotation.

2.3.1.3 SALLE

SALLE shares with PIL2 the bipartite annotation and with NOCE two of its three layers of information. In SALLE morphological and distributional PoS tags are annotated and the two layers of information coincide unless non-canonical forms appear, i.e. *makes* and *sound* in Example 7, drawn from Ragheb and Dickinson, 2011, p. 118. For what concerns *makes*, the morphological PoS indicates the third person singular verb in the present tense, while the distributional PoS indicates that it is located in a base form verb position. As *sound* is concerned, the morphological PoS indicates that it is a singular noun, while the distributional PoS underspecifies the information about number, since a plural noun would be needed in that position.

(7)

Tin Toy	can	makes	different	music	sound	
proper noun	modal	vb.3rd.sg	adjective	sg.noun	sg.noun	(PoS_m)
proper noun	modal	vb base	adjective	sg.noun	noun	(PoS_d)

In their paper, titled “Avoiding the Comparative Fallacy in the Annotation of Learner Corpora”, they refuse the approaches using error annotation and target hypotheses, guided from the “desire to annotate linguistic properties in a way which avoids the comparative fallacy”. However, the comparison—especially with the target language—is unavoidable, and they use it in order to annotate both morphological and distributional evidences. In addition, the comparison between PoS_m and PoS_d results in a different kind of error annotation, as better explained in 2.2.5. What matters in learner annotation is, in our opinion, to be consistent in the choices and to use the standard annotation rules of the learners’ target language so as to limit the annotator’s interpretation as much as possible.

2.3.1.4 ESL and CFL

Abandoning the pretension of SALLE’s authors, ESL and CFL—annotating only one PoS tag for learner language—take from them the principle that guides their annotation, i.e. the *literal reading* (Dickinson and Ragheb, 2009; Ragheb and Dickinson, 2012; Dickinson and Ragheb, 2013). When annotating PoS, ESL adheres as much as possible to the observed morphological forms of the words *or* to their distribution (e.g. in presence of adjectives having a plural suffix they mark them as adjectives, because the unnecessary agreement of number is marked in the error annotation). CFL in these cases associates the token with the correct (or existent) lemma and bases the PoS tag on it. This choice is probably due to the fact that, while ESL has released learner sentences and the corrected target hypotheses as a parallel treebank, CFL did not. For this reason their choices are different when tackling spelling errors. In ESL lemmas are not annotated, instead. In our opinion, the strategy adopted by ESL is the best one because it exploits parallel corrected version to extract the non-canonical structures and do not fall into the vicious circle of wanting to annotate the interlanguage avoiding the comparative fallacy, but still using the categories of the target language. However, in terms of what can be defined corpus (Sinclair, 2005), perhaps ESL treebank would not be recognized as such, being a selection of random sentences drawn from FCE (Yannakoudakis, Briscoe, and Medlock, 2011), a subcorpus of CLC consisting of First Certificate of English scripts.

It is worth noticing that the PoS tags used in SALLE are drawn from the SUSANNE tagset (Sampson, 1995), while ESL and CFL used the Universal Dependencies tagset.¹² The main differences between the two label sets are the number of tags they contain and the possibility for crosslingual studies. On the one hand, SUSANNE tagset is bigger, but encodes in the PoS labels also information that in Universal Dependencies is encoded separately (see Section 5.1.1 to know more about Universal Dependencies formalism). On the other hand, Universal Dependencies supplies a unified annotation framework for multiple languages and is targeted towards multilingual NLP. Hence, learner treebanks annotated using this formalism support computational analysis of learner language using not only target language-based but also multilingual approaches which seek to relate interlanguage phenomena

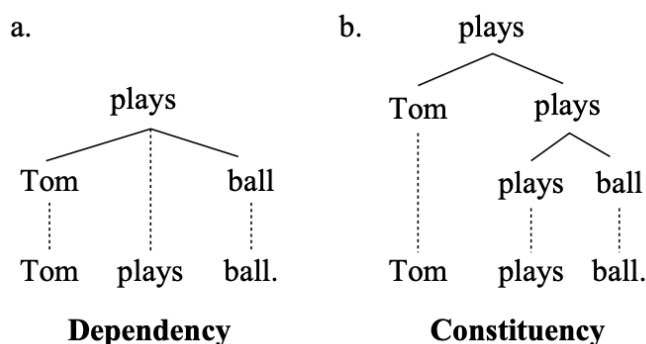
¹²Universal Dependencies Tagset available here: <https://universaldependencies.org/u/pos/index.html>.

to learners' L1 (and L2s) syntax. This is an important characteristic that contributed to the choice of this formalism also for treebanking VALICO (see Section 5.1).

2.3.2 Syntactic annotation

As a matter of principle, syntax can be represented mainly in two ways: constituency and dependency representations. Constituency or phrase structure representations are based on context-free grammar (Chomsky, 1956; Chomsky, 1965) and display how each building block is organized in the sentence, formally ordering from single words into phrases, clauses and, eventually, sentences. Dependency representations, instead, are based on dependency grammar, a theory introduced by Tesnière, 1959. In Example 8, drawn from Osborne, 2014, p. 605, the same sentence *Tom plays ball* is represented according to dependency (8.a) and constituency (8.b) tree representations. For both trees, the words are used to label the tree structures, as a convention. On the one hand, in the dependency tree, the words are linked directly to each other (e.g. *plays* and *ball*) following a binary, asymmetrical relation between head and dependent. On the other hand, in the constituency tree, the same relations (e.g. the one linking *plays* and *ball*) are mediated by higher nodes (e.g. *plays*), following a part-whole relation.

(8)



Since the theory by Tesnière, 1959, several frameworks of dependency grammar have been developed, such as Functional Generative Description (Sgall et al., 1986) and Word Grammar (Hudson, 1984).¹³ These different frameworks can be distinguished depending on what is considered as head, e.g. some frameworks consider the determiner as head of a noun phrase

¹³To know more about syntax see, e.g., Carnie, Siddiqi, and Sato, 2014.

(Hudson, 1984), others the content word (such as Universal Dependencies, see Section 5.1.1).

Notwithstanding the head-dependent different choices, contrarily to constituency annotation—which make a large use of non-terminal symbols—dependency annotation allow “words to constrain the learning and parsing process successfully” (Bunt, Merlo, and Nivre, 2010, p. 2). Dependency-based representations have been increasing their popularity (Bunt, Merlo, and Nivre, 2010, p.3) thanks to the fact that dependency-based structures are perceived as better suited for free or flexible word order languages (e.g. Italian) and for the promising results that models using features based on dependency annotation have obtained in many NLP tasks (e.g. in machine translation and information extraction). These features—particularly appreciated in the field of computational linguistics—have resulted in a growing interest in dependency grammar. As a consequence, in the last six years, a new framework based on dependency annotation, Universal Dependencies (see Section 5.1.1), has established itself as the *de facto* standard. To date, it contains nearly 200 treebanks, which are the result of the efforts of over 300 contributors from all over the world. Among these treebanks, two are the above mentioned ESL and CFL treebanks. Since May 2021, also the the core section of VALICO-UD (see Section 3.1.1) has been published in this repository.

Similarly to what happens for PoS-tagged learner corpora, syntactically-annotated learner corpora can be parsed manually or automatically, or in a mixed fashion (automatic annotation followed by manual revision). When automatically parsed, usually models trained on native language are used due to the lack of learner language gold standard (e.g. MERLIN with no manual revision; see Astaneh and Frontini, 2009; Corino and Russo, 2016 for some preliminary discussion about the issues in parsing learner corpora).

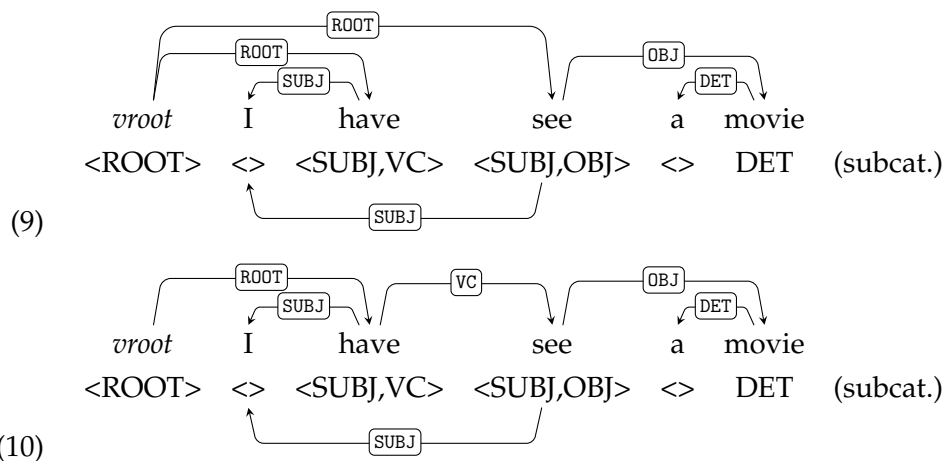
When manually parsed, scholars proposed to adapt native schemes to annotate learner language (Ragheb, 2014; Berzak et al., 2016; Lee, Leung, and Li, 2017). All the projects devoted to the treebanking of learner language make use of dependency annotation and try to annotate it as such, placing “a greater emphasis on word order, or positional information for determining grammatical relations” (Dickinson and Ragheb, 2009, p. 63).

In what follows we describe the methodologies adopted in three learner treebanks, i.e. SALLE, ESL and CFL. The major difference relies in providing one or two annotations per each sentence. In SALLE two dependency

trees guided by morphological and distributional evidences are annotated. In ESL and CFL these two kinds of evidence are used to produce only one dependency tree. Another difference concerns the annotation scheme used. The authors of SALLE used the CHILDES label set, originally developed for L1 English acquisition, adapting it to L2 English. The authors of ESL and CFL adapted the Universal Dependencies scheme to L2 English and Chinese, respectively.

2.3.2.1 SALLE

SALLE's authors set themselves an ambitious goal, i.e. "to be able to annotate any level of learner from any native language (L1) for any type of text" without considering "the context in which something was written" and avoiding learners' intended meaning (Ragheb and Dickinson, 2014a, p. 293). To do so, they implemented a double-layer scheme, one annotating the surface form of the tokens considering the morphological PoS tags, and the other considering the distributional ones. In addition they also annotate the dependencies expected for each token (defined by the authors as *subcaterization*, Ragheb, 2014, pp. 60–62). In Examples 9 and 10, we report the morphological and distributional dependency trees, respectively, as shown in Ragheb, 2014, p. 56.



As it can be noted from the examples, their annotation scheme allow for more than one root per sentence (Example 9). The major differences between the two trees lie in the double roots in 9 and in considering, in 10, *see* as verbal complement of *have*. Another particularity of the scheme concerns the double annotation of the subj relation, even though *see* is annotated distributionally as a past participle.

For what concerns the dependency relation labels, SALLE's authors chose the CHILDES set (Sagae et al., 2010) because it has been developed for language acquisition (L1 acquisition), "thus making distinctions that are relevant to learner language as well, such as the use of INCROOT [i.e. incorrect root] for sentences that do not have finite verbs as a head" (Ragheb, 2014, pp. 90 and 243). However, we believe that even the relation they use for supporting the decision towards the CHILDES label set counters with their aim of annotating interlanguage without falling in the comparative fallacy, because it implies that interlanguage must follow the target language rules when selecting the root of a sentence, else it is described as incorrect. To state clearly our point, we believe, together with other scholars (Tenfjord, Hagen, and Johansen, 2006; Rosén and De Smedt, 2010), that it is unrealistic to describe learner language without making references to the target language, as SALLE's authors try to do. In addition, we want to stress the importance of considering the context in which something was written and also other L2s (if any) known by the learners. And, in any case, it is also necessary to remain consistent throughout the annotation process.

2.3.2.2 ESL and CFL

ESL's authors, decided to use the inventory defined by the English UD formalism and opted for one tree annotating the learner sentence as it is, and another one to represent its correct version. In their paper, Berzak et al., 2016 describe the way they applied literal annotation, in particular to non-canonical English sentences. They rely to literal annotation when the argument structure of a verb is altered by the omission or the presence of a preposition (e.g. *to give to him*, thus annotating *him* as oblique argument and not indirect object), or a word is used in a non-canonical form (e.g. *necessaryiest* annotated as superlative). On the other hand, spelling errors make an exception to literal annotation (e.g. *we where invited to visit*, in which *where* is annotated as auxiliary and not as an adverb) together with wrong word formations (e.g. *they do not sale them*, in which *sale* is not treated as a noun but as a verb).

Similarly, CFL's authors, dealing with learner Chinese, followed basically what done for ESL, but having more difficulties, e.g. Chinese does not mark word boundaries; spelling confusion errors—such as *were* confused for *where*—include also words with different tones; their corpus had not had already annotated error tags.

2.3.2.3 Other approaches

A different choice was made by the authors of FALKO and by Rosén and De Smedt, 2010, who, instead, decided to annotate syntactically only the correct version of the learner language in order to avoid the problem of annotating non-canonical structures but enabling some searches that are not possible in a corpus tagged only at word level.

Chapter 3

VALICO-UD design

This chapter is divided into two sections. The first describes the data composing VALICO-UD. The second deals with the creation of target hypotheses, i.e. the normalized version of each learner sentence.

3.1 Data description

Inspired by existing resources developed for other learner languages and the wide literature about them, VALICO-UD has been designed as a novel treebank to be exploited for investigating Italian learner language from several different perspectives. In this section, we describe the typology of texts collected in the treebank and provide basic statistics about its composition.

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 and downloadable. Second, because it is a collection of non-native Italian texts elicited by comic strips, hence it is more reliable the reconstruction of THs when non-canonical words or structures occur, because lexical choices and semantic frames are circumscribed to the comic strip (Corino and Marello, 2009; Marello, 2011).¹ And third, because it collects a wide variety of metadata, hence enabling the creation of subcorpora following precise design criteria.

VALICO texts can be written by the learners directly using a computer, or can be transcribed manually by transcribers (usually students at the Department of Foreign Languages in Turin). On the dedicated website it is possible to download texts in doc format.² Asking to the project's scientific directors, it is possible instead to receive a copy of two different types of transcriptions

¹Comic strips available here: <http://www.valico.org/vignette.html>.

²VALICO website: www.valico.org.

```

1  <HEAD>
2  <doc-id>
3  <charset>ansi</charset>
4  <lingua>italiano</lingua>
5  <aut_NC>[REDACTED]</aut_NC>
6  <fornitore>[REDACTED]</fornitore>
7  <trascr>[REDACTED]</trascr>
8  <data>2011,0,0</data>
9  <luogo>Milano, IT</luogo>
10 <ist>scuola</ist>
11 <ist_nome>?</ist_nome>
12 </doc-id>
13 <set-id>
14 <corpus>valico</corpus>
15 <gruppo_num>01,g1</gruppo_num>
16 <gruppo_nome>al parco</gruppo_nome>
17 </set-id>
18 <autore>
19 <specifiche>f</specifiche>
20 <eta>19-25</eta>
21 <status>3</status>
22 <annualita>3</annualita>
23 <lingua1>tedesco</lingua1>
24 <lingue>inglese</lingue>
25 <scolarizzazione>un</scolarizzazione>
26 <permanenza>(5, Milano, IT)</permanenza>
27 <esposizione>sc, am</esposizione>
28 </autore>
29 <testo>
30 <tipo_forma>c-lib_narr</tipo_forma>
31 <tipo_produzione>did</tipo_produzione>
32 <topics>...</topics>
33 <keyw>([REDACTED], [REDACTED], [REDACTED], [REDACTED]);?</keyw>
34 <test></test>
35 <qualita>orig</qualita>
36 <esecuzione>ms</esecuzione>
37 <cap-min>0</cap-min>
38 </testo>
39 <ref>
40 <stel>[REDACTED].txt, [REDACTED].txt, al parco_G.txt, P.txt</stel>
41 <cons>love_C.txt</cons>
42 <txttext></txttext>
43 <imgext></imgext>
44 <txtint></txtint>
45 <imgint></imgint>
46 </ref>
47 </HEAD>

```

FIGURE 3.1: Header provided per each VALICO text. Proper names are darkened.

(i.e. diplomatic transcription, TD, and tokenized and marked-up transcription, TTM), both beginning with a header (see Figure 3.1 and the guidelines provided by Prof. Manuel Barbera and Prof. Elisa Corino) providing meta-data about e.g. supplier, transcriber, learner, text typology.³ We used TTM transcriptions because we got more transcription of this kind. TTM transcriptions provides tokenized punctuation and mark-up information of various kind.⁴

We preprocessed TTM transcriptions in order to maintain only the text, excluding information that was not useful for the analyses covered by this

³Header guidelines: <http://www.bmanuel.org/projects/br-g01.html>.

⁴Guidelines explaining in detail the collection and transcription of VALICO texts can be found here: <http://www.bmanuel.org/projects/br-g00.html>. See <http://www.bmanuel.org/projects/br-g02.html> for the transcription criteria.

```

1 <BODY>
2 <emph_sc>IL TESTO</emph>
3 <col_red?>Ieri al parco</col> <corr>ho letto</corr>
  stavo leggendo
4 il giornale quando che ho sentito
5 urlai forte . E dopo un po ho potuto
6 vedere l' origine di questo : Un <corr>uomo</corr>
7 ragazzo <ins><interlinea>di</interlinea></ins> alta
  statura e muscoloso
8 <corr>e</corr> panava con la sua fidanzata sula
9 spalla . Lei ancora stava urlando . <corr>Ho</corr>
  In
10 questo momento ho capito che la
11 fidanzata non voleva essere sula
12 spalla , perÃ² lei non ha potuto
13 opporsi a lui . Proprio per questo ho
14 deciso di intervenire . Si deve sapere che
15 anche io sono di alta statura e
16 mi alleno . CosÃ¬ era semplice per me
17 di sopraffare l' altro e liberare
18 la <corr>don</corr> ragazza . Mi ho sentito come
19 un' eroe .
20 Ma ei invece ha iniziato di urlare
21 da nuova . E non poteva finire .
22 All' inizio non ho capito perchÃ¨ ma
23 dopo un po ho potuto immaginare :
24 lei era innamorata di lui e tutto era
25 parte di un gioco tra di loro .
26 </BODY>

```

FIGURE 3.2: TTM text before preprocessing.

thesis. See Figure 3.2 and Figure 3.3 for an example of TTM before and after preprocessing respectively.⁵ The header is maintained as it is, only the body is edited.

In the preprocessing we cleaned encoding issues (texts were encoded in ASCII and we converted them into UTF-8)⁶ and information about the textual phylogenesis (i.e. textual criticism), such as diacritics indicating the separation point in a word wrapped onto the next line or tags indicating corrections (i.e. tag `corr` in Figure 3.2). The preprocessing was carried out manually with the help of regular expressions.

Since we did not have originals nor digital copies of handwritten texts, we could not verify transcription quality.⁷ Hence, we rely on transcription

⁵Note that metadata in Figure 3.1 refers to the text in Figures 3.2 and 3.3.

⁶In Figure 3.2 the words ‘perÃ²’, ‘CosÃ¬’ and ‘perchÃ¨’ present encoding issues normalized as *però*, *così* and *perché*. Ambiguity in the normalization occurs only for ‘perchÃ¨’ because, even though in Italian the correct spelling would be *perché* we cannot be sure that the learner actually used the right accent. We normalized ‘Ã¨’ as *è* because in other texts it was present in words for which ambiguity does not arise (i.e. third person singular present indicative of the verb *to be*).

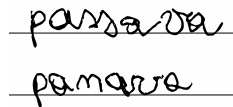
⁷Handwritten texts have the code *ms* in the tag `esecuzione`, line 36 in Figure 3.1.

```

1 <BODY>
2 IL TESTO
3 Ieri al parco stavo leggendo il giornale quando che
  ho sentito urli forte.
4 E dopo un po ho potuto vedere l'origine di questo:
  Un ragazzo di alta statura e muscoloso panava con
  la sua fidanzata sulla spalla.
5 Lei ancora stava urlando.
6 In questo momento ho capito che la fidanzata non
  voleva essere sulla spalla, però lei non ha potuto
  opporsi a lui.
7 Proprio per questo ho deciso di intervenire.
8 Si deve sapere che anche io sono di alta statura e
  mi alleno.
9 Così era semplice per me di sopraffare l'altro e
  liberare la ragazza.
10 Mi ho sentito come un'eroe.
11 Ma ei invece ha iniziato di urlare da nuova.
12 E non poteva finire.
13 All'inizio non ho capito perchè ma dopo un po ho
  potuto immaginare: lei era innamorata di lui e
  tutto era parte di un gioco tra di loro.
14 </BODY>

```

FIGURE 3.3: TTM text after preprocessing.


FIGURE 3.4: Handwriting of *passava*.

even when some doubts can be cast. For example *panava* (in Italian *panare* is a verb meaning ‘to dip in breadcrumbs’) in the sentence reported in Example 11 (lines 5–9 in Figure 3.2 and line 4 Figure 3.3) could be an error due to interpretation of handwriting (see Figure 3.4 in which we write *passava* in a way in which it cannot be confused with *panava* and another which is highly ambiguous).

(11) **LS:** Un ragazzo di alta statura e muscoloso **panava** con la sua fidanzata sulla spalla.

TH: Un ragazzo di alta statura e muscoloso **passava** con la sua fidanzata sulla spalla.

*A tall, muscular boy was **walking by** with his girlfriend on his shoulder.*

The final output of the preprocessing phase was a UTF-8 txt file containing the header and the cleaned text written one sentence per line.

This preprocessing was applied only to the selected texts. The texts included in the parallel treebank VALICO-UD were selected according to two main design criteria (i.e. L1 and comic strip—tag `lingua1` and `cons` line 23

and line 41 in Figure 3.1), chosen to obtain a resource suitable for training and testing models in context of tasks like Native Language Identification (NLI). A third criterion, the learners' year of study (tag `annualita` line 22 in Figure 3.1)), has been considered for the core section of the treebank (i.e. a section of the treebank featuring error annotation and fully-manually-revised syntactic annotation, as described in Section 3.1.1).

Regarding the first criterion, i.e. learners' L1, we selected texts written by German, English, Spanish and French native speakers. We made this decision for two reasons. The first is related to the available VALICO data. The mother tongues (i.e. L1s) of the largest groups of learners in VALICO are German, Spanish and French (Corino and Marelllo, 2017, p. 85). Then, we selected also English, because—even though it is not as large as the other three groups—is the most studied language and we wanted data by English native speakers. The second reason is related to computational reasons. Even though these four languages are all Indo-European languages, they represent two different families: while French and Spanish belong to Romance languages like Italian, English and German belong to Germanic languages. Thus, we believe that using this 4-language subcorpus for tasks like NLI, the task would be challenging enough because of the similarity between the languages.⁸ In addition, a further challenge for NLI concerns the fact that speakers of these four languages are expected to study or be in contact with the other L1s involved, as we will see in the paragraph in which we give a picture about the other languages known by the authors.

For what concerns the second design criterion, we tried to have a balanced number of texts per comic strip in order to exclude topic bias which could inflate the results in NLP tasks. We selected the data referred to two different comic strips, each about a different topic: “Ieri al parco...” (T1 in Table 3.1, reported in Figure 3.5) and “Stazione” (T2 in Table 3.1, reported in Figure 3.6). We selected these two comic strips because they are the two that have elicited most of VALICO's texts (644 and 658, respectively) (Corino and Marelllo, 2017, p. 89). While “Ieri al parco...” requires a narrative development and is relatively poor in details, “Stazione” is richer in details and elicits mainly descriptive texts. Eventually, following the two design criteria above mentioned, we obtained a subcorpus of VALICO consisting of 237 texts (2,234 sentences) as shown in Table 3.1.

⁸However, a NLI task making use of VALICO-UD data would be simplified because it uses data from four L1s instead of the eleven represented in the TOEFL11 corpus.

L1	# Texts	# Sentences	# LS Tokens	# TH Tokens
	T1 T2	T1 T2		
German (DE)	58 29 29	622 280 342	8,729	8,838
English (EN)	60 42 18	662 474 188	9,834	10,029
Spanish (ES)	59 42 17	381 266 115	8,270	8,361
French (FR)	60 30 30	569 249 320	8,623	8,686
EN+FR+DE+ES	237	2,234	35,456	35,914

TABLE 3.1: Summary of VALICO-UD composition. T1 and T2 stands for the two different topics eliciting the texts.

Among the 237 authors, 202 are adults from 19 years old and up, 12 are children between 8 and 13 years old, 23 are teenagers from 14 to 18 years old. As far as authors' background is concerned, 187 authors have a university education, whilst the remaining are studying at high school or elementary and middle school.

Looking at other languages known by the authors, among the 177 non-native speakers of EN,⁹ 150 out of 177 know EN as second language.¹⁰ As far as DE native speakers are concerned, the most popular second language is EN (known by 49 out 54 authors of whom we know this metadata), followed by FR (27), Latin (7) and ES (6). Please note that more than one language can be referred to one author. As far as EN native speakers are concerned, FR is the most popular second language (23 out of 33 authors of whom we know this metadata), followed by Spanish (12), and German (7). As far as ES native speakers are concerned, the most popular second language is EN (41 out of 43 authors of whom we know this metadata), followed by FR (15), Catalan (7) and DE (6). As far as FR native speakers are concerned, all declared to know EN, of whom 13 know also ES, 3 DE.¹¹ This information about authors' known L2s is important to understand the substantial cross-linguistic influence that can be present in VALICO-UD data, hence resulting in an increased challenge in NLI task, for example.

⁹Number obtained subtracting to the totality of the authors, 237, the totality of the anglophones, 60.

¹⁰This metadata is not known or left blank for 47 authors (not known: EN = 26, ES = 16, total = 42; blank: DE = 4, EN = 1, total = 5).

¹¹Note that we are considering only the number of authors whose L2s are known, excluding the 47 authors whose L2 metadata is not known or blank.

Each sentence of VALICO-UD has been normalized following Italian standard variety norms, creating *de facto* a parallel, sentence-aligned treebank.¹² In fact, for each learner sentence, a meaning-based target hypothesis has been written as explained in detail in Section 3.2. Then each of the two parallel corpora encompassed in the resource (i.e. original sentences and target hypotheses) has been automatically parsed using a model of UDPipe (Straka, 2018) trained on two Italian UD treebanks: ISDT (Simi, Bosco, and Montemagni, 2014) and PoSTWITA (Sanguinetti et al., 2018), the same model used in (Cignarella et al., 2020). We used this model because ISDT represents the standard written Italian and proved to be suitable for training (Zeman et al., 2018), while PoSTWITA, collecting tweets, contains a more non-standard variety of Italian which can be more similar to the text genre included in VALICO-UD.

Two parts can be identified in the VALICO-UD, characterised by different sizes, and different richness and quality of annotation. The smallest part, i.e. the gold standard, is the core section of the treebank which features error annotation and linguistic annotation that has been wholly manually revised and adapted to the VALICO-UD annotation scheme (described in Section 5.1.2). The remaining part of the parallel treebank, the silver standard has been automatically parsed and a quantitative evaluation of the quality of automatic annotation is provided in Section 6.1. In what follows we describe these two parts of the resource.

3.1.1 Gold standard

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

TABLE 3.2: Summary of VALICO-UD core section.

As usual in the development of resources, in order to test and validate the design of the annotation schema, we applied it on a portion of data (see Table 3.2), already mentioned above as the core gold standard of the treebank.

¹²Note that for the normalization the totality of the text is considered.



FIGURE 3.5: “Ieri al parco...” (*Yesterday at the park...*) comic strip from VALICO.

It is composed of a selection of VALICO-UD texts elicited by one comic strip, namely the one entitled “Ieri al parco...” (*Yesterday at the park...*, corresponding to T1 in Table 3.1), reported in Figure 3.5.

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, 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. The reason of her madness can be explained by the balloon over her head in which a heart is depicted. As far as the criterion of the learner’s year of study of Italian language is concerned, in Table 3.3, we report a summary of the texts sorted according to it—mean and standard deviation in brackets.

As shown in Table 3.3, 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 3.3 and

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)

TABLE 3.3: Core section summary according to selection criteria (mean and standard deviation in brackets).

marked with the asterisk in Year of Study column). For what concerns FR texts, we could not find 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 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 these 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.

The automatically obtained parsed output of the gold section data has been entirely manually revised. In addition, this treebank data feature also error annotation (see Chapter 4) in one CoNLL-U additional field (see Section 5.1.1).

3.1.2 Silver standard

The silver standard is composed by 201 texts, 107 elicited by the comic strip shown in Figure 3.5 and described in the previous subsection, and 94 elicited by the comic strip shown in Figure 3.6. These texts are grouped as shown in Table 3.4.

This comic strip is clearly different from the first one. In this single cartoon many different things happen. In the centre, there is a man dressed in



FIGURE 3.6: “Stazione” (*Station*) comic strip from VALICO.

black stealing a suitcase from a kissing couple. In the foreground, an old man smoking a cigarette looks at the thief. He is probably his accomplice. A mechanic to the right of the kissing couple seems to be looking in the direction of the old man in the foreground. In the meantime, a little white dog runs away, knocking over a table with everything on it. The woman who was sitting at the table gets dirty and screams. She is probably the dog’s owner. A couple behind the thief drops a shoe box. This adds to the confusion already present in the scene. Everything in this station is very dirty and messy. In the background there are various people entering or leaving the shops in the station and three trains from which people get on and off. Much more could be added by describing in detail the objects in the scene.

Compared to the texts elicited by “Ieri al parco...”, the average number of sentences in the texts elicited by this comic strip is higher. We performed an unpaired t test to check if this difference was statistically significant. The resultant two-tailed P value equals 0.0008, which means that this difference is considered to be extremely statistically significant.

The silver data are unrevised, i.e. we released the automatic output obtained by the application of the parser UDPipe. These data have been used to perform some data exploration experiments reported in Chapter 6.

L1	# Texts	# Sentences	# LS Tokens	# TH Tokens
	T1 T2	T1 T2		
DE	49 20 29	529 187 342	7,538	7,637
EN	51 33 18	512 324 188	7,452	7,641
ES	50 33 17	304 189 115	6,406	6,483
FR	51 21 30	491 171 320	7,276	7,321
EN+FR+DE+ES	201	1,836	28,672	29,082

TABLE 3.4: Summary of VALICO-UD silver data.

3.2 Target Hypothesis writing

In VALICO-UD we have been developing a parallel treebank made of learner texts and Target Hypotheses (THs). These two corpora are sentence aligned, i.e. to each Learner Sentence (LS) there is one target hypothesis (TH). A TH is essentially a normalized version of what the learner wrote. Depending on what is normalized, it is possible to distinguish form-based and meaning-based THs (see Section 2.2.3.1). Form-based THs normalize only grammatical errors, not considering the context or what the author wanted to say. Usually form-based THs do not consider semantic and collocational anomalies (e.g. FALKO or MERLIN). On the other hand, meaning-based THs take context and learners’ intention into account.¹³

Both types of THs are difficult to be performed reliably (e.g. Tetreault and Chodorow, 2008b; Rosen et al., 2014; Dahlmeier, Ng, and Wu, 2013), however form-based THs associated to detailed guidelines can be more reliable and easier to be generated (e.g. FALKO, see Table 2.1; Lüdeling, 2008; Reznicek et al., 2010; Lüdeling and Hirschmann, 2015). On the other hand, a more explicit task context (e.g. comic strip elicited texts, like in VALICO or COMIGS) can facilitate a reliable annotation of meaning-based THs (Meurers, 2015; Köhn and Köhn, 2018).

Since we developed VALICO-UD bearing NLP tasks in mind, we decided to include in the normalization of the single TH associated to the LS also lexical and context-dependent errors—errors that are usually considered in GEC

¹³Both FALKO and MERLIN have two THs per deviant learner sentence: the first TH called *minimal target hypothesis* and can be equated to the form-based TH; the second is called *extended target hypothesis* and can be equated to the meaning-based TH.

tasks—, despite these are not normalized in the form-based TH of similar learner projects (i.e. FALKO, MERLIN).

Initially, following what has been done for the CLC, we left this task to a single annotator under the supervision of another one, in order to test the impact of subjectivity in the task and to evaluate the cases that can be discussed in the guidelines. The annotator had to write a corresponding TH per LS normalizing all that they believed needed normalization.

It is well known in the literature that lexical errors create the most disagreement among annotators (Rosen et al., 2014; Del Río Gayo and Mendes, 2018b), including some lexico-grammatical errors (e.g. use of prepositions; Tetreault and Chodorow, 2008b). As far as lexical errors are concerned, in VALICO-UD this issue was partially resolved because of the use of comic strips to elicit texts. Comic strips, in fact, constrain the context so that lexis is circumscribed. For the remaining sources of disagreement, at first, we left the decision to the annotator.

As a result, the TH was highly invasive and subjective, because also acceptable forms were normalized. In Example 12 we report a learner sentence (LS) and the first invasive version of the TH.

(12) **LS:** Mi ho sentito **come** un'eroe.

TH: Mi sono sentito un eroe.

I felt like a hero.

The normalized version normalizes the choice of the auxiliary of *sentire*, using the *essere* auxiliary verb ('to be') instead of *avere* ('to have') in forming the *passato prossimo* (which formally corresponds to the English present perfect, except that in Italian both *essere* and *avere* can be used to form it), because the verb is used reflexively, thus it requires the *essere* auxiliary. Then the indefinite article *un'* (which correspond to *una*, feminine) is normalized using the masculine. To this point, no disagreement arises. The choice to delete the adverb *come* in the TH can be questionable, since there are contexts in which *sentirsi come* are actually used, and in this context, in particular, it is acceptable, despite less common. To reach this conclusion, we used Google as a corpus and constrained the query using quotation marks. Looking for "sentito un eroe" returned about 78,500 results, whilst only 16 looking for "sentito come un eroe". However, among these 16 occurrences, one is from a theater script, and another from a novel.¹⁴ For this reason, even though it is less used, in a less invasive TH, we decided not to normalize cases like this

¹⁴The novel is titled *La notte dello scorpione*, written by Antonella Scarfagna.

which are due to the sensibility of the annotator. The choice of deleting it in the TH was probably due to the fact that this construction was felt by the annotator as a syntactic calque from Germanic languages. In fact, the annotator, aware that the learner was a young adult native speaker of German who also knew English, may have thought that the use of *come* was influenced by learner's L1 or L2 and therefore, feeling it to be foreign, decided to normalize it.¹⁵

Another example of invasive and subjective TH construction is reported in Example 13.

(13) **LS:** Ce anche un orologio **grande** che indica le **Ore** 2:00.

TH: C'è anche un **grande** orologio che indica le 2:00.

*There is also a **large** clock indicating 2 o'clock.*

In the example above, the invasive normalization concern two parts of the sentence. The first is the position of the adjective in the nominal phrase *un orologio grande* ('a large clock'), the second regards the deletion of *Ore* in the TH. The qualifying adjective *grande* ('large', 'big') in Italian is one of those adjectives that depending on the position in which they are located can change their function or meaning.

As far as the function is concerned, in Italian qualifying adjectives can have a restrictive function or an appositive function depending on their position in relation to the noun they qualify. When positioned to the left of the noun, they have usually a restrictive function; to the right of the noun an appositive function (Renzi, Salvi, and Cardinaletti, 2001, vol. 1, p.316–321). As far as the meaning is concerned, if we look at De Mauro's dictionary,¹⁶ when used as adjective and referred to someone, *grande* has a different meaning depending on its position. If positioned to the left of the noun that qualifies indicates someone distinguished by quality, merit, gifts, genius and the like; whilst if positioned to the right of the noun that qualifies, it indicates someone who has a body size, a physique larger than ordinary. When referred to an object and positioned to the right of the noun it qualifies, as in the Example 13, it indicates a larger size than something else of the same species. Thus, the annotator when normalizing *orologio grande* into *grande orologio* is saying that *grande* indicates a characteristic of the clock in question and not that the watch belongs to a specific class (i.e. that of large clocks) nor bigger

¹⁵The same sentence can be rendered in DE: *Ich fühlte mich wie ein Held*; in EN: *I felt like a hero*.

¹⁶Url of *grande* in De Mauro's dictionary: <https://dizionario.internazionale.it/parola/grande>.

dimensions than usual clocks. In addition, Conte, Bosco, and Mazzei, 2017, in a study based on Italian UD treebanks, analysed the position of descriptive adjectives, distinguishing them into six categories. In the ISDT treebank analyzed by the authors, dimension-related adjectives, like *grande*, are usually positioned before the noun they qualify, thus confirming the annotator's instinct to normalize its position in Example 13. However, this is a subjective decision, and since the order in this sentence can be acceptable in both positions, the less invasive, and final TH do not normalize cases like these.

As far as the noun *Ore* ('hours') is concerned, its deletion in the TH can be justifiable only if also stylistic and register errors are normalized. In this case the use of *ore* can be accepted in formal or bureaucratic register. Also in this example, such as in Example 12, the annotator knowing that the learner's L1 is German, the choice to write *Ore* for indicating the time in Italian could be influenced by their L1. In German, indeed, the substantive *Uhr* is always used when giving information about time.¹⁷ Hence, a less invasive TH for the same sentence would be: *C'è anche un orologio grande che indica le ore 2:00*, normalizing only grammar and orthography (i.e. *Ce* into *C'è*, from pronoun to existential construction, and *Ore* into *ore*).

3.2.1 Guiding principles

The purpose of the first writing of THs is to define guidelines. The reproducibility of a task (in this case, the annotation task) is crucial to ensure that the data is reliably and consistently annotated regardless of the annotator performing the task. Consequently, subjectivity, since it is unavoidable in a task like this, must be as circumscribed as possible. To do so, it is necessary to establish precise criteria and reference resources.

The first principle is to normalize following the formal criterion (Andorno and Rastelli, 2009, p. 58; Barni and Gallina, 2009): the normalization should be as far as possible near to what the learner wrote. The nearness is evaluated counting the TH features (lexical, morphological, syntactic, and semantic) in common with the LS.¹⁸ In Example 14, we report a case in which learner's intended meaning is clear, but it can correspond to at least two different THs.

¹⁷The sentence in Example 13 can be translated into German as: *Es gibt auch eine große Uhr, die 2 Uhr anzeigt.*

¹⁸Note that this process could be automated using for example the Levenshtein distance, a raw count of morphological and syntactic features in common, and semantic similarity measures.

- (14) **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 formal criterion, when dealing with the error reported in Example 14, we decided to choose TH2, because the learner's *signifier* (using de Saussure's term) is maintained, along with other grammatical features, such as the continuous aspect of the verb using the same verb form.

The formal criterion applies also to lexical issues, and it is valid both for non-existent and existent words used in a wrong context. So that, in case of non-existent words, as reported in Example 15, the non-existent word *benco* is reconducted to *bianco*, because of the similarity of the *signifier*.

- (15) **LS:** Poi, il uomo con gli vestito **benco** da una al grande ragazzo [...]
TH: Poi, l'uomo con il vestito **bianco** dà un pugno al grande ragazzo [...]
Then, the man in the white suit gives a punch to the big guy [...]
- (16) **LS:** [...] Giulio ha capito subito che quel giorno farebbe l'eroe della città o piùtostto del parco, salvando una ragazza **indefessa**.
TH: [...] Giulio ha capito subito che quel giorno avrebbe fatto l'eroe della città o piuttosto del parco, salvando una ragazza **indifesa**.
*Giulio immediately realised that that day he was going to be the hero of the town, or rather of the park, saving a **vulnerable** girl.*

Similarly, in cases of existent words used in a wrong context, the normalization should be selected following the formal criterion. In Example 16, we reported part of a sentence in which the learner wrote *indefessa* (meaning 'untiring') in the context of 'vulnerable'. For the principle discussed so far, the annotator should normalize into *indifesa*—because it is formally similar to the learner's *signifier*—and also because it has the right meaning as requested from the context (i.e. 'vulnerable'). This is to say that, even though other Italian words can fulfill the requirements of meaning for the context, such as *vulnerabile*, these must be excluded due to the first principle just described.

The second principle concerns error gravity (see Section 2.2.1) and error count. We decided to consider error gravity and error count when normalizing learner sentences so that, if possible, a TH involving less serious errors is selected, if the counts of errors remain the same (e.g. if to normalize one error we can choose between one serious error or one error less serious, we select the second one).¹⁹ The error gravity hierarchy we followed was in part

¹⁹We are aware that error gravity perception depends on the raters.

inspired by the results reported by Vann, Meyer, and Lorenz, 1984. Only partially because since they examine errors in written academic English, we had to exclude errors like spelling errors due to a different variety of English (i.e. the use of British instead of American spelling) and add errors that were not taken into account in their study (because they selected only twelve errors).

The hierarchy sees spelling errors as the less serious, syntactic errors (e.g. word order) as the most serious, passing through morphological, and lexical errors. To improve agreement between annotators, this hierarchy is used to normalize errors occurring in a token following a precise order, as explained in Section 4.1.

(17) **LS:** *Qualchi minuti fra* il ragazzo ha renduto conto che il brutto sognava..

TH1: *Alcuni minuti dopo* il ragazzo si è reso conto che il brutto sognava.

TH2: *Qualche minuto dopo* il ragazzo si è reso conto che il brutto sognava.

A few minutes later the boy realised that the ugly one was dreaming.

In Example 17 we normalize the learner's phrase *qualchi minuti fra* into *alcuni minuti dopo* (i.e. TH1 in the example) instead of *qualche minuto dopo* (i.e. TH2 in the example) because we consider not only error count but also error gravity. That is to say that, if we count the edits between the learner's phrase and TH1 and TH2 versions, the nearest TH to the learner's phrase would be TH2, but the errors' number and gravity would be increased than if we use TH1 as normalization. In fact, TH2 not only normalizes *fra*—an Italian preposition, which resembles the adverb *fa* ('ago'), i.e. the opposite meaning of *dopo* ('later')—and the number of the indefinite determiner *qualchi*—which in Italian is used only in singular form (i.e. *qualche*), whilst in plural forms *alcuni/le* is used—but also the number of the substantive *minuti*, resulting in three errors, including one concerning the number of a substantive, which is considered as a heavier edit than number issues involving dependents such as determiners or adjectives. This is due to the fact that changing the number of a substantive, we change the learner's sentence more than normalizing grammatical issues (e.g. agreement between the determiner and the noun).²⁰

This second principle is in addition to the first, it does not replace it. Rather, if necessary, the first principle overrides the second. So, in Example 18 the learner wrote *pugnalato*, which means 'stabbed', but it is clear from the context (bear in mind that learners' texts are elicited from comic strips,

²⁰Of course, the number of a substantive is normalized when it is wrongly used in a sentence.

thus context is highly circumscribed) that they meant ‘punch’ which in Italian is rendered as *dare un pugno*. However, an underspecified normalization (e.g. *colpito* in TH2) would also be contextually plausible and in compliance with the second principle. But since, the first principle overrides the second, in cases like these, a normalization maintaining learner’s intention is chosen (in the Example it is TH1).

(18) **LS:** “Perché hai **pugnalato** il mio ragazzo?”

TH1: “Perché hai **dato un pugno** al mio ragazzo?”

TH2: “Perché hai **colpito** il mio ragazzo?”

“Why did you *stab/punch/hit* my boyfriend?”

The third principle is used in order to avoid annotators’ subjectivity in the normalization. In fact, an annotator in order to normalize something must verify that the contentious forms can be considered wrong after consulting the resources used as reference. These selected reference resources comprise Italian reference corpora and treebanks, an Italian descriptive grammar and a dictionary. The first resource selected as reference is the corpus VINCA (Corino and Marellò, 2017). VINCA is a small reference corpus specifically compiled for VALICO.²¹ It includes texts elicited by the same comic strips used for VALICO, the difference being that VINCA texts are written by Italian native speakers. In particular, from VINCA we extracted the subcorpus of 181 texts (123 elicited by the comic strip “Stazione”, Figure 3.6, 58 by “Ieri al parco...”, Figure 3.5), the same two comic strips selected in VALICO-UD. Moreover, in order to have a greater coverage of structures, comprising also those that do not occur in VINCA (because the only fact that they do not occur in VINCA does not make them wrong), we decided to refer to the Italian reference corpus CORIS (Rossini Favretti, Tamburini, and De Santis, 2002), and to the Italian treebanks available in the UD repository—i.e. ISDT (Simi, Bosco, and Montemagni, 2014), ParTUT (Sanguinetti and Bosco, 2015), VIT (Alfieri and Tamburini, 2016), PoSTWITA (Sanguinetti et al., 2018) and TWIT-TIRÒ (Cignarella, Bosco, and Rosso, 2019). In addition, we referred to the De Mauro’s Dictionary²² (De Mauro, 2016) and to the Italian reference grammar *Grande Grammatica Italiana di Consultazione* (Renzi, Salvi, and Cardinaletti, 2001).

All these resources were exploited by the Italian native speaker for writing the THs. In particular, the TH differs from the corresponding LS only

²¹VINCA is available here: <http://www.valico.org/vinca.html>.

²²The dictionary is accessible here: <https://dizionario.internazionale.it>.

if grammatical errors are encountered—considering as grammatical also orthographical and semantic well-formedness, and acceptability (James, 1998, pp. 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 normalized version must be written, bearing in mind the intended meaning and actual forms used in the learner’s sentence.

It is worth noticing 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, the choice to annotate it linguistically (i.e. including morphological and syntactic annotation) enables the contrastive analysis of learner data and improves the replicability of the analysis (Lee, Li, and Leung, 2017; Doval and Nieto, 2019). In Chapter 5 we deal with the issues in treebanking a learner corpus. In the next one, Chapter 4 we describe the error annotation applied to the core section of VALICO-UD.

Chapter 4

Error Annotation

Error annotation is a practice that has been embraced by various learner corpora (e.g. ICLE (Granger et al., 2002), CLC (Nicholls, 2003), ASK (Tentford, Meurer, and Hofland, 2006), NUCLE (Dahlmeier, Ng, and Wu, 2013), COPLE2 (Del Río Gayo and Mendes, 2018a); see Table 2.1), and it is an important step in order to carry out linguistic research but also to develop automatic systems performing author profiling (e.g. Native Language Identification using error tags as feature) or Grammatical Error Identification (GEI) and Grammatical Error Correction (GEC) (see Section 2.2 for a brief overview). Thus, in order to deploy a resource useful for both (computational or not) linguists and computer scientists, we applied this annotation to the core section of the treebank. In this chapter we describe the methodology and the taxonomy applied, provide error statistics and, finally, conclude the chapter reporting on three inter annotator agreement experiments.

4.1 Methodology

Once the THs were obtained (as detailed in Section 3.2), we used Transcript'omatic (Costantino, 2009), a software developed for the transcription of texts in the VALICO project, adapting it in order to display each LS and the corresponding TH in parallel and to subsequently obtain a visualisation that allows immediate detection of errors in the LS. A new file was then created in which the errors (i.e. the differences between the LS and its TH) were annotated in XML format. This level of annotation is provided inside the *err* field in the CoNLL-U file of both LSs and THs,¹ currently available for the

¹CoNLL-U is a revised version of the CoNLL-X format (Buchholz and Marsi, 2006) and it is used in the Universal Dependencies formalism to store annotations. It contains comment lines that provide information about sentence id, sentence text, and in VALICO-UD's core section also an additional comment line called *err* containing the error tagged sentence. See Section 5.1.1 to know more about the format.

core section of the treebank, as published in the UD repository.²

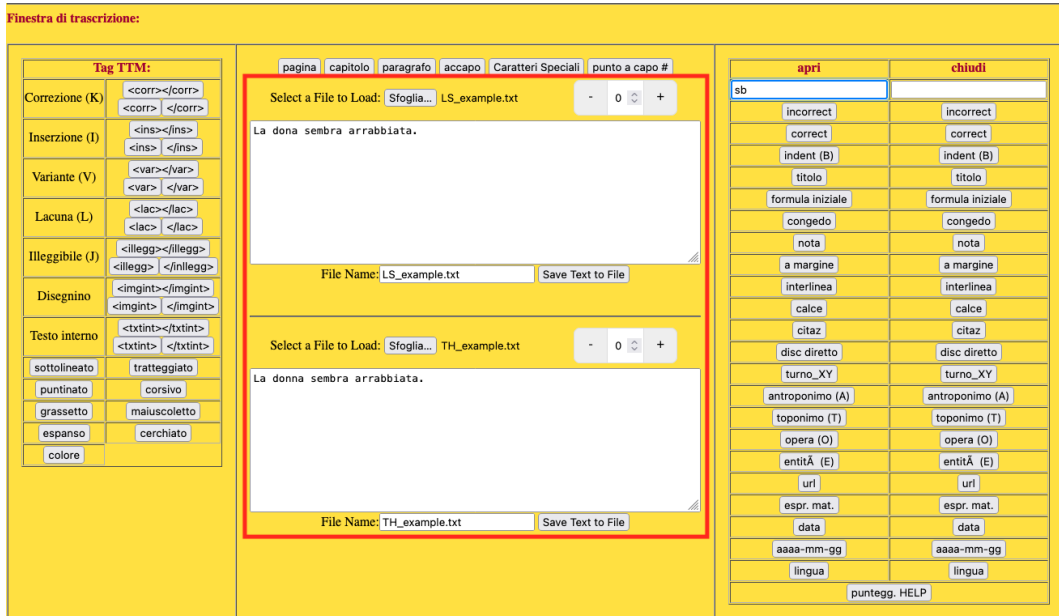


FIGURE 4.1: Visualization of LS and TH using the edited version of Transcript'o-matic.

In Figure 4.1 we report the visualization as provided in the edited version of Transcript'o-matic. In the figure, with the red frame we draw the attention on two text boxes that are used by annotators to visualize LS (in the Example in the upper box) and TH (in the lower box) in parallel. Annotators can choose to edit either the box containing LSs or the one with THs as long as they remain consistent throughout the annotation session. When they want to finish their session, they can save the annotated sentences as a new file clicking on the Save Text to File button right under the text box they edited. Each error-annotated sentence is then inserted as a comment line after the text comment line of CoNLL-U files of both LSs and THs, as shown in Figures 4.2 and 4.3, respectively. In what follows we describe the error coding system to explain how it works and how error annotation was performed.

Our error coding system is based on that developed by Nicholls, 2003 applied to the CLC. In adapting it, we tried to follow as much as possible the requirements stated by Granger, 2003 (see this paper for a detailed description). Using her words, an error coding system to be fully effective should be: *informative but manageable, reusable, flexible, and consistent*. Thanks to its

²UD repository available here: <https://github.com/UniversalDependencies/UD-Italian-Valico>.


```

1 # sent_id = LS_example
2 # text = La dona sembra arrabbiata.
3 # err = La <SB><i>dona</i><c>donna</c></SB> sembra arrabbiata.
4 1 La la DET RD Definite=Def|Gender=Fem|Number=Sing|PronType=Art 2 det _ _
5 2 dona dona NOUN S Gender=Fem|Number=Sing 3 nsubj _ _
6 3 sembra sembrare VERB V Gender=Masc|Number=Sing|Tense=Past|VerbForm=Part 0 root _ _
7 4 arrabbiata arrabbiato ADJ A Gender=Fem|Number=Sing 3 xcomp _ SpaceAfter=No
8 5 . . PUNCT FS _ 3 punct _ SpacesAfter=\n
9

```

FIGURE 4.2: *Err* field in the LS CoNLL-U file.

```

1 # sent_id = TH_example
2 # text = La donna sembra arrabbiata.
3 # err = La <SB><i>dona</i><c>donna</c></SB> sembra arrabbiata.
4 1 La la DET RD Definite=Def|Gender=Fem|Number=Sing|PronType=Art 2 det _ _
5 2 donna donna NOUN S Gender=Fem|Number=Sing 3 nsubj _ _
6 3 sembra sembrare VERB V Gender=Masc|Number=Sing|Tense=Past|VerbForm=Part 0 root _ _
7 4 arrabbiata arrabbiato ADJ A Gender=Fem|Number=Sing 3 xcomp _ SpaceAfter=No
8 5 . . PUNCT FS _ 3 punct _ SpacesAfter=\n
9

```

FIGURE 4.3: *Err* field in the TH CoNLL-U file.

structure, this error coding system easily copes with these requirements. In fact, it is informative (the more letters the more fine-grained the annotation) but manageable (only three positions and no more than eighteen letters to remember), it is reusable (it can be applied to languages different than English or Italian), it is flexible (it is potentially expandable), and consistent (the same phenomenon cannot be tagged with different codes).

Differently from the two-letter tag used in the CLC, our tagset consists of maximum three letters. Following the principle explained by Nicholls, 2003, p. 573 in her article, “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.” We added a third letter which represents the grammatical category involved (Simone, 2008, pp. 303–341), to provide a finer-grained description of the error. With this principle in mind, all errors are encoded using a fixed set of letters which can occur in the first, second and third position:

1. The first letter indicates the general error category or the type of edit necessary in order to pass from the marked LS sequence (i.e. a single token, a phrase, a clause, a sentence; it depends on the involved error) to the TH; the types are D (derivation), F (form), I (inflection), M (missing), R (replace), S (spelling/mechanical), U (unnecessary), and W (word order).
2. The second letter indicates the orthographic, grammatical or syntactic category of the required word. The letters allowed in this position are: A (pronoun), B (double consonants), C (conjunction), D (determiner), E (apostrophe), I (graphic accent), J (adjective), N (noun), O (interjection),

P (punctuation), R (adverb), T (adposition), V (verb), X (auxiliary), and W (more than one token).

3. The third optional letter specifies further features of the error category as indicated by the first letter. Depending on the first letter, in the third position it is possible to have the following letters indicating: A (aspect with I in first position), B (co-occurring tense and mood or double letters, depending if in first position there is I or S, respectively), G (gender related errors, distinguished with F or I in first position), L (foreign word distinguished with F or R in first position), M (mood with I in first position), N (number with I in first position), O (collocation error or gerund error if the first letter is R or I, respectively), P (person with I in first position), S (capitalization with S in first position), T (tense or tokenization³ with I or S in first position, respectively), W (multi-word expression with F, M, R, S or U in first position), and X (existential construction with F, M, R, S or U in first position).

The error coding system provides also special tags for dealing with clauses and multiple complex errors, as provided in Nicholls, 2003. However, these special tags account for the 0.56% of all errors marked in the core section.

The mark-up of errors in VALICO-UD follows the XML annotation standard. Consequently, error tags, which by our choice are in capital letters, are written within angle brackets, and must be encoded twice, once as *start-tag* (e.g. <TAG>) and once as *end-tag* (e.g. </TAG>).⁴ Inside each tag, in our XML scheme, two mandatory tags indicate the incorrect form (i.e. <i>...</i>) and its correct counterpart (i.e. <c>...</c>). If dealing with missing tokens, the tag indicating the incorrect token contains an underscore (i.e. <i>_</i>); conversely, if unnecessary tokens occur, the tag indicating the correct token contains the underscore (i.e. <c>_</c>).

Therefore, the pattern of the annotation of an error is as in Example 19, where the incorrect LS word *dona* occurs within <i> and </i> while the corresponding correct form *donna* occurs within <c> and </c>; they are in turn inside the <SB>...</SB> tag, which indicates the error type, i.e. spelling (i.e. S), double letters (i.e. B).

(19) **LS:** La dona sembra arrabbiata.

TH: La donna sembra arrabbiata.

³Tokenization errors are described in Section 4.2.1.7 and Section 5.1.2.2.

⁴XML specification: <https://www.w3.org/TR/REC-xml/>.

ERR: La <SB><i>dona</i></c><donna</c></SB> sembra arrabbiata.

The woman seems angry.

As previously mentioned, this mark-up is obtained using the edited version Transcript'o-matic.

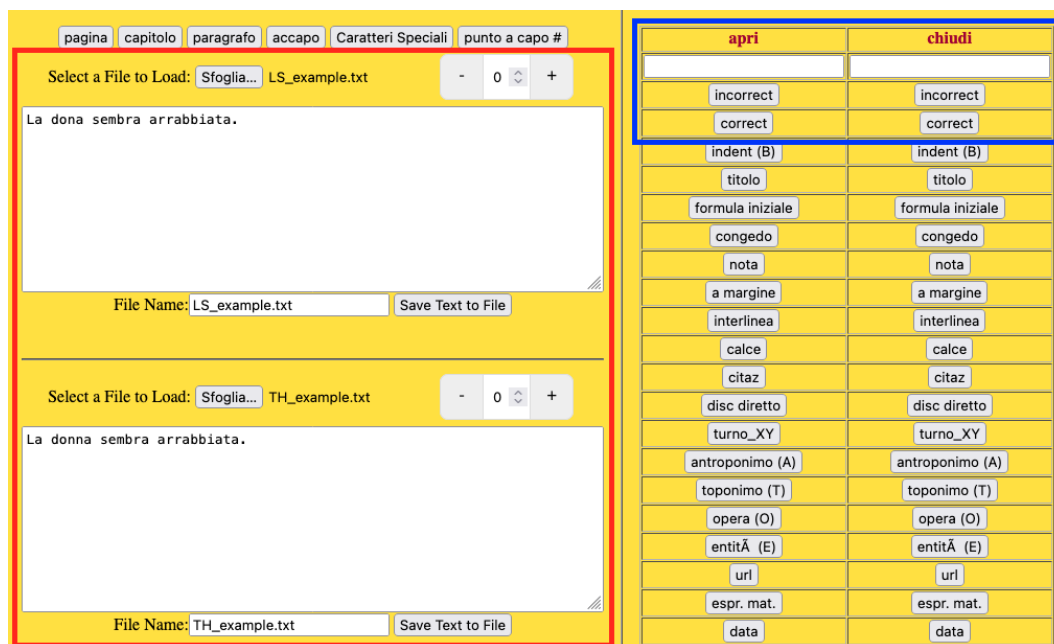


FIGURE 4.4: Adding error tags in sentences with the edited version of Transcript'o-matic.

The region inside the blue frame on the top right shown in Figure 4.4 is used by annotators to mark-up the text. As said, annotators can decide to edit one of the two text boxes, inside the red frame in the figure. First they have to write the error tag inside the text box under *apri* ('open'), on the top right the error tag—they can write it also in lowercase because it will be automatically converted into uppercase—then click on the place where they want to insert the written tag. Second, they can click on the tag *incorrect* under *apri* to open the tag indicating the beginning of the wrong form, and then with the cursor on the end of the wrong form they can click on *incorrect* under *chiudi* ('close'). Third they can proceed with the same steps, but using the button *correct* to signal the normalized form and then closing the error tag writing its letter inside the text box under *chiudi*. Since this procedure is slow and error prone, the second annotator used WebAnno (Castilho et al., 2016) to perform the second error annotation used for inter-annotation agreement (see Section 4.4).⁵ Both tools, however, are error-prone because annotators

⁵WebAnno webpage: <https://webanno.github.io/webanno/>.

have to write the tag themselves and cannot choose it from a drop-down list (see Section 4.4.3).

Note that *dona*, in Example 19, is a case of spelling error resulting in a real word. In fact, *dona* exists in Italian only as verb meaning ‘to donate’ at the third person singular present indicative. However, for the principle explained in Section 3.2, when creating a TH and when annotating an error, it is selected the most probable normalization involving the less serious error, i.e. spelling error.

If a word contains more than one error, or if more steps are needed to go from the LS to the TH form, nested tags are used, as shown in Example 20.

(20) **LS:** Ma Sophia non è andato **disposto**.

TH: Ma Sophia non è andata **bendisposta**.

ERR: [...] <RJ><IJG><i>disposto</i><c>disposta</c></IJG>
<i>disposta</i><c>bendisposta</c></RJ>.

But Sophia did not go willingly.

First, the tag IJG indicates a gender agreement error involving an adjective (i.e. I inflection, J adjective and G gender), from masculine *disposto* to feminine *disposta* (meaning ‘arranged’) referring to a feminine referent (i.e. Sophia). Then, the adjective was replaced with *bendisposta* (translated into ‘willingly’) as indicated by the tag RJ, i.e. replace adjective. In this way, the changes necessary to go from the LS to the TH are retained step by step.⁶

In order to improve the consistency across annotations provided by different annotators, in the error annotation guidelines we provide a hierarchical order to be followed in dealing with multiple errors and nested tags. The errors are organized in a pyramid with at the bottom less serious errors (see Sections 2.2 and 3.2), i.e. spelling errors (including tokenization, capitalization, and punctuation errors) and, proceeding towards the apex, more serious errors, i.e. morphological (e.g. derivation and inflection), lexical (e.g. form and replace), and syntactic (e.g. missing, unnecessary, word order) errors. Thereby, following this hierarchical order, in Example 20 the inflection error (IJG) was encoded before the lexical error involving the replacement of an adjective (RJ).

In addition, we decided to distinguish errors that are not learners’ own, but are caused by the normalization of another error, also called *cascade errors* in the literature (Andorno and Rastelli, 2009, p. 52), as introduced in Section 2.2.2. As far as we know, the distinction between learners’ errors and

⁶This choice implies the possibility of having non-words in the transition from LS to TH.

cascade errors has never been encoded within any annotation project, but we think that it can be useful in the analysis of LSs. In Example 21, we report an error (i.e. wrong gender of the article, tag IDG) which occurs in cascade when a noun is substituted by another noun having a different lexical gender (*banco*—considered as a negative transfer from the learner’s L1, thus it is marked with the tag RNL—is replaced with *panca* having a different lexical gender).

- (21) **LS:** Ieri al parco, ero seduto su un banco [...]
TH: Ieri al parco, ero seduto su una panca [...]
ERR: Ieri al parco, ero seduto su <IDGcascade><RNL><i>un banco</i>
 <c>un panca</c></RNL><i>un panca</i><c>una panca</c></IDGcascade>
Yesterday in the park, I was sitting on a bench [...]

As shown in Example 21, in case of a cascade error, the order in which errors are normalized changes with respect to the hierarchical order described above, because cascade errors must be encoded right after the error triggering them. In fact, in Example 21, following the hierarchical order, the gender agreement error (i.e. IDG), being a morphological error, should have been normalized before the lexical one (i.e. RNL). However, since the lexical error triggered the gender agreement error, the latter is normalized after the former.

For the sake of LS adherence—i.e. formal criterion, as explained in Section 3.2—sometimes we sacrifice naturalness (e.g. we keep repetitions or unnatural-sounding sentences). This criterion is applied also to normalize lexical errors. For instance, in Example 21 the formal criterion applied to the replacement of a lexical item drives our choice in favor of *panca* rather than *panchina* for substituting *banco*—despite *panchina* being the most used term in VINCA (i.e. the paired corpus of VALICO collecting comic strip-elicited texts written by native speakers of Italian, used to write the THs as explained in Section 3.2).

4.2 Tagset description

As introduced in the previous section, our tagset consists of three letters, in which the first indicates the macro-category of the error, the second refers to the word class or the phenomenon involved, the third further specifies the error category if allowed by the tagset restrictions.

Combining the letters allowed in the three positions, in the core section of the treebank we used 120 unique tags plus 28 of these tags marked as *cascade*, for a total of 148 tags.

To summarize in Table 4.1 we report all the letters in their position, their possible combination and the resulting tag meaning. Position 1 indicates the general error category. Position 2 indicates word class (i.e. part of speech) or a specification of the error category. Position 3 is optional and can be used in combination of certain letters in position 1 and 2 to further specify the error category, mainly using the grammatical categories defined in Simone, 2008, pp. 303–341.

In the next subsections we describe the tags organizing them into general error categories: spelling errors, derivation errors, form errors, inflection errors, unnecessary, missing and replace word or phrase errors, word order errors and complex errors.

1	2	3	Meaning
S	ACDJNORTVX ⁷	-	Generic spelling error affecting one of the word classes indicated by the letter in position 2.
	ACDJNORTVX	ST	Capitalization or tokenization errors of a word whose class is identified by the letter in position 2.
	B	-	Spelling error affecting double consonants.
	EIP	MRU	Apostrophes, accents punctuation errors further specified as Missing, Replacement or Unnecessary.
D	ACDJNORTVX	-	Derivation error affecting one of the word classes indicated by the letter in position 2.
F	ACDJNORTVX	-	Form errors affecting one of the word classes indicated by the letter in position 2.
	ACDJNORTVX	G	Alternative form errors affecting gender.
	ACDJNORTVX	L	Foreign forms that do not exist in Italian.
	V	-	Conjugation errors.
I	V	ABMT	Aspect, mood and tense, mood or tense errors.
	V	O	Gerund instead of relative clause.
	ADJNV	GN	Gender or number errors affecting one of the word classes indicated in position 2.
	ADVX	P	Person errors affecting one of the word classes indicated in position 2.
U	ACDJRTW	-	An unnecessary word, or more than one (W), whose word class is specified in position 2.
	JNR	GN	Unnecessary inflection of gender or number in the word class specified in position 2.
M	ACDNRTVX	-	A missing word whose word class is indicated by the letter in position 2.
R	ACDJNRTVX	-	Replace word with another whose word class is indicated by the letter in position 2. ⁸
	ACDJNORTVX	L	Foreign words that do exist in Italian but with another meaning.
W	ACDJNRVW	-	Word order errors involving one of the word classes indicated in position 2.

TABLE 4.1: Summary of the letters allowed in position 1, 2 and 3 to form error tags, and their meaning.

⁷In position 2: A stands for pronoun, C stands for conjunction, D stands for determiner, J stands for adjective, N stands for noun, O stands for interjection, R stands for adverb, T stands for adposition, V stands for verb X stands for auxiliary.

⁸Note that the word classes of the replaced and replacing words can correspond.

4.2.1 Spelling errors

Spelling errors is used with a broad meaning to indicate errors involving spelling or orthographic issues, punctuation, and tokenization. The tags used in the core section that pertain to this category all begin with S. Of all marked errors, mechanical errors account for 32.7% of the total (see Table 4.2). Among mechanical errors we marked thirty-seven unique tags, plus seven unique cascade errors.

In this category we distinguish seven error types, i.e. generic spelling errors, spelling errors involving capitalization, spelling errors involving double letters, punctuation errors, spelling errors involving apostrophes, spelling error involving graphic accents, word boundary errors.

4.2.1.1 Generic spelling errors

The tag used to refer to generic spelling errors is formed by the S in first position plus a letter in second position indicating *the word class of the corrected word* (Nicholls, 2003), as shown in Example 22.

- (22) Ma <SN><i>sorpesa</i><c>sorpresa</c></SN>!
But surprise!

In the core section we marked spelling errors involving ten different word classes—i.e. pronouns (SA), conjunctions (SC), determiners (SD), adjectives (SJ), nouns (SN), interjections (SO), adverbs (SR), prepositions (ST), verbs (SV), and auxiliary verbs (SX). These errors account for 25% of all spelling errors.

4.2.1.2 Issues with capitalization

Capitalization issues are indicated by adding in third position the letter S as shown in Example 23.

- (23) Un altro uomo al <SNS><i>Parco</i><c>parco</c></SNS>
Another man at the park

In the core section we marked capitalization errors involving the same ten word classes of generic spelling errors—i.e. pronouns (SAS), conjunctions (SCS), determiners (SDS), adjectives (SJS), nouns (SNS), interjections (SOS), adverbs (SRS), prepositions (STS), verbs (SVS) and auxiliary verbs (SXS). These errors account for 10.05% of all spelling errors.

4.2.1.3 Issues with double letters

Spelling errors involving double letters, consonants in particular, are marked with the letter B in second position, as shown in Example 24.

- (24) l'altra <SB><i>personna</i><c>persona</c></SB> [...] era il suo amore
*the other **person** [...] was her love*

We decided not to include also the word class of the word featuring this kind of spelling error because we considered it not necessary. This type of error might be further specified indicating the letter involved (e.g. N in the example). These errors account for 16.91% of all spelling errors.

4.2.1.4 Issues with punctuation

Punctuation errors are marked with the letter P in second position. They are further divided into missing (i.e. SPM), unnecessary (i.e. SPU) and replacement (i.e. SPR) punctuation errors. In Example 25 a missing punctuation error is reported.⁹

- (25) Lei era terrorizzata <SPM><i>_</i><c>,</c></SPM> gridava e urlava
She was terrified, she was screaming and shouting

These errors account for 28.68% of all spelling errors.

4.2.1.5 Issues with apostrophes

Errors involving apostrophes are marked with the letter E in second position, as shown in Example 26.

- (26) a cercare un <SEM><i>po</i><c>po'</c></SEM> di calma
*looking for a **bit** of calm*

Also in this case, we use the third letter to indicate a missing apostrophe (i.e. SEM), an unnecessary apostrophe (i.e. SEU) or the replacement of the apostrophe with a graphic accent (i.e. SER). These errors account for 3.43% of all spelling errors.

⁹Note that in the examples, in case of errors which are not the ones we are interested in for the explanation, we report the normalized version—e.g. *terrorizzata* in Example 25 was written as *terrozzata*, a word that does not exist in Italian.

4.2.1.6 Issues with graphic accent

Errors involving graphic accents are marked with the letter I in second position, as exemplified in Example 27.

(27) non lo so <SIR><i>perchè</i><c>perché</c></SIR>
I don't know why

Also in this case, we use the third letter to indicate a missing accent (i.e. SIM), an unnecessary accent (i.e. SIU) or the replacement of the grave accent with the acute accent or *vice versa* (i.e. SIR). These errors account for 13.23% of all spelling errors.

4.2.1.7 Issues with word boundaries

Issues with word boundaries can result in hypersegmentation or hyposegmentation errors (see Section 5.1.2.2) and are marked with the letter T (i.e. tokenization) in third position, as shown in Example 28.

(28) non lo so <SRT><i>per ché</i><c>perché</c></SRT>
I don't know why

Similarly to generic spelling and capitalization issues, we used a letter in second position to indicate the word classes of the words involved. These are: conjunctions (i.e. SCT), adjectives (i.e. SJT), nouns (i.e. SNT), adverbs (i.e. SRT), prepositions (i.e. STT), or, in case of hyposegmentated words, more than one word class (i.e. SWT). These errors account for 2.70% of all spelling errors.

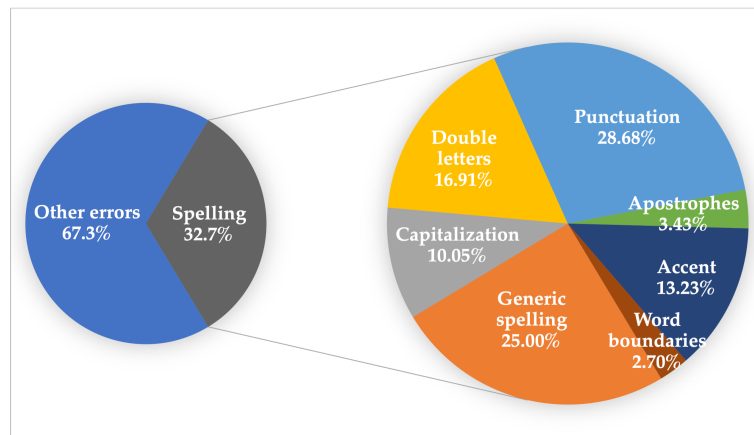


FIGURE 4.5: Distribution of spelling errors and its subcategories in the core section of VALICO-UD.

The subcategories of spelling errors, accounting for 32.7% of all the errors marked in the core section of VALICO-UD, are distributed as shown in Figure 4.5

4.2.2 Derivation errors

Derivation errors—marked with D in first position plus another letter indicating the word class—indicate issues in the morphological formation of a new word from an existing word by the addition (or deletion) of affixes (i.e. prefixes, suffixes and infixes), as shown in Example 29, in which the suffix *-ato* is replaced with *-oso* to form the adjective *muscoloso* from the noun *muscolo*.

- (29) anche lui era <DJ><i>moscolato</i><c>muscoloso</c></DJ>
he too was muscular

It is not always straightforward to distinguish a derivation error from a spelling error (Examples 30 and 31). For example in 30 we report a word with an error that could be interpreted both as spelling or derivation. On the one hand, in 30 the word *desperata* could be considered a derivation error because of the etymological prefix *de-* used instead of *di-*. On the other hand, *e/i* confusion is highly common, hence also spelling error could be a valid tag. Similarly, in Example 31 drawn from the FCE dataset (Yannakoudakis, Briscoe, and Medlock, 2011), a publicly available subcorpus of the CLC, *practice* is considered a derivation error even though it could be also a spelling error (i.e. *practice* as American spelling of *practise*). In case of doubt, in VALICO-UD we mark the least serious error (i.e. spelling in Example 30).¹⁰

- (30) era <SJ><i>desperata</i><c>disperata</c></SJ>
she was desperate

- (31) I would still like to <DV><i>practice</i><c>practise</c></DV>.

Depending on the project, in this category can also be included errors related to wrong selection of pronouns, as reported in Example 32 drawn from the FCE dataset. In VALICO-UD, this error type is marked as replacement (see Section 4.2.5.3).

¹⁰We decided to tag the least serious error because we give the learner the benefit of the doubt, choosing, where possible, the least serious error, i.e. the solution in which they are most aware of the language. Error gravity is introduced in Section 2.2.1 and in Section 3.2, in which we provide the error hierarchy followed in VALICO-UD.

- (32) But Lily never tell her what she really want and what **her** really thought the thing.

*But Lily never tells her what she really wants and what **she** really thinks about things.*

The annotation choices impact on the distribution of the errors, making derivation errors the least commonly marked (excluding complex errors), i.e. 0.8% of the total (see Table 4.2). The tag used are DJ (i.e. wrong derivation of adjective), DN (i.e. wrong derivation of noun) and DR (i.e. wrong derivation of adverb).

4.2.3 Form errors

Form errors are the errors that change the most from project to project. This is due to the fact that whilst projects as ICLE or CLC deal with English as L2, when a richly morphological language is involved—e.g. Italian—a finer-grained distinction between morphological issues might be required. In fact, in ICLE form errors involve morphology (i.e. derivation and inflection) and spelling. In the CLC form errors involve inflection (i.e. degree of adjectives—e.g. *the **better** time* instead of *the **best** time*—, verb form selection—e.g. *which lessons and activities are better to **be filmed** in order to...* instead of *which lessons and activities are best to **film** in order to...*—, plural—e.g. *some **kind** of food* for *some **kinds** of food*). In VALICO-UD spelling, derivation and inflection have a tag of their own, and form error tag is used for identifying:

1. Verb forms in wrong distributional slot (Example 33);¹¹
2. Wrong selection of closed-class words (Examples 34, 35 and 36);
3. Issues with words that fulfil the same grammatical function but have two alternative forms according to spelling rules (e.g. conjunctions or masculine determiners having a different form depending on the following word) (Examples 37 and 38);
4. Foreign forms that do not exist in Italian, distinguishing them from those resulting in real Italian words. We mark the former with F in first position followed by the word class and the letter L in third position, as shown in Example 39. The latter is marked as replacement error (see Example 55 in Section 4.2.5.3).

¹¹Note that it is different from verb tense errors, which in VALICO-UD are instead marked as inflection errors.

- (33) quando ha <FV><i>ascolto</i><c>ascoltato</c></FV> la donna gridare
*when he **heard** the woman screaming*
- (34) Mi ha detto di <FR><i>no</i><c>non</c></FR> intromettermi
*she told me **not** to interfere*
- (35) ha iniziato a <FA><i>urlare a me</i><c>urlarmi</c></FA>
*she started **shouting at me***
- (36) <FA><i>Me</i><c>Mi</c></FA> portava dal dottore
*He was taking **me** to the doctor*
- (37) Litigavano <FC><i>ed</i><c>e</c></FC> l'uomo portava [...]
*They were fighting **and** the man was carrying [...]*
- (38) era <FDG><i>il</i><c>l'</c></FDG> amico della donna
*he was **the** woman's friend*
- (39) pensava che fosse un <FNL><i>crime</i><c>crimine</c></FNL>
*he thought it was a **crime***

Of all marked errors, form errors account for 6.0% of the total (see Table 4.2). Of these, 45.33% involve alternative forms (i.e. Point 3), 34.67% foreign forms (i.e. Point 4), 13.33% wrong selection of closed-class words (i.e. Point 2) and 6.67% verb forms in wrong distributional slot (i.e. Point 1). The distribution of form errors is shown in Figure 4.6.

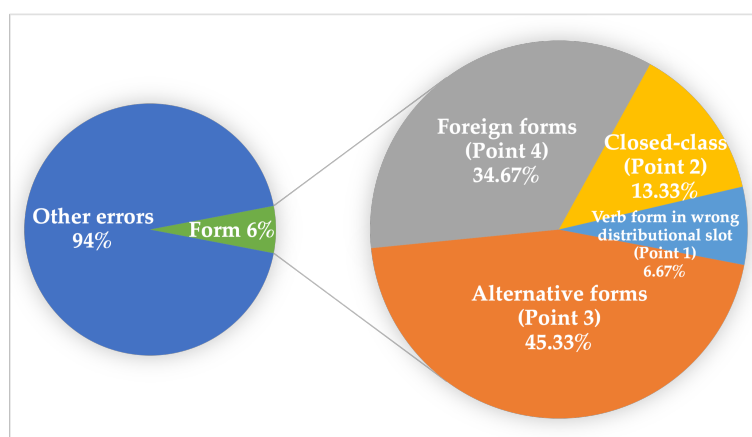


FIGURE 4.6: Distribution of form errors in the core section of VALICO-UD.

With regard to consistency, this is the category in which we experienced the greatest uncertainty among annotators who had to choose between form or replacement (see Section 4.2.5.3). In general R is used in combination to

closed-class words when the grammatical function changes. For instance, in Example 36 we use form and not replacement because the grammatical function is in both cases indirect object. On the contrary, in Example 54 in Section 4.2.5.3 the function of the normalized form is different, thus we marked it as a replacement error.

4.2.4 Inflection errors

Inflection tag, together with derivation tag, indicates issues involving the morphology of a word. Whilst derivation deals with affixes used to create new words from existent ones, inflection tag refers to issues in the mechanism used to express different grammatical categories maintaining the meaning of the inflected word.

In VALICO-UD inflection errors are marked with the letter I in first position, followed by the word class and by a third letter which specifies the grammatical category involved as defined in Simone, 2008, pp. 303–341—i.e. P for person, N for number, G for gender, D for definiteness (not used because in Italian, definiteness is not expressed via inflection),¹² C for case (not used),¹³ T for tense, A for aspect, M for mood, V for voice (not used).¹⁴

The only inflection error tag formed by only two letters indicate conjugation errors. In Italian there are three conjugations (i.e. *-are*, *-ere* and *-ire*) and depending on the conjugation there are different inflection rules. Errors involving conjugation—due to irregular verbs (Example 40) or to a wrong conjugation rule applied (Example 41)—are marked with the tag IV.

(40) quando all'improvviso è <IV><i>apparito</i><c>apparso</c></IV> un uomo
when suddenly a man appeared

(41) povero bambino <IV><i>piangiava</i><c>piangeva</c></IV> molto
poor child, he cried a lot

¹²Plurals make an exception. However, the use of plural indicates definite referents characterized more by 'inclusiveness' (Hawkins, 1978) than 'uniqueness' (Russell, 1905). See Lyons, 1999 to know more about definiteness.

¹³Case is not used because in Italian they are not expressed via inflection. In Italian case is an isolate category, because it is present in personal pronouns and relative pronouns (Simone, 2008, p. 306). See also Malchukov and Spencer, 2008 to know more about the grammatical category of case.

¹⁴Voice was not used only because errors involving verb voice were not found in the error-annotated portion of the treebank. We do not exclude the possibility to use it when the error annotation will be expanded to more and more texts.

If the error involves other categories which are expressed in Italian via verb conjugation (e.g. tense) we use the correspondent tag made of three letters (Example 42).

- (42) Non potevo crederci; <IVT><i>ha detto</i><c>diceva</c></IVT> «il 99 per cento degli adulti trova l'amore prima dei 28 anni di età»
*I couldn't believe it; it **said**: «99 per cent of adults find love before the age of 28»*

Of all errors marked, inflection errors account for 21.00% (see Table 4.2). Of these, gender issues account for 37.40% (i.e. IDG, IVG, IAG, IJG, ING), verb tense errors for 22.90% (i.e. IVT), number issues for 11.07% (i.e. IVN, IDN, INN, IAN, IJN), conjugation errors for 9.92% (i.e. IV), person issues for 5.72% (i.e. IVP, IDP, IXP, IAP), mood for 4.96% (i.e. IVM), both tense and mood errors for 3.82% (i.e. IVB), use of gerund instead of relative clause for 3.43% (i.e. IV0) and aspect conveyed through verbal periphrases for 0.76% (i.e. IVA). The distribution of inflection errors is shown in Figure 4.7. Aspect errors are described in the next subsection.

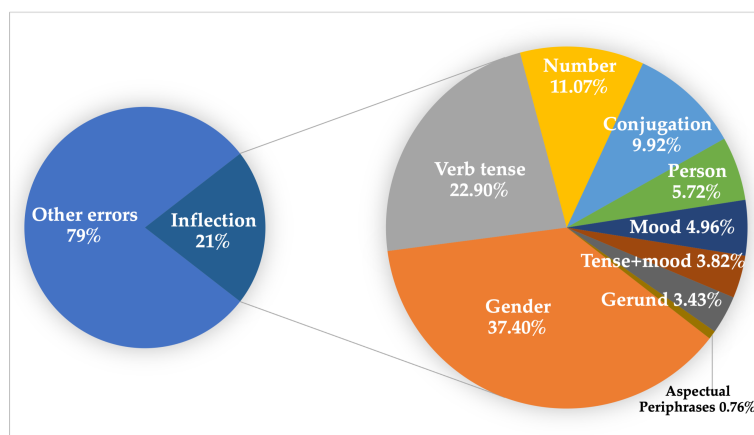


FIGURE 4.7: Distribution of inflection errors in the core section of VALICO-UD.

4.2.4.1 Aspect

In Italian aspect is a covert category, since it has not a formal mark valid only to express aspect (Whorf, 1945). Aspect in Italian can be expressed mainly thorough tense, verbal periphrases or lexis, i.e. verb tenses can convey both tense and aspect, or verbal periphrasis are used (e.g. *comincio a dormire* ‘I start to sleep’ to indicate inchoative aspect) or some verbs can convey aspect semantically—e.g. inchoative aspect of ‘to sleep’ is rendered lexically as *ad-dormentarsi* while continuity is lexicalized into *dormire*, thus it can be said *ho*

dormito per tre ore (I slept for three hours) but not **mi sono addormentato per tre ore* (*I fell asleep for three hours)—or also specific account for different aspect.

In VALICO-UD we used the letter T in third position when the aspect is rendered via verb tense (Example 43), the letter A only when periphrases are involved (Example 44).

- (43) *l'ha colpito e Marco* <IVT><i>cadeva</i><c>è caduto</c></IVT>
he hit him and Marco fell

- (44) *Ieri al parco...* <IVA><i>stavo sedendo</i><c>ero seduto</c></IVA> *e leggevo*
il giornale
Yesterday in the park... I was sitting and reading the newspaper.

4.2.5 Unnecessary, missing and replace word errors

Using a surface strategy taxonomy, in VALICO-UD the majority of errors concerning lexis and grammar are tackled using three different letters in first position: U for unnecessary, M for missing and R for replace. Generally, lexical errors are marked as replacement errors, whilst unnecessary and missing errors target grammatical errors. Unnecessary, missing and replacement errors account for 37% of all errors (see Table 4.2). In the following subsections we describe them individually.

4.2.5.1 Unnecessary word errors

The unnecessary tag has been used to mark tokens that although written by the learners are not necessary. Unnecessary tokens account for 8.4% of all marked errors. Errors involving unnecessary words are usually formed by two letters, U in first position plus as second letter A (i.e. pronoun), C (i.e. conjunction), D (i.e. determiner), J (i.e. adjective), R (i.e. adverb), T (i.e. preposition) or W (i.e. more than one word). Their unnecessary is mainly due to grammatical reasons (92.38% of all unnecessary token errors)—e.g. the presence in Example 45 of a non-required preposition—or to semantic repetition or inconsistency (7.62% of all unnecessary token errors)—e.g. the adverb in Example 46 used to form a non-existent phrasal verb or *mai* ‘never’ which is inconsistent with *da molti anni* ‘for many years’ in Example 47.

- (45) *La ragazza gridava* <UT><i>per</i><c>_</c></UT> *aiuto*
The girl cried for help

(46) Sophia stava camminando <UR><i>avanti</i><c>_</c></UR>

*Sophia was walking **along***

(47) Non riescivo <UR><i>mai</i><c>_</c></UR> a trovare l'amore da molti anni

*I had not been (**ever**) able to find love for many years*

When this tag is formed by three letters, indicates unnecessary morphological features. For instance, tokens which although invariable were inflected by learners (e.g. the gender-invariable adjective *fragile* in Example 48, inflected in *fragila* displaying the feminine singular morpheme, so the tag marks the unnecessary gender inflection of the adjective).¹⁵

(48) una <UJG><i>fragila</i><c>fragile</c></UJG> donna

*a **fragile** woman*

Distribution of errors involving words tagged as unnecessary is shown in Figure 4.8.

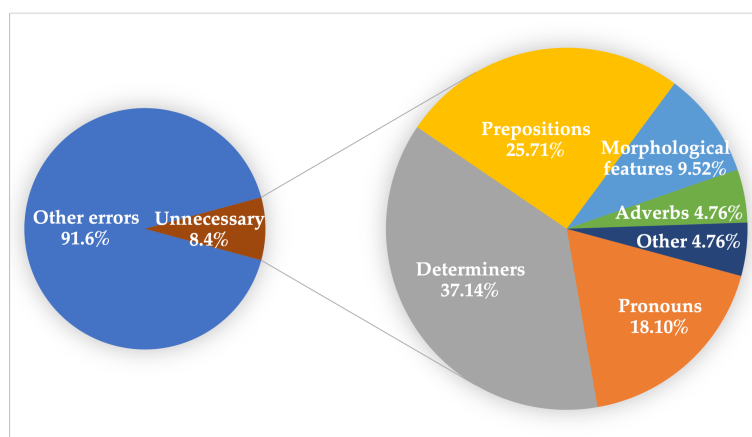


FIGURE 4.8: Distribution of errors marked as unnecessary in the core section of VALICO-UD.

4.2.5.2 Missing word errors

Missing can be seen as the antonym of unnecessary word errors. Instead of unnecessary words there is their absence. The totality of missing word errors are due to grammatical reasons (e.g. the missing definite article in Example 49 or the incomplete argument structure in Example 50).

¹⁵In FCE and consequently in ESL these errors are marked as IJ, inflection adjective, because in English you can never inflect the adjective. In Italian, on the other hand, since there can be both problems of inflection (agreement) and problems of unnecessary inflection (in the case of invariable tokens), we have decided to distinguish between them.

(49) Luca era <MD><i>_</i><c>il</c></MD> suo fidanzato

Luca was her boyfriend

(50) il fratello interrompe <MN><i>_</i><c>la conversazione</c></MN> e prende la sorella

*the brother interrupts **the conversation** and grabs the sister*

Note that in Example 50 we are only showing the missing noun for layout issues, but in the treebank we marked also the missing determiner as cascade error.

Differently from unnecessary word tags which adding a third letter are used to indicate unnecessary morphological features, missing word tags are not used to mark lack of inflection (which is handled with the letter I in first position).

In the core section of VALICO-UD we encountered missing token errors involving pronouns (MA accounts for 23.16% of all missing token errors), conjunctions (MC = 9.47%), determiners (MD = 30.53%), nouns (MN = 2.11%), adverbs (MR = 3.16%), prepositions (MT = 23.16%), verbs (MV = 2.11%) and auxiliaries (MX = 6.32%), as shown in Figure 4.9.

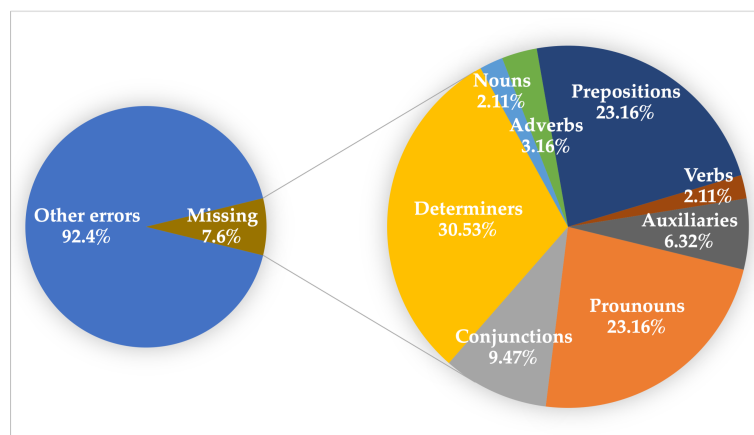


FIGURE 4.9: Distribution of missing word errors in the core section of VALICO-UD.

4.2.5.3 Replace word or phrase errors

Differently from unnecessary and missing word errors, replacement errors are due to lexical or lexico-grammar (following functional approaches)—Example 51 and Example 52, respectively—but also to grammatical errors (Example 53). In the core section of VALICO-UD we found replacement errors involving pronouns (i.e. A), conjunctions (i.e. C), determiners (i.e. D),

adjectives (i.e. J), nouns (i.e. N), adverbs (i.e. R), prepositions (i.e. T), verbs (i.e. V) and auxiliaries (i.e. X). Depending on the part of speech of the *corrected* form involved, we can predict with high confidence the type of error, i.e. (lexico-)grammar (A = 13.74%, C = 3.05%, D = 5.72%, T = 20.61%, X = 9.54%) or lexical (J = 4.96%, N = 16.03, R = 2.29%, V = 18.32%).

(51) Lo faceva per il mio proprio <RN><i>buono</i><c>bene</c></RN>
*He did it for my own **good***¹⁶

(52) Sono rimasto <RT><i>di</i><c>a</c></RT> leggere
*I kept **on** reading*

(53) <RX><i>Aveva</i><c>Era</c></RX> passato
 <RD><i>il</i><c>del</c></RD> tempo quando ha visto che [...]
*Some time **had** passed when she saw that [...]*

As introduced in Section 4.2.3, the difference between FA and RA is that when the function of the pronoun changes between the learner and the correct forms, replacement is used. In Example 54, the function of *l'* is different from that of *gli* (i.e. direct object and indirect object, respectively). For this reason, it is marked as replacement error and not as form.

(54) Invece <RA><i>l'</i><c>gli </c></RA>ha detto parolacce
*But instead she swore at **him***

Differently from negative transfer resulting in non-existing words (which are marked within the form macro-category), negative transfer resulting in real words are treated as replacement errors. Indeed, in Example 39 we saw a negative transfer resulting in a non-existent word in Italian that has been marked as a form error. In Example 55, instead, the word used by the learner is a real Italian word, then the letter R is used in first position.

(55) ha detto che era pericoloso e stupido cominciare un
 <RNL><i>argomento</i><c>litigio</c></RNL> con un uomo così
 grande e brutto
*she said it was dangerous and stupid to start an **argument** with such a big and ugly man*

In the example *argomento* is a semantic calque because in the sentence it acquired the meaning of the English term *argument*. *Argomento*, on the contrary, can indicate: 1. what is used to support an assertion, a thesis; 2. a pretext, motive; 3. the object of a discourse.

¹⁶Note that *il mio proprio bene* is very rare, whilst *il mio bene* is more natural. *Proprio* could be a calque from the English *my own good*.

When a third letter is added, replacement errors usually involve clauses (e.g. in Example 56 a finite clause is replaced with a non-finite clause) or phrases issues (e.g. in Example 57 the learner accidentally produced *the man who had rabies*). These errors account for 5.72% of all replacement errors. From the most frequent to the least, these mark replacement errors involving more than one word (i.e. RWP), replacement of multi-word expressions (i.e. RWW), replacement of finite/non-finite clauses (i.e. RSE), replacement of subordinate (i.e. RS), replacement of coordinate to subordinate (i.e. RCS), and replacement of correlative structure (i.e. RRC).

(56) Gli ho detto <RSE><i>che lasciasse la donna</i>
 <c>di lasciare la donna</c></RSE>
I told him to leave the woman

(57) l'uomo che <RWP><i>aveva la rabbia</i>
 <c>era arrabbiato</c></RWP>
the man who was angry

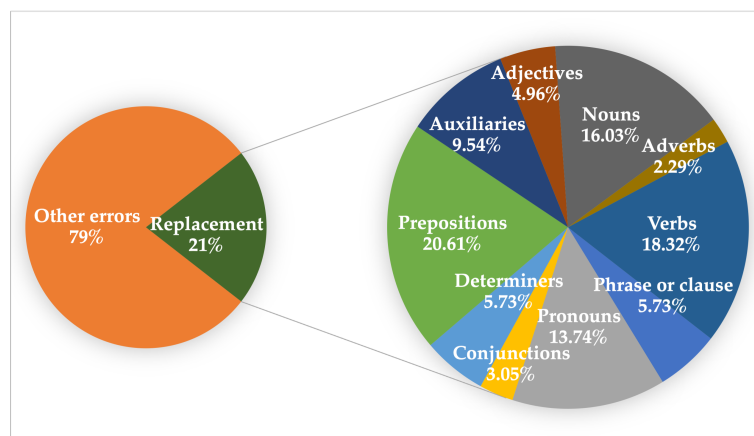


FIGURE 4.10: Distribution of replacement errors in the core section of VALICO-UD.

Replacement errors are distributed in the core section of VALICO-UD as shown in Figure 4.10.

4.2.6 Word order errors

Word order errors concern misplaced words or phrases. In VALICO-UD word order errors account for 2.3% of all marked errors. In particular, these word order errors involve pronouns (31.03%), conjunctions (3.45%), determiners (3.45%), adjectives (3.45%), numerals (3.45%), adverbs (44.83%), verbs (3.45%), and more than one word (6.90%), as shown in Figure 4.11.

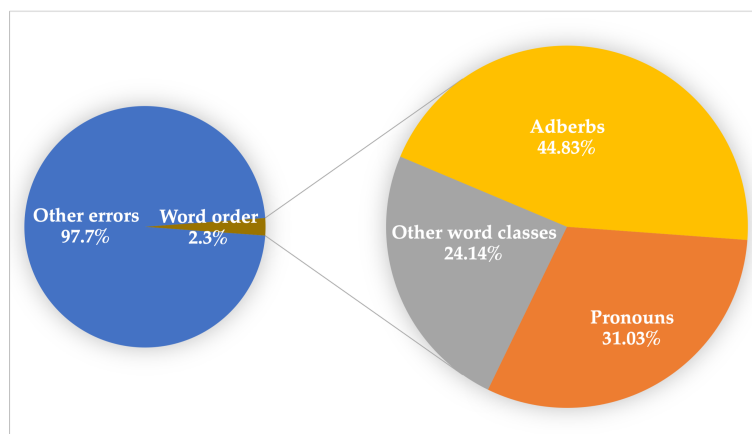


FIGURE 4.11: Distribution of word order errors in the core section of VALICO-UD.

As evident from Figure 4.11, the majority of word order errors encountered in the core section concern adverbs and pronouns. In Example 58 we report an error involving the misplacement of the adverb *sempre* ('always').

- (58) la scusa che mi avevano <WR><i>detto le ragazze sempre</i>
 <c>sempre detto le ragazze</c></WR>
 *the pretext that girls had **always** told me*

As can be seen in the example, we put in the error all the necessary tokens to go from the wrong phrase to the correct one. Initially, word order errors were marked twice, as shown in Example 59 in which the tag WR is wrote once in the place in which the adverb should have been written (and contains the tag c indicating the correct form) and once in the place in which the learner actually wrote the misplaced token (and contains the tag i indicating the incorrect form).

- (59) la scusa che mi avevano <WR><c>sempre</c></WR> detto le ragazze
 <WR><i>sempre</i></WR>

Although in Example 59 it is easier to identify the token that needed to be modified, such an annotation would create count issues and slow down the annotation process. For this reason we opted for the correction shown in Example 58.

4.2.7 Complex errors

Complex errors (i.e. CC) is a generic code to cover multiple errors involving words that cannot be clearly classified in one of the previously described

categories, in line with Nicholls, 2003, p. 575. In VALICO-UD we used it once in the sentence reported in Example 60.

(60) **LS:** Un altro uomo che si siede su un banco del parco la ha vista che la donna rovesciare l'eccedenza equipaggia la spalla e che è andato conservare.

TH: Un altro uomo che stava seduto su una panca del parco ha visto la donna portata sulla spalla e è andato a salvarla.

ERR: [...] <CC><i>rovesciare l'eccedenza equipaggia la spalla</i>
<c>veniva portata sulla spalla</c></CC> [...]

*Another man who was sitting on a bench in the park saw the woman **being carried on his shoulder** and went to rescue her.*

In the marked phrase, the learner wrote literally *spilling the surplus equips the shoulder*, which meaning cannot be established. However, we can suppose from the comic strip that a possible target is *to be carried on the shoulders*. We used the singular for *shoulder* and the definite article to maintain part of the learner's forms. The learner concludes the sentence with *conservare*, likely a non-contextualized translation of 'to save'.

4.3 Error distribution per macro-categories and L1s

In the core section of VALICO-UD, we marked 1,247 errors, including 72 cascade errors, distributed in 325 out of 398 sentences, that is almost 81.7% of sentences contain at least one error (mean number of errors per sentence = 3.13; standard deviation $\sigma = 3.58$), and only 18.3% of sentences are error-free. In Table 4.2 we report error distribution per error macro-category (calculated on the total of errors) and per learners' L1 (calculated on the total of errors per L1).

Error category	Tag	Learners' L1	# L1	% L1	#	%
Spelling	S	DE	55	25.5%	408	32.7%
		EN	77	24.2%		
		ES	178	40.8%		
		FR	98	35.4%		
Derivation	D	DE	2	0.9%	10	0.8%
		EN	3	0.9%		
		ES	1	0.2%		
		FR	4	1.4%		

Error category	Tag	Learners' L1	# L1	% L1	#	%
Form	F	DE	10	4.6%	75	6.0%
		EN	5	1.6%		
		ES	35	8.0%		
		FR	25	9.0%		
Inflection	I	DE	59	27.3%	262	21.0%
		EN	85	26.7%		
		ES	68	15.6%		
		FR	50	18.0%		
Unnecessary	U	DE	23	10.6%	105	8.4%
		EN	28	8.8%		
		ES	35	8.0%		
		FR	19	6.9%		
Missing	M	DE	18	8.3%	95	7.6%
		EN	30	9.4%		
		ES	26	6.0%		
		FR	21	7.6%		
Replace	R	DE	44	20.4%	262	21.0%
		EN	78	24.5%		
		ES	87	19.9%		
		FR	53	19.1%		
Word order	W	DE	5	2.3%	29	2.3%
		EN	11	3.5%		
		ES	6	1.4%		
		FR	7	2.5%		
Complex	C	EN	1	0.3%	1	0.1%
Total		DE	216	17.3%	1,247	100%
		EN	318	25.5%		
		ES	436	35.0%		
		FR	277	22.2%		

TABLE 4.2: Distribution of errors per macro-categories and learners' L1s in the core section of VALICO-UD.

See also Figure 4.12 to observe the data reported on Table 4.2 from a different point of view.

On the one hand, considering all the L1s together (see Table 4.2 and Figure 4.12), we can observe that derivation errors are the least common type of errors. Spelling errors are the most common type of errors (tag S), followed by inflection and replacement errors (tag I and R), the presence of unnecessary tokens (tag U), the absence of tokens (tag M), and wrong form selection errors (tag F).

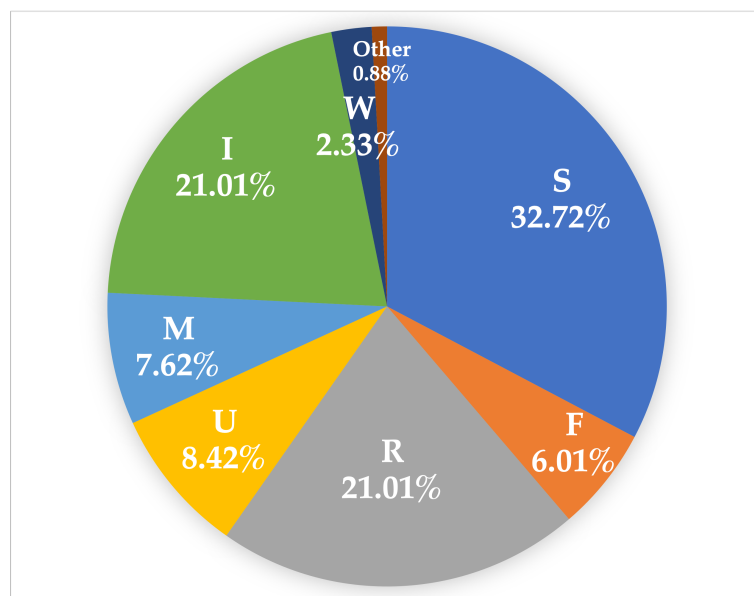


FIGURE 4.12: Distribution of errors per macro-categories in the core section of VALICO-UD.

On the other hand, considering the L1-wise quantitative distribution reported in Table 4.2 (columns 3–5) and in Figure 4.13 we observe a coarse-grained description of the macro-categories as they are distributed in the four learner groups considered (i.e. DE, EN, ES and FR native speakers). Considering each L1 at a time, we can observe that:

- **DE:** Inflection is the most common error type, followed by spelling and replacement errors. These three error types together account for up to 73.2% of errors. Unnecessary and missing tokens account for almost 20% of errors.
- **EN:** Similarly to DE native speakers, also in EN L1 texts inflection is the most common error type. Replacement and spelling errors are the second and the third most common error types; the three types accounting for up to 75.4% of errors. Missing and unnecessary tokens account for almost 20% of errors.
- **ES:** Spelling errors are the most common type of errors, followed by replacement and inflection: together accounting for up to 76.3% of errors. Unnecessary and missing tokens account for less than 15% of errors.
- **FR:** A similar pattern of ES error distribution is found in FR L1 texts. In fact, mechanical errors are the most common type of errors, followed by replacement and inflection. Differently from ES texts, here these three

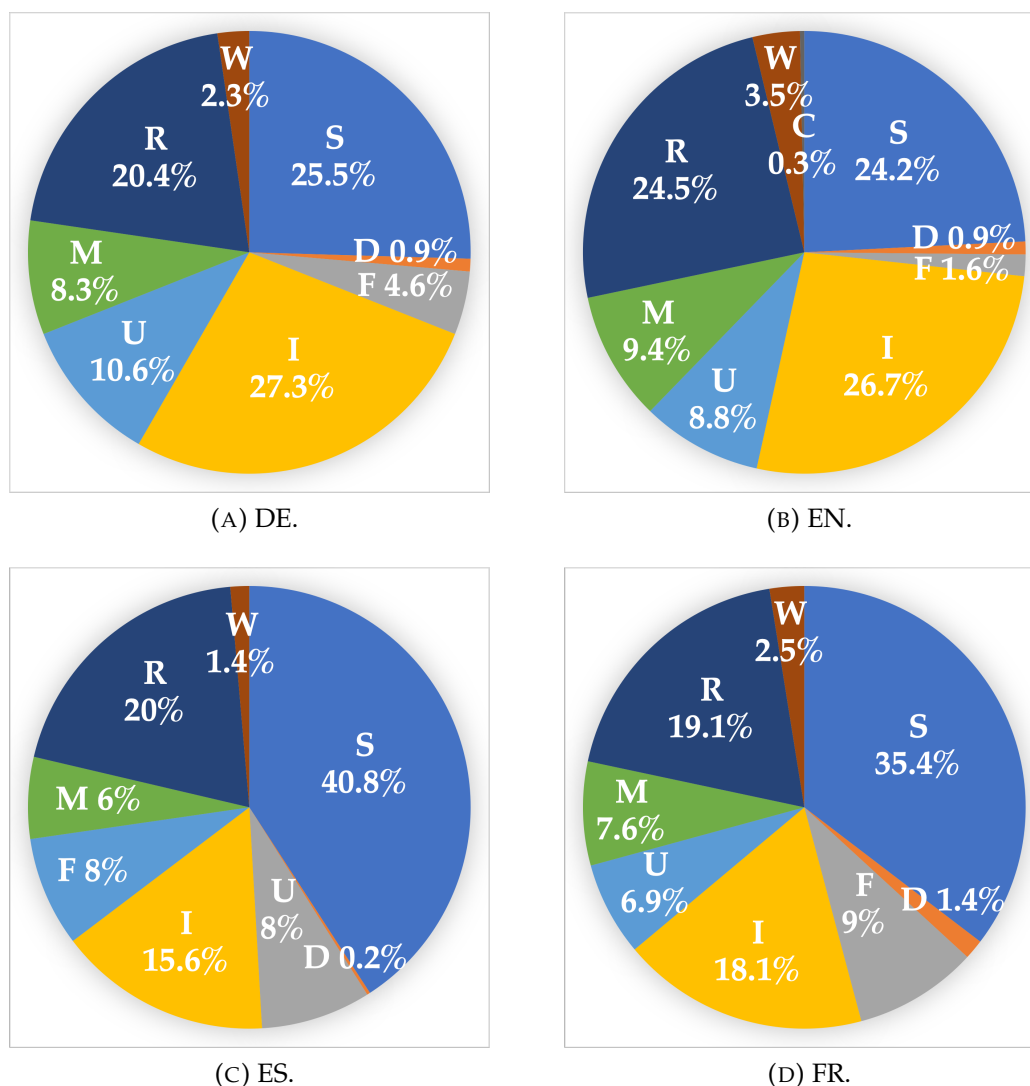


FIGURE 4.13: Error distribution per L1s in the core section of VALICO-UD.

error types together account for up to 72.5% of errors. The fourth most common type of errors, differently from the other three L1s described, is wrong form selection, followed by missing and unnecessary tokens. These three error types account for up to almost 24% of errors.

Although with some minor differences, it is evident from Figure 4.13 that DE and EN, being two Germanic languages, share a similar error pattern, as do ES and FR, both being Romance languages.

However, to better describe the four learner groups, error density must be taken into account. To calculate error density, it is necessary to consider not only the number of errors, but also the number of tokens per L1 (reported in Table 3.2 and in Table 4.3 for clarity). From those with the most errors to

those with the least, texts produced by the learners from different L1s can be ordered as: ES (23.4%), FR (20.6%), DE (18.1%) and EN (13.3%). Further details about error density (ED) are reported in Table 4.3.

Tag	DE	EN	ES	FR	Total
S	4.62	3.23	9.55	7.28	6.01
D	0.17	0.13	0.05	0.30	0.15
F	0.84	0.21	1.88	1.86	1.11
I	4.95	3.57	3.65	3.71	3.86
U	1.93	1.18	1.88	1.41	1.55
M	1.51	1.26	1.39	1.56	1.40
R	3.69	3.27	4.67	3.93	3.86
W	0.42	0.46	0.32	0.52	0.43
Tokens	1,191	2,382	1,864	1,347	6,784
ED	18.14	13.31	23.39	20.56	18.37

TABLE 4.3: Error density (ED) per error macro-category (Tag) and L1.

Despite the small size of the sample (i.e. 36 texts, consisting of 398 sentences and 6,784 tokens), we wanted to measure if these four samples have a statistically significant different distribution of errors when grouped according to their L1 or year of study. To do so, we carried out two tests: the first one is a one-way ANOVA statistical test, the second an unpaired t test, as resumed in Table 4.4. With the first we compared the error distributions between the 4 L1s; with the second we compared the error distributions between the 2 groups aggregating the L1s together: in the first group we selected all the texts written by learners at their first year of study and in the second all texts written by learners at their third or fourth year of study.

Statistics			Population								
Test 1	one-way ANOVA	L1	DE		EN		ES		FR		
		# Texts	9		9		9		9		
Test 2	unpaired <i>t</i> test	YoS	1					3-4			
		L1	DE	EN	ES	FR	DE	EN	ES	FR	
		# Texts	3	3	3	3	3-0	3-0	1-1	3-3	
		Total	12					14			

TABLE 4.4: Summary of the two statistical tests performed. In the second column we have the grouping criteria (i.e. in the first test it is the L1, in the second test it is Year of Study (YoS)) used to distinguish the populations.

In order to be able to test statistically the different distributions of errors between L1s or groups of learners at different years of study, we normalized

the number of errors dividing the per number of sentences composing the text. As far as the one-way ANOVA test is concerned, the p value obtained comparing the distribution of the errors between the four L1s is 0.000085, which means that the result is extremely statistically significant. We then carried out a Post Hoc Tukey HSD to facilitate pairwise comparisons within our ANOVA data. The results confirmed that the ES group significantly differs from each of the other three L1s involved. In particular, when comparing ES learners with FR learners, the resulting p is 0.01315. Comparing ES learners with EN learners, the resulting p is 0.00171. Comparing ES learners with DE learners, the resulting p is 0.00010.

As far as the unpaired t test is concerned, the error type distribution encountered between the two groups of learners (initial and advanced) resulted not to be statistically significant ($p = 0.8452$). Then, we carried out the same test but focusing singularly on each error type. None of them resulted to differs significantly between the two groups (S $p = 0.7469$; D $p = 0.1339$; F $p = 0.9620$; I $p = 0.5214$; U $p = 0.8621$; M $p = 0.8353$; R $p = 0.4854$; W $p = 0.9302$).¹⁷

In the next section we report on three inter annotator agreement experiments carried out to evaluate what is considered to be error by different annotators, then, if they agree on the error, how they normalize it and, eventually, when error and normalization are provided, if they agree on the error tag.

4.4 Inter Annotator Agreement

Error annotation is a complex task which requires time and specific skills as described till here, hence finding suited annotators is not an easy task. In this section we report on three Inter Annotator Agreement (IAA) experiments, measured using Cohen's κ —germane to other IAA studies in the field of learner corpora (see Section 2.2.4).

¹⁷Note that the statistical tests carried out in Section 6.2, in which we use the silver data (see Section 3.1.2), are not comparable to the tests reported here, because here we are using the core section of VALICO-UD (i.e. gold data as reported in Section 3.1.1).

4.4.1 Methodology

Each of the three IAA experiments was performed by two annotators. The two annotators, Bianca Maria De Paolis and myself,¹⁸ are both PhD students in Digital Humanities at University of Turin and are graduated in Foreign Languages with a thesis in Applied Linguistics. They are both native speakers of Italian (i.e. myself from Sicily and Bianca from Piedmont) and proficient in English plus another language (i.e. I speak also Spanish and Bianca French). In addition, they are both beginner learners of German. No training was carried out, only guidelines were provided to the second annotator.¹⁹ This choice is motivated by both annotators' skill with similar tasks and the willingness to evaluate the quality of guidelines.

The texts used to perform the three experiments are the ones of the core section of the treebank (see Section 3.1.1). We decided to use these texts because they form a subcorpus balanced for L1s, prompt, and proficiency levels (expressed in year of study of Italian).

We carried out three experiments, one for assessing the agreement in defining what should be considered as error (i.e. error **identification**), one considering also the normalization associated to the error (i.e. error **normalization**) and the last one considering the error tag (i.e. **error coding system evaluation**). It must be observed that the first two experiments are not focused on error coding but on the errors themselves and their normalization provided by the two annotators. These two experiments show how much the basic tasks that preceded the error coding—i.e. that of detecting the presence of an error and that of normalizing it—can be objective (or subjective).

The first experiment validates the agreement between two annotators in deciding if a token needs to be edited, hence measuring the agreement on error identification (see Section 2.2.2). If the annotators agree on error identification, the second experiment assesses if they also agree on its normalization, hence checking if both annotators provided the same solution to avoid the error. The third experiment measures the agreement on the error tags with explicit THs provided, hence validating the error coding system. Since TH annotation is difficult to perform reliably, we wanted to get rid of this source of disagreement and measure IAA on the application of the tagset, thus validating it. This last experiment was inspired by the results reported by Rosen et al., 2014. In their experiment almost half of the disagreement

¹⁸I would like to thank Bianca for her time and for the valuable talks.

¹⁹Error annotation guidelines are available here: <https://bit.ly/3xB2WJ3>.

corresponds to annotators considering differing target hypotheses. To our knowledge, this is the first study aiming at validating an error coding system providing explicit THs to annotators.

4.4.2 Experiment 1 and 2: error identification and normalization

Error annotation and TH writing are not deterministic tasks for which there is only one right output. For this reason, it is necessary to quantify agreement and disagreement between different annotators in, firstly, identifying the presence of an error and, secondly, its normalization. Hence, we set up two experiments, the first experiment consists in the identification of tokens which should be edited, the second evaluates agreement in the normalization annotators provided. Comparable experiments were performed by Dahlmeier, Ng, and Wu, 2013, Köhn and Köhn, 2018 and Boyd, 2018. Since they reported Cohen’s κ , we use the same measure for comparability reasons. Like Boyd, 2018, we used the script by Lippincott to calculate Cohen’s κ .²⁰

Like in the comparable experiments cited above, we exploited spaces to tokenize LSs and wrote one token per line. Multi-token words not divided by spaces are in one line, indeed. The first experiment is treated as a binary task, thus the annotators are asked to mark 1 or 0 if a token should be edited or not, respectively (see Example 61).

(61)

<i>You</i>	0	
<i>are</i>	1	are the
<i>best</i>	1	best.

In Example 61 *are* is marked as to-be-edited because of a missing token after it (i.e. *the*). The same applies to *best*, which is marked because the sentence lacks the full stop. In other studies (e.g. Köhn and Köhn, 2018; Boyd, 2018) missing tokens are marked in the following token, except for missing tokens at the end of a sentence which are normalized in the previous and sentence-final token actually written by the learner. In our experiment, the annotators were asked to mark missing tokens in the previous token. Thanks

²⁰Lippincott’s script is available here: <https://cswww.essex.ac.uk/Research/nle/arrau/Lippincott/>.

to this choice, we treat all the missing tokens throughout the text in the same way.

Experiment 1, being treated as a binary task, does not add information about the number of errors occurring in a token. If we want to validate also if different annotators find the same number of errors in one token, one possible solution could be that of marking the number of errors in the token and not merely their presence (i.e. 1) or absence (i.e. 0).

We marked as to-be-edited also those tokens that due to the editing of another token should be in turn edited (i.e. *cascade errors*, see Section 2.2.2 and Section 4.1). For example, if an annotator marks *banco*, a masculine common noun, as to-be-edited and change it for a feminine common noun, e.g. *panca*, also its dependents, if any, must be edited to agree with the morphological features of the normalization, such as shown in Example 62.

(62)

<i>seduto</i>	0	
<i>sul</i>	1	<i>sulla</i>
<i>banco</i>	1	<i>panca</i>

In the example both the tokens *sul* and *banco* are marked as to-be-edited, because the articulated preposition *sul* agrees with *banco*, masculine singular, but not with the hypothesised normalization, *panca*, feminine singular. As a result, normalizing also *cascade errors* we can obtain complete THs. This makes the task more challenging because, in this way, identification is influenced by normalization.

Using the space-based tokenization we obtained 5,602 tokens. The annotators worked independently and both marked the same 950 tokens as to-be-edited. In addition to these 950 tokens, the first annotator marked also 159 tokens not marked by the second annotator, for a total of 1,109 tokens marked as to-be-edited. The second annotator instead marked as to-be-edited 1,148 tokens, including 198 not marked by the first annotator. Their agreement, expressed in Cohen's kappa, is $\kappa = 0.82$. A similar result was reported in Köhn and Köhn, 2018 who report for token identification a $\kappa = 0.79$.²¹ This similar result can be explained by the fact that also Köhn and Köhn, 2018 used

²¹Since they followed the guidelines of FALKO (Reznicek, Lüdeling, and Hirschmann, 2013), they have two THs normalizing different types of errors. For this reason, we report the average of their reported κ on TH1 and TH2. At TH1 level they normalize only grammatical errors, excluding lexical or contextual errors, which are instead addressed in the second level, TH2, together with stylistic and pragmatical errors.

a picture-elicited corpus. Lower results are instead reported by Boyd, 2018 ($\kappa = 0.68$) and Dahlmeier, Ng, and Wu, 2013 ($\kappa = 0.39$). In these two studies, texts are not elicited by comic strips, so the context is not circumscribed, thus errors are harder to be identified. For instance, in Example 62 the annotators identified an error only because they knew that the text is elicited by that precise comic strip (i.e. the one in Figure 3.5) in which the events take place in a park and not in a church or a school. In fact, *banco* would have been appropriate to refer to a bench in a church, or a desk in a classroom. Thus, it could have been a valid oblique of the verb *sedere* ('to seat') and without the comic strip nor the defined context (i.e. a man in a park) it would not be identified as error.

Experiment 2 is meant to verify if the two annotators agree in the normalization of a token providing the same edited version. To do so, the annotators were asked to mark the normalization next to the tokens that in their opinion needed to be edited, as shown in the third column of Examples 61 and 62. We measured the agreement of the normalization when both annotators marked the token as to-be-edited. The agreement obtained is $\kappa = 0.69$ (74.74% of the time the annotators wrote the same word or phrase), with only 240 normalization differing (out of 950 tokens marked as to-be-edited by both annotators).²² Also in this task, Köhn and Köhn, 2018 reported a similar result ($\kappa = 0.64$). The results reported in Dahlmeier, Ng, and Wu, 2013 are not comparable because they calculated the agreement of the normalization considering also the associated error tag. Boyd, 2018, instead, reported only the percentage of cases in which both annotators provided the same normalization (i.e. 70%).

In the next subsection, we quali-quantitatively analyse the results obtained in this experiment.

4.4.2.1 Quali-quantitative disagreement analysis

Since in Experiments 1 and 2 we do not have information about the error category, in order to analyse IAA, we exploited the normalized texts written by the annotators and classified disagreement empirically. We call this analysis quali-quantitative because also quantitative information about the qualitative analysis is reported.

²²Please bear in mind that in Experiment 2 we calculate the agreement considering only the tokens that both annotators marked as to-be-edited.

As far as Experiment 1 is concerned, we identified 17 sources of disagreement. We identified as major source of disagreement punctuation errors. In fact, 34.76% of the time, disagreement involved a different judgement about punctuation. In particular, when punctuation disagreement is concerned, more than a quarter of the time (27.73%) commas were involved. On the one hand, there are cases in which the learner used a comma and one of the two annotators marked it as to-be-edited (25.25%). On the other hand, there are cases in which one of the two annotators added one comma where the learner did not use it (66.67% and 8,08%). It is interesting to notice that comma omissions are more frequent and that one of the two annotators tended to normalize these instances more than the other annotator.

Also in Experiment 2 when the annotators disagreed on the normalized text (i.e. 25.26%), 25% of the time the difference concerned punctuation (e.g. the presence of a comma or its substitution with another punctuation mark, the use of different quotations marks). In particular, looking at commas—which are the 68.51% of all the differences due to punctuation—70.27% of the time one annotator used a comma where the other did not. This annotator coincides with the one who normalized more missing commas in the previous experiment.

This substantial difference between the two annotators in using punctuation, and especially commas, could suggest that this is a fuzzy area—such as preposition selection as reported in Tetreault and Chodorow, 2008b, in which two annotators had an agreement of $\kappa = 0.63$ in making judgments on preposition acceptability—and the reasons for this should be investigated in depth, involving first of all more annotators in order to be able to generalize the results. Then, if this would be confirmed, it would be interesting to verify how this subject is treated in textbooks that learners use, or if it is covered at all, as it was done by McEnery and Kifle, 2002 on strong modality markers. This study can be performed using complementary corpora, called by Meunier and Gouverneur, 2009, *pedagogical corpora*. Meunier and Gouverneur, 2009 in their article present the annotation scheme used to mark up the data and argue that these textbook corpora are important resources in learner corpus studies.

Going back to the most common sources of disagreement in Experiment 1,

after punctuation errors (34.76%), lexical issues are the second most common sources of disagreement (16.67%), mistakes (due to annotator distraction or format induced) are the third (12.80%),²³ and different normalization involving tense, mood and aspect are the fourth (10.98%). The fifth most frequent sources of disagreement are due to prepositions and to a different TH in mind (each 6.10%). The remaining disagreement (i.e. 12.40%) is divided between determiners (2.64%), orthography (2.03%), word order (1.63%), clitics (1.22%), marked constructions (0.81%), conjunctions (0.81%), pronouns (0.81%), valency (0.81%), deixis (0.61%), finite/non-finite clause (0.61%), and agreement (0.41%). These sources of disagreement are shown in Figure 4.14 A.

Disagreement on lexical errors can occur when one of the two annotators normalized a lexical error and the other did not (see Example 63)—disagreement on error identification—or when both annotators normalized a lexical error but providing different solutions (see Example 64 and Example 82 for a discussion on the verb *battere*)—disagreement on error normalization.

(63) **LS:** Matteo non poteva **vederla** senza fare niente. [*see her*]

Annotator 1: Matteo non poteva **stare a guardare** senza fare niente.

Annotator 2: *No changes.*

*Matteo could not **stand by** and do nothing.*

(64) **LS:** Mi alzai e **battai** questo tizio arrabbiato.

Annotator 1: Mi alzai e **battei** questo tizio arrabbiato.

Annotator 2: Mi alzai e **colpii** questo tizio arrabbiato.

*I got up and **beat** this angry guy.*

As introduced above, disagreement can also be due to mistakes: unintentional errors induced by human distraction or by the experiment methodology (i.e. format). In general, distraction (probably due to the rapid pace at which the experiment was carried out) caused disagreement in normalizing spelling errors (see Example 65) which with a more thorough look would be identified as errors by both annotators. Disagreement induced by format can be easily described referring to Example 66, in which both annotators marked

²³Note that we used a class called mistakes that inside it can contain corrections of other classes. However, we opted for mistakes because in this cases it was evident that the disagreement was not caused by any of the other classes, but a mistake. In addition, since we revised the annotation to confirm and identify apparent disagreement, as discussed further on, mistakes are identified by both annotators.

the same tokens as to-be-edited, but in normalizing the word-boundary error, they wrote the normalized word in a different slot.

(65)

quando	0		quando	0
all'improvviso	1	all'improvviso	all'improvviso	0

(66)

ha	1	è	ha	1	è
suceso	1	successo	suceso	1	successo
un	0		un	0	
distrasto	1	disastro	distrasto	1	disastro,
per	1	perché	per	1	–
che	1	–	che	1	perché

Other errors induced by the format are caused by the difficulty of consistently following the annotation rules (e.g. missing tokens to be edited in the preceding one).

There is also format induced disagreement that is strictly linked to the nature of errors. In Example 67, for instance, the two annotators wrote the same normalization but in different slots because of the error type they had in mind: the first annotator normalized the gerund using a relative clause but did not marked the relative pronoun as a missing token in the previous token, as the second annotator did. In Example 68, in order to normalize an error affecting two tokens, one annotator opted for a replacement of the auxiliary preserving the verb, the other changed the verb instead.

As a result we counted disagreement of error identification, but the disagreement is not in the presence or absence of errors but in the way in which these are encoded in the experiment.

(67)

una	0		una	0	
ragazza	0		ragazza	1	ragazza che
chiedendo	1	che chiedeva	chiedendo	1	chiedeva
aiuto.	0		aiuto.	0	

The fourth most frequent sources of disagreement concerned verb inflection (in particular tense, mood and aspect). In this case the disagreement can be due to different perception of the text as a whole (e.g. the use of certain

verb tenses due to coherence and cohesion in the text) or to valid alternatives to normalize the same text span (e.g. depending on the speakers' intentions, sometimes it is possible to use both indicative or subjunctive) or different verbal periphrases conveying aspect.

(68)

Anche	1	Comunque	Anche	1	Comunque
ho	1	sono	ho	0	
rimasto	0		rimasto	1	continuato
de	1	a	de	1	a
leggere,	0		leggere,	0	

(69) **LS:** Ho pensato che questa situazione **era** la mia opportunità.

Annotator 1: Ho pensato che questa situazione **fosse** la mia opportunità.

Annotator 2: Ho pensato che questa situazione **era** la mia opportunità.

I thought this situation was my opportunity.

For example, the sentence reported in Example 69, one of the two annotators normalized the mood of the verb using the subjunctive in a subordinate headed by the verb *pensare* ('to think'). In Italian, it is possible to use both subjunctive and indicative mood, but depending on the mood the meaning changes: using the former, it means 'to suppose'; using the latter, it means 'to be sure'. This is a case in which then the annotator subjectivity influences their perception of what it is an error or not. Since the aim of VALICO-UD is to normalize only ungrammatical sentences (see Section 2.2 for the definition used in this dissertation), in the final version of the THs and of the error-annotated sentences, this sentence is not normalized (i.e. TH corresponds to LS).²⁴

Also disagreement involving prepositions or a different TH are mostly caused by valid alternatives. As far as different THs are concerned, they can be due to at least four reasons:

1. Only one annotator identified an error and normalized it (see Example 70);
2. Both annotators identified an error but normalized it providing a different solution for the same text span (see Example 71);

²⁴Note that this IAA was calculated after VALICO-UD was released in the UD repository, so it will be updated in the next release scheduled for May 15, 2022.

3. The number of errors normalized by the two annotators does not coincide (see Example 72);
 4. The learner's sentence is not clear enough and leaves a considerable room for interpretation (see Example 60 described in Section 4.2.7).
- (70) **LS:** questo uomo con i grandi muscoli che **si è sdraiato sulla** terra era il suo fidanzato [*laying on the*]
Annotator 1: questo uomo con i grandi muscoli che **era svenuto a** terra era il suo fidanzato
Annotator 2: *No changes to LS*
this man with big muscles who was unconscious on the ground was her boyfriend
- (71) **LS:** Qualchi minuti fra il ragazzo si è reso conto che il brutto sognava. [**Few-PLUR minutes between*]
Annotator 1: Alcuni minuti dopo il ragazzo si è reso conto che il brutto sognava. [*Some-PLUR minutes later*]
Annotator 2: Qualche minuto dopo il ragazzo si è reso conto che il brutto sognava. [*Few-SING minute later*]
A few minutes later the boy realised that the ugly one was dreaming.
- (72) **LS:** Il brutto uomo **dormendo** era l'amico della donna. [*sleep-GERUND*]
Annotator 1: Il brutto uomo **dormiente** era l'amico della donna. [*sleep-PRESENT_PARTICIPLE*]
The ugly sleeping man was the woman's friend.
Annotator 2: Il brutto uomo **svenuto** era l'amico della donna. [*faint-PAST_PARTICIPLE*]
The ugly unconscious man was the woman's friend.

As far as Experiment 2 is concerned, we analysed disagreement when both annotators provided normalization or, in other words, when both annotators agreed that at least one error was present in the token (i.e. 950 tokens with 240 differing). Being a subsection of the disagreement resulting in Experiment 1, we did not observe new categories. However, as can be expected, disagreement in this second experiment sees the same categories but with a different frequency. These are: 1. Punctuation (22.50%), 2. Different TH (17.50%), 3. Lexis (14.58%), 4. Mistakes (10.00%), 5. Verb inflection (9.55%), 6. Preposition (5.83%), 7. Lexico-grammar (i.e. determiners, pronouns, valency—2.92%), 8. Orthography (2.08%) and 9. more than one category (15.00%), as shown in Figure 4.14 B.

The above-mentioned sources of disagreement can be divided into two major classes: real or apparent disagreement. To distinguish apparent from

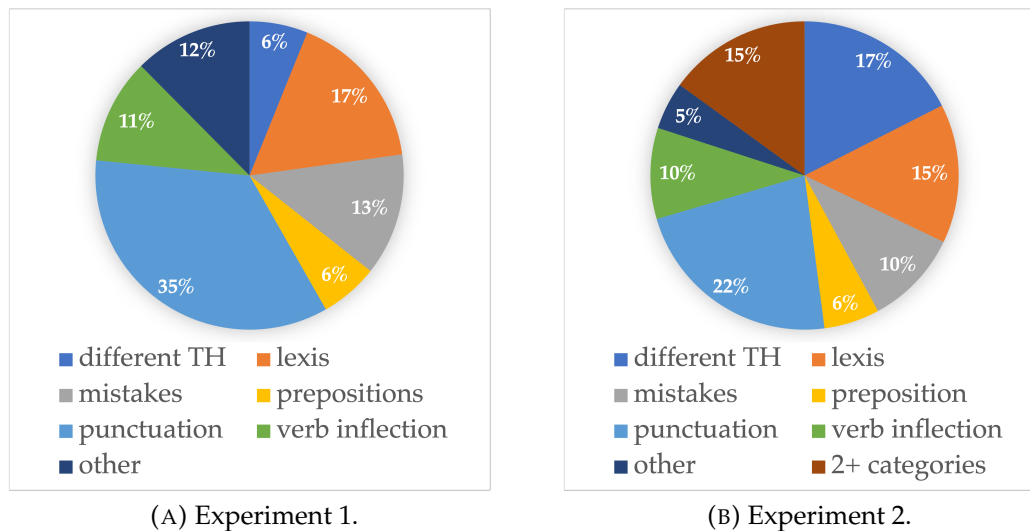


FIGURE 4.14: Sources of disagreement in Experiment 1 and 2.

real disagreement for both experiments, it is necessary to carry out a second round of annotation in which annotators are involved to solve apparent disagreement. We carried out this second round of annotation and found 40.12% of apparent disagreement. A part from mistakes (i.e. distraction and format issues)—which obviously are apparent disagreement—apparent disagreement involved also 75.00% of disagreement on adverbs, 75.00% of disagreement on conjunctions, 34.23% of disagreement on punctuation, 33.33% of disagreement on word order, 26.67% of disagreement on prepositions, 25.43% of disagreement on verb inflection, 25.00% of disagreement on clitics, 22.22% of disagreement on lexis, 15.38% of disagreement on different THs, and 8.70% of disagreement on determiners. In Figure 4.15 the distribution of these categories of apparent disagreement is shown.

During the revision of the disagreement, which was conducted by the two annotators together, it became apparent that the fast pace at which the experiment was carried out had an influence on the identification of errors and their correction. The fast pace affected the individuation of spelling, textual (such as verbal tenses), and lexical errors. In fact, spelling errors require a careful reading, while correcting textual errors, such as tense errors, requires a second reading of the text. A second reading would also be necessary in order to pay attention to consistency. Lexical errors are those most affected by annotators' inconsistency. See for example, the word *suolo* that it was corrected by the same annotator into *terra* in Example 73, being argument of the verb *cadere* ('to fall') and left unchanged in Example 74 (again as argument

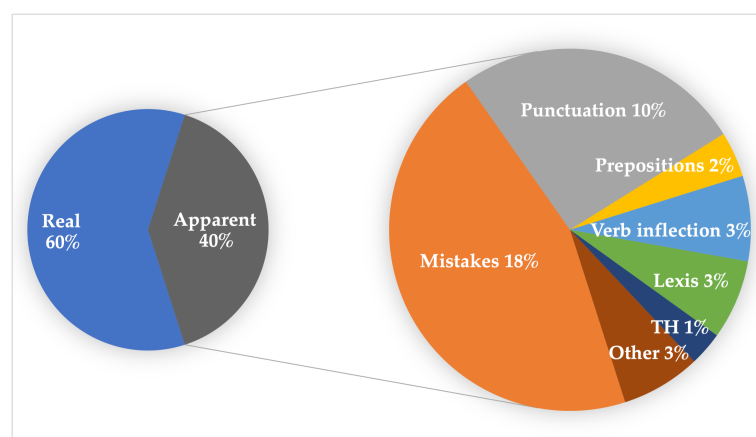


FIGURE 4.15: Experiments 1 and 2: distribution of categories involved in apparent disagreement.

of *cadere*, even though in this case the verb was corrected into *cadere* by the annotator).

(73) **LS:** L'uomo **cade** al **suolo** con la sua lingua fuori e senza spirito.

TH: L'uomo **cade** a **terra** con la lingua fuori e senza sensi.

*The man **falls** to the **ground** with his tongue out and unconscious.*

(74) **LS:** L'uomo è stato sull'**suolo** e Giacomo si ha sentito molto bene

TH: L'uomo è **caduto** al **suolo** e Giacomo si è sentito molto bene

*The man **fell** to the **ground** and Giacomo felt very well*

For both experiments, apparent disagreement is caused by distraction (see Example 65) and format (see Examples 66 and 67) issues. Distraction usually affects spelling or punctuation, but also lexis and verbal inflection, especially with co-occurring errors in the involved tokens or sentence. For instance, see Example 75, in which for distraction one of the two annotators normalizing the spelling of *gritava* forgot to rewrite the comma.

(75)

La	0		La	0
donna	0		donna	0
gritava,	1	gridava	gritava,	1 gridava,
così	0		così	0
Io	1	io	Io	1 io
trovava	1	ho provato a	trovava	1 cercavo di
rescatarla.	1	liberarla.	rescatarla.	1 salvarla.

In the same example, the verb *trovava* (meaning 'to find') is normalized by both annotators into 'to try', 'attempt to', but one of the two annotators

changed also the verb tense from *imperfetto* into *passato remoto*, locating the action in a puntual past, and not in an imperfective past which do not specifies start, end or duration.²⁵ Eventually, *rescatarla*, a non existent word in Italian, was normalized into two valid alternatives.²⁶

Format issues can create apparent disagreement when the authors decide to correct the same error providing the correction in a different place (see Example 67). When the annotators identify the same error but it can be solved differently, it creates apparent disagreement for Experiment 1, real disagreement for Experiment 2 (see Example 68). Also agreement errors corrected on the head or on the dependents, as in the case reported in Example 76, produce apparent disagreement for Experiment 1, real disagreement for Experiment 2.

(76)

portando	0		portando	0
una	0		una	0
donna	0		donna	0
sulla	1	sulle	sulla	0
sua	1	sue	sua	0
spalle.	0		spalle.	1
			spalla.	

To reduce apparent disagreement due to distraction and format some solutions could be provided. As far as distraction issues are concerned, one possible solution to reduce it would be that of considering less texts or the possibility of doing the experiment in several sessions. However, the latter may have had an impact on the annotators' consistency. As far as format issues are concerned, guidelines could better describe how to deal with these issues.

In both experiments, real disagreement occurs when only one of the two annotators corrects an error—in Experiment 2 also if the provided correction is different (in cases in which the same error can be corrected into more than a valid solution, such as it happens with prepositions or lexis)—or have a different TH in mind.

²⁵*Imperfetto* and *passato remoto* both are used to indicate past tense, but they are used to indicate different aspect: the former merges habitual and progressive aspects, the latter perfective and perfect.

²⁶Note that *rescatarla* is a Spanish verb plus clitic pronoun whose meaning can be rendered in Italian with the two corrections provided by the annotators, i.e. *liberarla* ('to free her') and *salvarla* ('to save her'). Even though only one of the two annotators speaks Spanish, it was easy to recover the meaning of *rescatarla*, not only from the context, but also for the similarity with the Italian verb *riscattarla*, whose meaning helps in the normalization.

Register was a source of real disagreement because the two annotators, correcting punctuation but also lexical issues, had different degrees of tolerance. See Examples 77–79.

(77) **LS:** sono io anche un po unamora ta del bel uomo!!

Annotator 1: anche io sono un po' innamorata del bell'uomo!

Annotator 2: anche io sono un po' innamorata del bell'uomo!!

I too am a little in love with the handsome man!!

(78) **LS:** l'uomo era caduto a piombo **sul suolo pavimentato** auuch!

Annotator 1: l'uomo è caduto a piombo **sul suolo pavimentato**, auuch!

Annotator 2: l'uomo è caduto a piombo **a terra**, auuch!

the man fell perpendicularly on the paved floor/ on the ground.

(79) **LS:** mi sono **levato** in piedi.

Annotator 1: mi sono **alzato** in piedi.

Annotator 2: mi sono **levato** in piedi.

I stood up.

Register plays a role in the normalization of both the issues reported in Examples 77–79. In fact, the presence of the error can be detected only if register is taken into account. In Example 77, the normalization of the two exclamation marks into one is mandatory in formal register, whilst *suolo pavimentato* in Example 78 might be used in user manuals or legal texts. Also the normalization (or not) of *levato* in Example 79 depends on register. The verb *levarsi*, although marked as a term of common use in De Mauro, 2016, was normalized into *alzato* by one annotator due to register issues (*levarsi* is felt as a term of literary use).

Very often disagreement due to lexis and to a different TH depends on a LS that is not clear enough (see Example 60) or requires specific language skills (e.g. borrowing from L1/L2 as reported in Examples 80 and 81). In Example 80 it is clear that *lasciare* ('to give up') is a wrong, although in some contexts plausible, translation of the verb 'to leave'. Both annotators, knowing English, recognized easily this semantic calque and normalized it with *sono andato via*.²⁷ Conversely, in Example 81, one of the two annotators, not knowing the Spanish verb plus clitic pronoun *derribarle*, normalized it using a distributional valid verb (considering also *salvare* after it). The other annotator, knowing Spanish, normalized *derribarle* using *batterlo*, meaning 'to

²⁷This text is full of semantic and syntactic calques. As an example of syntactic calque, let us consider *Come imbarazzando!*, literal translation of *How embarrassing!* instead of *Che imbarazzo!*

take him down’, preserving the meaning of the Spanish verb plus clitic but adding a further error (the missing coordinating conjunction *e*, ‘and’).

(80) **LS:** Ho chiesto scusa e **ho lasciato**.

I apologised and left.

(81)

ma	0		ma	0	
Io	1	io	Io	1	io
può	1	potevo	può	1	potevo
derribarle	1	cercare di	derribarle	1	batterlo e
salvare	0		salvare	0	
a	1	–	a	1	–
la	0		la	0	
donna.	0		donna.	0	

Disagreement caused by different language skills can also reveal itself where there is apparently no linguistic borrowing, as shown in Example 82.

(82) **LS:** Ha **battuto** l’uomo che è caduto.

He defeated the man who fell.

Only the annotator knowing French identified *battere* as a semantic calque, from the French *Il a battuto l’uomo qui est tombé*. In fact, the French verb *battre*, is a false friend of *battere*, because it means ‘to hit’ and not ‘to defeat’, so in Italian should be rendered as *picchiare* or better *colpire* in this sentence.

Some of the above mentioned real disagreement could be avoided by better clarifying guidelines. Disagreement due to a different way of correcting a sentence being the least invasive as possible, perhaps, could be avoided in some cases. In particular, cases in which disagreement arises from the annotator decision of normalizing a sentence being the least invasive as possible (e.g. changing the verb to favour the syntactic dependency or leaving the verb and correcting the argument is the least invasive? When dealing with disagreement errors is least invasive to normalize the head or its dependents?) could be avoided if a decision is clearly stated in the guidelines. However, this type of disagreement is very rare and the majority of real disagreement due to plausible alternatives cannot be controlled in the guidelines.

4.4.3 Experiment 3: error coding system evaluation

As described in Sections 4.1 and 4.2, the error coding system applied to the core section of the treebank is complex. Its complexity is given by the fact that it is potentially expandable *ad infinitum*—i.e. the combinations of letters in the three positions are not strictly fixed and can be creatively used by annotators to describe different error nuances, although there are linguistic constraints that limit certain combinations (e.g. in Italian gender errors can affect only some word classes). In the error-tagged core section of the treebank 120 different tags (plus 28 tags marked as cascade) are used. Since its complexity, we wanted to evaluate the error coding system and to do so we devised Experiment 3.

The purpose of the Experiment 3, as introduced in the previous sections, is to validate the error tagset providing to the two annotators LSs and THs. In this way, annotators had to annotate the errors already identified by the difference between the LS and its TH. In particular, the aim is to verify that the tagset is unambiguous (i.e. there should be no option to mark the same error using two different tags) and that the guidelines are clear enough to provide assistance if/where needed. Another objective of this experiment is to verify that explicit THs ensure the reliability of the analysis and to test what kind of errors can be problematic to annotate despite explicit THs.

We calculated the IAA in Cohen’s κ twice. It is worth noticing that the agreement is calculated on pairs of tags, and when one of the two annotators marked more tags than the other, a zero is added to indicate that only the other annotator marked the error. First, annotators reached a moderate agreement ($\kappa = 0.50$), in Landis and Koch, 1977 terms. Looking at the disagreement, there was a high percentage of apparent disagreement due to annotators’ mistakes and format. As can be easily predictable, and as it has been commented for previous experiments, also in this case, some errors can be missed (e.g. spelling). In addition, as mentioned in Section 4.1, annotators used different annotation tools and their output was in different formats: one produces a txt file containing XML-like tags, the other a special type of tsv file. Thus, a conversion was necessary to compare the two annotations. For this reason, differently from previous experiments in which both annotators produced the annotation in the same format, in this experiment also conversion errors might occur. The two annotators revised together the disagreement, solving apparent disagreement. In total 1,203 error tags were marked by both annotators (more than one tag can be applied to one token).

Annotator 1 marked 1,247 tags (including 44 not marked by Annotator 2), Annotator 2 marked 1,241 tags (including 38 not marked by Annotator 1). This corresponds to an almost perfect agreement ($\kappa = 0.95$).

From these results, we can state that THs alone does not ensure replicability, because in a highly complex task, revision to solve apparent disagreement is absolutely necessary. However, the results obtained after solving apparent disagreement confirm what hypothesised in the literature (Lüdeling, 2008; Reznicek, Lüdeling, and Hirschmann, 2013; Meurers, 2015; Rosen et al., 2014), i.e. explicit THs strongly improve the replicability and reliability of the analysis.

As far as comparability is concerned, these results cannot be compared with those of the other experiments reported in the literature because there are not experiments in which the annotators have to annotate errors with explicit THs provided. The results obtained in experiments without THs provided, scores are meaningfully lower in Boyd, 2018 ($\kappa = 0.47$) and in Dahlmeier, Ng, and Wu, 2013 ($\kappa = 0.55$). A different result was achieved by Del Río Gayo and Mendes, 2018b in which the annotators reached an almost perfect agreement ($\kappa = 0.85$).

Another important difference between our experiment and those reported in the literature is the number of error tags used. While in the NUCLE corpus (Dahlmeier, Ng, and Wu, 2013) the used tags are 27, in COPLE2 38 tags (Del Río Gayo and Mendes, 2018b), our annotators used 148 tags, including those marked as cascade error tags (in an error coding system potentially expandable almost *ad infinitum*). This means that despite the complexity (considering complex a tagset with more than 100 tags), explicit THs ensure error annotation reliability.

4.4.3.1 Qualitative disagreement analysis

After a revision useful to remove apparent disagreement, the results obtained using the κ statistic confirm that the guidelines are suitable for the annotation task and quite clear to ensure a reasonable objective interpretation of their content by two different annotators.

After the comparison of the annotations performed by the two annotators, we identified as real disagreement:

- Annotator's sensitivity in deepening the error annotation by marking each step (see Example 83);

- Annotator's identification of foreign borrowings (see Example 84);
- Different interpretation of the error (see Example 85);
- Identification of cascade errors (see Example 86).

(83) **LS:** Si sentiva come un buono carrabiniere e ha cominciato a tigare con il brutto uomo in modo **forzo**.

TH: Si sentiva come un buon carabinieri e ha cominciato a litigare con il brutto uomo in modo **violento**.

*He felt like a good policeman and started to fight with the ugly man in a **violent** way.*

Annotator 1: DJ: *forzo* → *forzoso*; **RJ:** *forzoso* → *violento*.

Annotator 2: RJ: *forzo* → *violento*.

(84) **LS:** *trattava* di ricordare dove la ha lasciato ma non ricordava niente si è disperata moltissimo, ma non li diceva niente al suo marito.

TH: *cercava* di ricordare dove la avesse lasciata ma non ricordava niente e si è disperata moltissimo, ma non diceva niente a suo marito.

*she was **trying** to remember where she had left it but she couldn't remember anything and she was very desperate but she didn't say anything to her husband.*

Annotator 1: SB: *trattava* → *trattava*; **RVL:** *trattava* → *cercava*.

Annotator 2: SB: *trattava* → *trattava*; **RV:** *trattava* → *cercava*.

(85) **LS:** il ragazzo li ha detto que lui non sapeva e per quello **l'ho** aveva fatto.

TH: il ragazzo le ha detto che lui non lo sapeva e per quello **l'**aveva fatto.

the boy told her that he didn't know and that's why he did it.

Annotator 1: UX: *l'ho* → *l'*.

Annotator 2: SA: *l'ho* → *l'*.

(86) **LS:** Quello, è **la mia** amante stupido.» ha detto.

TH: «Quello è **il mio** amante, stupido» ha detto.

*"That's **my** lover, stupid" she said.*

Annotator 1: IDG: *la* → *il*; **IDG:** *mia* → *mio*.

Annotator 2: IDGcascade: *la* → *il*; **IDG:** *mia* → *mio*.

Disagreement reported in Examples 83–85 cannot be solved improving the guidelines, because it cannot be comprehensively defined. In Example 83, one of the two annotators only marked the replacement of *forzo* (a non-existent word in Italian) with *violento*, whilst the other annotator also marked a derivation issue (from the noun *forza*, meaning 'strength', into the adjective *forzoso* instead of *forzo*) before signalling replacement. Disagreement in Example 85 is due to a different way of classifying the same error.

Probably annotator 2 classified the pronoun plus auxiliary as a spelling error, whilst annotator 1 as an unnecessary auxiliary. In this case the different classification depends on the error substance (see Section 2.2.2). Annotator 1 considered only the auxiliary *ho* as substance, whilst annotator 2 must have considered *l'ho* and classified it as spelling error involving the pronoun (i.e. a spelling confusion error, such as the confusion between *where* and *were* in English). Similarly, guidelines cannot comprehend lists of possible foreign borrowings in order to inform annotators, thus disagreement in this case is inevitable. The identification of foreign borrowings—*tratava* in Example 84 which is probably a borrowing from Spanish (*tratar*, meaning ‘to try’)—depends on annotators’ personal knowledge and a possible solution to avoid this type of disagreement would be that of underspecifying the error using the two-letter tag (i.e. RV) instead of the three-letter tag (i.e. RVL). Eventually, in Example 86 disagreement is caused by the notion of cascade error, i.e. an error caused by the correction of another error. Annotator 1 considered the two determiners depending from the noun *amante*—a noun that depending on the extra-linguistic referent can be feminine or masculine and we know from the comic strip that it is masculine—as two distinct agreement errors. Annotator 2, instead, counted only one agreement error, and considered the other determiner child of the head *amante* as an error caused by the correction of the first agreement error. This kind of disagreement can be solved better clarifying in the guidelines how to deal with agreement errors.

Chapter 5

Linguistic annotation

This chapter is divided into four sections: the first, starting with a brief description of Universal Dependencies formalism, presents the annotation scheme followed in treebanking VALICO-UD; the second reports inter annotator agreement results; the third provides treebank statistics, and the fourth incremental parsing evaluation making use of the gold data.

5.1 Treebanking VALICO-UD

In this section, before describing in detail the annotation scheme followed in VALICO-UD, we provide a brief introduction of the UD format, essential for its understanding.

5.1.1 Universal Dependencies formalism

As stated in the introductory page of the project¹, Universal Dependencies (UD) is a project that has been “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 annotation scheme is based on an evolution of (universal) Stanford dependencies (De Marneffe et al., 2014), Google universal part-of-speech tags (Petrov, Das, and McDonald, 2011), and the Intersect interlingua for morphosyntactic tagsets (Zeman, 2008)”.

The UD format, usually shown in CoNLL-U encoding, starts with meta-data 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

¹Universal Dependencies web page: <https://universaldependencies.org/introduction.html>.

separated by a single tab. Thereby, a sentence consists of word lines, as many as the tokens in the sentence, and word lines are composed of the ten 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 87).
2. **FORM** contains the word/token form (i.e. signifier) or punctuation symbol (see second column in Example 87).
3. **LEMMA** contains the lemma of the word form (see third column in Example 87).
4. **UPOS** contains the PoS tag (see fourth column in Example 87).²
5. **XPOS** contains the language-specific PoS tag (see fifth column in Example 87).
6. **FEATS** contains a pipe separated list of morphological features from the universal feature inventory or from a defined language-specific extension (see sixth column in Example 87).³
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 87).
8. **DEPREL** contains the universal dependency relation to the HEAD (see eighth column in Example 87).
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 (each word line contains an underscore in the ninth column) like in other resources where enhanced dependency relations are not annotated (for space reasons it is deleted in Example 87).
10. **MISC** contains any other annotation, including information about the absence of spaces after the token (see ninth column in Example 87).

²The complete list of Universal PoS tags is available here: <https://universaldependencies.org/u/pos/index.html>.

³Universal 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>.

To summarize, in UD morphological annotation is included in four columns (i.e. LEMMA, UPOS, XPOS, FEATS), syntactic annotations in three (HEAD, DEPREL, DEPS).

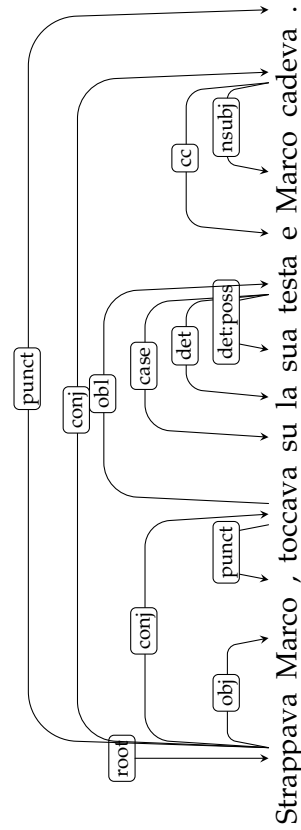
In Example 87, a sample of CoNLL-U is presented (where the ninth column is deleted for layout reasons). Underscores indicate unspecified values.⁴ In case of multi-word tokens (e.g. the articulated preposition *sulla* in sent_id 1), the ID column contains a range (in the example, 5-6), while all the other columns except FORM are left empty (i.e. they contain an underscore). The tokens composing the multi-word token are then separately analyzed in other word lines (i.e. *su* and *la*). The absence of the space between two words that do not compose together a multi-word token (i.e. between the proper noun *Marco* and the comma and the verb *cadeva* and the full stop) is marked in the MISC column by adding `SpaceAfter=No`. While the end of the sentence is marked in the same column with `SpacesAfter=\n`. The tree visualization of the same example is displayed in Example 88.⁵

⁴All details about UD format are available at <https://universaldependencies.org/guidelines.html>.

⁵Note that in the tree representation, text is tokenized.

(87)

# sent_id = 5-06_fr-3									
# text = Strappava Marco , toccava sulla sua testa e Marco cadeva .									
1	Strappava	strappare	VERB	V	Mood=Ind Number=Sing Person=3 Tense=Imp VerbForm=Fin	0	root		
2	Marco	Marco	PROPN	SP	-	1	obj		SpaceAfter=No
3	,	,	PUNCT	FF	-	4	punct		
4	toccava	toccare	VERB	V	Mood=Ind Number=Sing Person=3 Tense=Imp VerbForm=Fin	1	conj		
5-6	sulla	-	-	-	-	-	-		
5	su	su	ADP	E	-	8	case		
6	la	la	DET	RD	Definite=Def Gender=Fem Number=Sing PronType=Art	8	det		
7	sua	suo	DET	AP	Gender=Fem Number=Sing Poss=Yes PronType=Prs	8	det:poss		
8	testa	testa	NOUN	S	Gender=Fem Number=Sing	4	obl		
9	e	e	CCONJ	CC	-	11	cc		
10	Marco	Marco	PROPN	SP	-	11	nsubj		SpaceAfter=No
11	cadeva	cadere	VERB	V	Mood=Ind Number=Sing Person=3 Tense=Imp VerbForm=Fin	1	conj		SpacesAfter=\n
12	.	.	PUNCT	FS	-	1	punct		



(88)

5.1.2 Annotation scheme

In this subsection we describe the annotation we applied to VALICO-UD, both LSs and THs, to actually make VALICO-UD a parallel treebank.

We decided to linguistically annotate it following an hybrid process involving both automatic and manual annotation/correction. First, we applied on the corpus the UDPipe parser (Straka, 2018), using a model trained on two Italian UD treebanks, ISDT (Simi, Bosco, and Montemagni, 2014) and PoSTWITA (Sanguinetti et al., 2018) (see Section 3.1 for the reasons behind this choice). Second, we manually corrected the core section of the treebank, in order to obtain a first gold standard dataset to test our approach, which will be extended to more texts in the future.

In the next subsections we describe the annotation challenges and consequent annotation choices made to adapt the UD format to learner Italian, making use of the experience gained from the manual correction of the core part of VALICO-UD (see Section 3.1.1).

5.1.2.1 Segmentation

To parse a text it is necessary to segment it into sentences. This process can be done automatically using NLP pipelines (e.g. UDPipe) or can be done manually. Since learner texts do not always contain clear signals of sentence boundaries (such as a full stop followed by a word with the first letter capitalized), we carried out sentence segmentation using regular expressions and then we took advantage of the writing of THs to manually correct any segmentation errors as well.

In VALICO-UD we split sentences considering full stops, exclamation marks, and question marks as sentence boundaries. We also considered meaning and sentence completeness to segment sentences which are not divided by any sentence boundary, such as in Example 89.

- (89) # sent_id = 15-02_de-1 Durante un ragazzo è passato.
 # sent_id = 15-03a_de-1 Questo ha portato una ragazza bella
 # sent_id = 15-03b_de-1 questa ha gridata e è defenduta contra il ragazzo.
In the meantime a boy passed by.
This one was carrying a beautiful girl
she shouted and defended herself against the boy.

We did not consider colons, and semi-colons as sentence boundaries. This choice is in line with the segmentation rules followed in VIT, PoSTWITA and

TWITTIRÒ Italian UD treebanks, but not with ISDT and ParTUT. It is worth noticing that PoSTWITA and TWITTIRÒ, being corpora collecting tweets, keep in one sentence the whole tweet, even if the tweet contains more than one *canonical* sentence. In particular, we decided not to consider semi-colons as sentence boundaries because most of the time we used them in the TH to replace a weaker punctuation mark. In Example 90 we report a sentence in which the learner used a comma instead of a stronger punctuation mark. In order to keep the corpus aligned 1:1, we decided to use a semi-colon and not a full stop.

(90) # sent_id = 8-04_es-2

LS: Va bene continuerò, è stato sufficiente di scrivere sul carattere di questo ragazzo che alla fine non fa un'altra cosa che leggere il giornale in cerca di **la-voro**, proprio stupido, chi cerca lavoro in questi tempi?

TH: Va bene continuerò, è stato sufficiente scrivere del carattere di questo ragazzo che alla fine non fa altra cosa che leggere il giornale in cerca di **la-voro**; proprio stupido, chi cerca lavoro in questi tempi?

In the example above, the annotator replaced a comma with a semi-colon. If we had considered the semicolon as sentence boundary, the LS and TH would not be aligned 1:1.⁶

However, if the above mentioned sentence boundaries are contained in direct speech, we consider the whole direct speech as one sentence. Among the Italian UD treebanks, the choice to keep together direct speech containing more than one sentence is only valid for VALICO-UD. Of course, since PoSTWITA and TWITTIRÒ analyse each tweet as one sentence, direct speech containing more than one sentence in one tweet would be considered as a single sentence. So for example, the sentences drawn from ISDT reported in Example 91, in VALICO-UD would have been considered as one sentence.

(91) # sent_id = isst_tanl-757 Clemente Mimun, direttore del Tg 2, si scusa;

sent_id = isst_tanl-758 spiega:

sent_id = isst_tanl-759 "In questo momento mi sento troppo sovraesposto, preferisco non parlare";

sent_id = isst_tanl-760 passa il testimone al suo vice Bruno Socillo.

For this reason, when comparing different treebanks, it is important to take segmentation choices into account. For example, comparisons of tokens

⁶It is worth noticing that the 1:1 alignment is not mandatory. As a matter of fact, in sentence-aligned corpora it is also possible that each source sentence may correspond to one or more target sentences or *vice versa*.

per sentence or maximum sentence depth (i.e. the number of dependency links from the root to the farthest leaf) between VALICO-UD and ISDT treebanks would not be reliable because of the different segmentation choices.

5.1.2.2 Tokenization

Errors involving tokenization can be encoded in the text or due to parsers.⁷ The latter can occur in presence of multi-word tokens, which in Italian mainly are articulated prepositions, 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. The first are defined as performance errors—those called in the literature also as mistakes (Corder, 1967; Corder, 1971)—and can occur also in native language. The second are defined 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, like the one we used for building VALICO-UD (i.e. UDPipe), because tokenization is the starting point for all other annotations. Therefore, a preliminary check was carried out on UDPipe’s output, focusing first on tokenization issues and their correction.

(92)

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

The UD format provides some recommendations to deal with both types of tokenization errors, thus in VALICO-UD we followed them.⁸ As reported

⁷See Section 4.2.1.7 to know more about error annotation and tokenization issues encoded in the text.

⁸Available here: <https://universaldependencies.org/u/overview/typos.html>.

in Example 92, for dealing with wrongly merged words, we split them in different word lines (such as we do for *c'era*) and add in the MISC column of the first word involved (*nessuno* in the example), the `SpaceAfter=No` attribute accompanied by `CorrectSpaceAfter=Yes`.

On the other hand, in case of wrongly split words, as shown in Example 93, morphology and syntax are annotated in the first token in which the word is wrongly split (*co*, word line 10); the other tokens composing the wrongly split word (*si*, word line 11) we leave all the columns empty except for ID, FORM, UPOS (it has to be *X*), HEAD (the ID of the first wrongly split token), and DEPREL (it has to be *goeswith*).

(93)

text = Ma quando la ragazza ha visto il suo ragazzo **co si**, era disperata.

[*But when the girl saw her boyfriend like this, she was desperate.*]

1	Ma	ma	CCONJ	CC	_	14	cc	_
2	quando	quando	SCONJ	CS	_	6	mark	_
3	la	la	DET	RD	[...]	4	det	_
4	ragazza	ragazza	NOUN	S	[...]	6	nsubj	_
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	ragazzo	ragazzo	NOUN	S	[...]	6	obj	_
10	co	cosi	ADV	B	_	6	advmod	_
11	si	_	X	_	_	10	goeswith	SpaceAfter=No

5.1.2.3 Lemmatization

Apart from parsers' lemmatization errors—which usually involve open-class words, e.g. nouns, verbs, adjectives⁹—in learner 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, usually the mother tongue of the learner. In the 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 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

⁹See the UD Universal POS tags page for a complete list: <https://universaldependencies.org/u/pos/>.

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.¹⁰ Thus, in VALICO-UD, misspelled words have their own lemma, according to the PoS assigned, as shown in Examples 94 and 95.

(94) **LS:** Lui Era in **colera**, Lei era **terrozzata** [...]

Lemma: [...] colera [...] terrozzato [...]

PoS: [...] NOUN [...] ADJ [...]

*He was **furious**, she was **terrified** [...]*

(95) **LS:** La **dona** ringraziava suo salvatore [...]

Lemma: [...] dona [...]

PoS: [...] NOUN [...]

*The **woman** thanked her saviour [...]*

In Example 94, *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 95—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 96 and 97.

(96) **LS:** [...] ma Io può **derribarle** salvare a la donna.

Lemma: [...] derribar [...]

PoS: [...] VERB [...]

[...] *but I can beat him and save the woman.*

¹⁰Borderline 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.

(97) **LS:** [...] perchè non l’aveva fatto il **discaount** del 10% [...]

Lemma: [...] discaount [...]

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 lemma associated—following the standard lemmatization rules—remains the correct one. In Example 98, 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 99, 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.

(98) **LS:** [...] gente che soltanto **voglie** chiamare un pò l’atenzione.

Lemma: [...] volere [...]

PoS: [...] AUX [...]

[...] *people who just want to call some attention to themselves.*

(99) **LS:** [...] la ragazza voleva **pertire** ma il ragazzo la teneva.

Lemma: [...] pertire [...]

PoS: [...] VERB [...]

[...] *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.

5.1.2.4 PoS 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 (see Section 2.3.1 for a literature review). 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, in VALICO-UD only one PoS per token is annotated.

To deal with non-canonical forms, two complementary criteria are followed, that is 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, such as exemplified in Example 100, in which a preposition (*Durante*) is used instead of an multi-word expression functioning as an adverb (*Nel mentre*, meaning ‘meanwhile’).

(100) **LS:** **Durante** un ragazzo è passato.

Lemma: [...] durante [...]

PoS: [...] ADP [...]

Meanwhile, a boy passed by.

(101) **LS:** [...] per la esattezza del **relato** devo descrivere quello che ho visto: Un uomo portava una donna sulle spalle e questa quiedeva aiuto.

Lemma: [...] relato [...]

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.*

(102) **LS:** Sono **cerca** della città, ci sono due ragazzi e una ragazza.

Lemma: [...] cerca [...]

PoS: [...] ADP [...]

*They are **next** to the city, there are two boys and a girl.*

(103) **LS:** [...] ha pensato il fratello è stato un **rapinato** e ha salvato la ragazza.

Lemma: [...] rapinato [...]

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, we let distributional properties prevail on lexical features, as shown in Example 95 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 101 and 102. In the former, *relato* is likely borrowed from Spanish with the meaning of ‘story’, and it is not the unusual Italian adjective meaning ‘related’,¹¹ 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’) plus articulated preposition (*della*, meaning ‘to the’). Finally, other existent word (except for closed-class words), used in unusual contexts for the original PoS, are annotated distributionally, as in Example 103, in which the learner used a past participle (*rapinato*, meaning ‘robbed’, functioning also as adjective) in a distributional context of noun.¹²

Lemmatization, as any other piece of information encoded in a resource, must be adapted to the annotation goal. We followed the interlanguage principle, i.e. describe learner language considering a language by itself. However, since lemmatization might be useful also to create alignment seeds between LSs and THs, we developed also a version of the core section with a normalized lemmatization (e.g. *dona* is lemmatized as *donna*).¹³

5.1.2.5 Morphological annotation

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 depending on the language. In Italian, the FEATS column of a verb contains information about the verb form (e.g. finite), mood (e.g. subjunctive), tense (e.g. present), gender (e.g. feminine), number (e.g. singular) and person (e.g. third). And depending on these features, some are excluded—e.g. gender is available only for partitive verbs. As we seen in previous examples in which we reported the CoNLL-U (Examples 87, 92 and 93) proper names, prepositions,

¹¹This meaning is probably unknown to most native speakers together with the Latin loanword *de relato* used in legal settings to refer to an indirect testimony, which might be reported in a legal dictionary but not in the our reference dictionary.

¹²Existent words used in unusual context in Díaz-Negrillo et al., 2010 correspond to the cases in which the lexical stem is inconsistent with the distributional one or in Berzak et al., 2016 should be marked in the typo field.

¹³This differently lemmatized treebank has been used for a case study reported in Section 5.3.2 and is available here: https://github.com/ElisaDiNuovo/VALICO-UD_NP_parallelReading.

conjunctions, punctuation, adverbs all have an underscore in the FEATS column. This means that, in Italian, no morphological features can be specified with this UPOS. Let us consider the word *contra* in Example 104. *Contra* is a Latin preposition meaning ‘against’ which is used in French, Spanish and German maintaining meaning and function. In Italian it corresponds to *contro*. Naively, one could think that in Italian *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.

(104) **LS:** Stà furiosa **contra** il ragazzo che non comprende quello che pasò.

Lemma: [...] contra [...]

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.

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 105, in which we do not annotate gender and number of *gle* (orthographically correct *glie*) because the referent cannot be identified with certainty.

(105) **LS:** Lui era un ragazzo buono e ardito: si è alzato, e li è seguito; quando se li ruscito, ha detto: «Lasciag**le**la»

TH: Lui era un ragazzo buono e ardito: si è alzato e li ha seguiti; quando li ha raggiunti, ha detto: «Lasciala».

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 106, 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 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.

(106) **LS:** Avevo sentito **delle parole forte**, una donna sta gridando et un uomo la portava con lui brutalmente.

*I heard **shouting**, a woman was screaming and a man was brutally carrying her.*

We marked other interesting phenomena to analyze in learner language in the MISC columns to avoid format issues. One of these is the presence of evaluative suffixes. They are 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 107 in which the learner used *canino* instead of *cagnolino*, producing a word having a different meaning ('canine tooth' instead of 'doggy').¹⁴

(107)

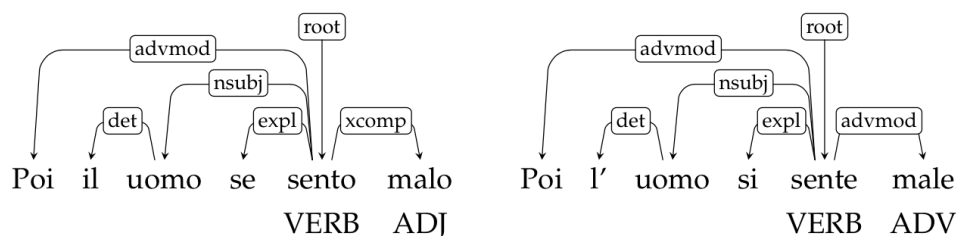
```
# sent_id = 34-07_en-3
# text = [...] tale come camminare il canino al parco. [such as walking the dog in the park.]
9      tale      tale      ADJ      A      [...]  11  mark  _
10     come      come      CONJ     CS      _      9  fixed  _
11     camminare camminare VERB     V      [...]  4  acl    _
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
```

¹⁴Note that in Italian masculine diminutives are formed adding *-ino* to the word stem. *Cane* makes an exception.

5.1.2.6 Dependency 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 focusing on morphological and distributional features, excluding semantics if necessary to prioritize syntax (Ragheb and Dickinson, 2014b).

(108)



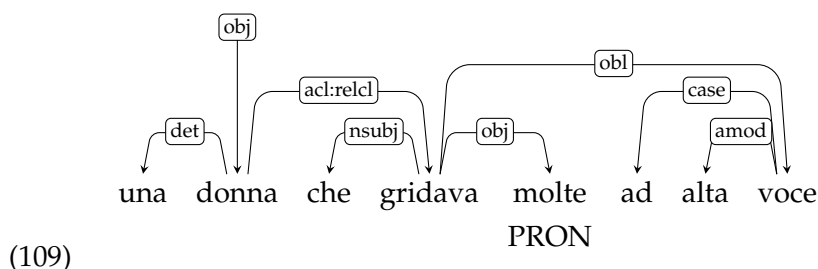
LS: Poi il uomo se sento malo.

TH: Poi l'uomo si sente male.

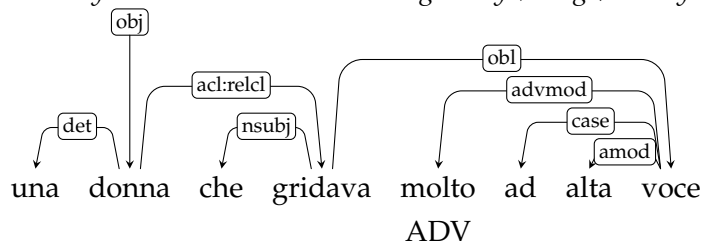
Then the man feels ill.

Annotating sentences as they are means that even a single vowel can change the syntactic tree, as shown in Example 108, in which *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*.

A similar case is the one reported in Example 109, 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*).

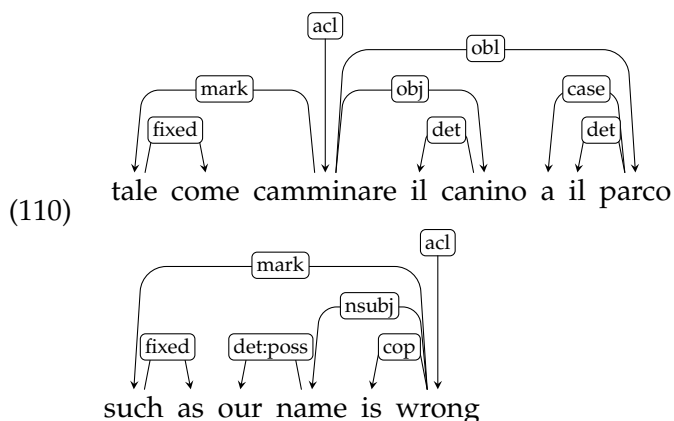


LS: All'improvviso ha sentito una donna che gridava molte ad alta voce.
Suddenly he heard a woman shouting many (things) loudly.



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 107, in which the subordinate clause is a word-for-word translation of the English structure (translation reported 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 110, we show two comparable structures retrieved from VALICO-UD and the English Web Treebank (EWT) (Silveira et al., 2014): they both feature a (non-)finite clause modifying a nominal (acl), introduced by *such as* (literally, *tale come*).

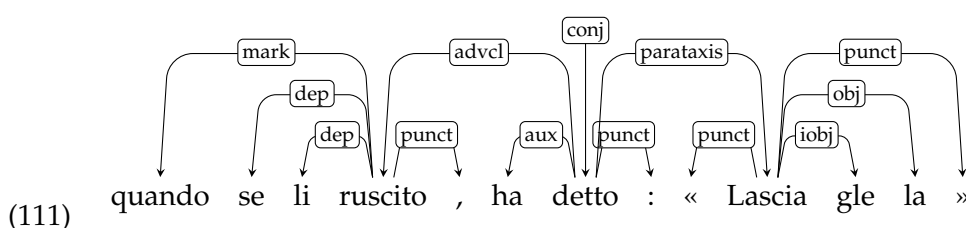


LS: Ho provato molte strategie per attrarre le ragazze tale come camminare il canino al parco.

I have tried many strategies to attract girls such as walking the doggy in the park.

EWT:¹⁵ I looked at the UEComm Master and had some comments—such as our name is wrong, [...]

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 111.



LS: Lui era un ragazzo buono e ardito: si è alzato, e li è seguito; quando **se li ruscito**, ha detto: «Lasciaglela»

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: «Lasciala».

He was a good and bold guy: he got up and followed them; when he reached them, he said: «Leave her».

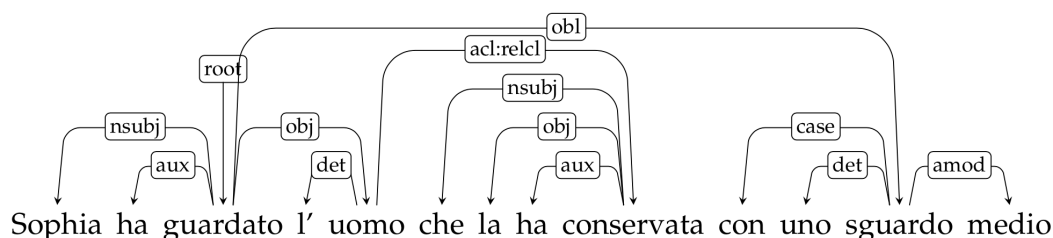
We annotated the subordinate clause starting with one conjunction—*quando*, even though *se* could also be annotated as a conjunction—and then, since it was not possible to understand the syntactic function of *se* and *li*, we used the general dependency relation *dep*. Perhaps, the verb in the subordinate 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 111, 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

¹⁵Found looking for *such as* here: http://match.grew.fr/?corpus=UD_English-EWT@2.8.

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 112. Even though the sentence makes no sense at all, for an Italian native speaker it should be easy to annotate it syntactically. 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, also because it is unlikely that someone (the man) can keep someone else (Sophia) with a medium look.



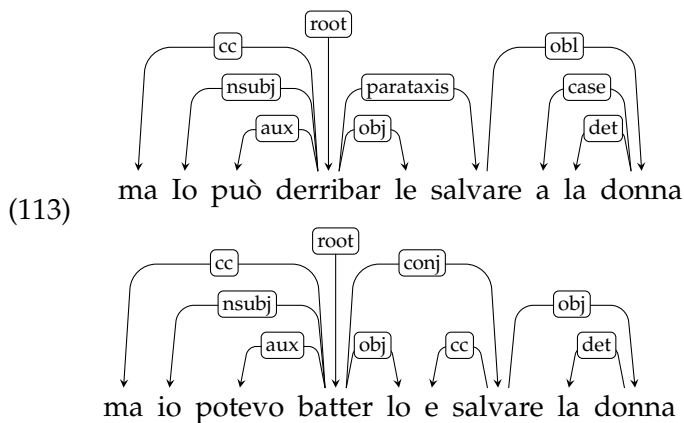
(112) **LS:** Sophia ha guardato l'uomo che la ha **conservata** con uno sguardo **medio**.

Sophia looked at the man who stored her with a medium look.

TH: Sophia ha guardato l'uomo che la ha **salvata** con uno sguardo **cattivo**.

Sophia looked at the man who saved her with a mean look.

Another challenge—although less problematic than the presence of forms which do not belong to any known language, as the case of *se li riuscito* reported in Example 111—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 113, 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 morphological information (prioritizing syntax over morphological features).



LS: Il uomo era alto, forte e molto muscoloso, 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 113 with the correspondent TH, we can obtain the interpretation of the learner' errors. In Example 113, 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.

The literal principle is used, also, in cases in which learners calque syntactic structures from languages other than Italian, but these structures can correspond to Italian marked structures, such as the case of *dislocated* relation. This relation is used in UD for "fronted or postposed elements that do not fulfill the usual core grammatical relations of a sentence."¹⁶ Dislocation can be used for emphasis, such as in the Example 114 drawn from ISDT.

(114) **Marked:** La voglia di prendersi una fetta di Bosnia Tudjman non l'ha mai nascosta, anzi, l'ha condivisa con Milosevic e insieme hanno discusso un piano di spartizione.

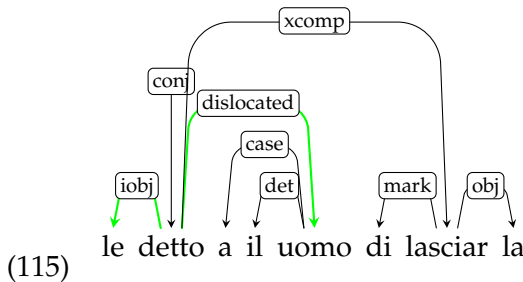
His desire to take a slice of Bosnia, Tadjman never hid it; on the contrary, he shared it with Milosevic and together they discussed a partition plan.

Unmarked: Tadjman non ha mai nascosto la voglia di prendersi una fetta di Bosnia, anzi, l'ha condivisa con Milosevic e insieme hanno discusso un piano di spartizione.

¹⁶For more details and examples see the UD web page: <https://universaldependencies.org/u/dep/dislocated.html>.

Tudjman never hid his desire to take a slice of Bosnia; on the contrary, he shared it with Milosevic and together they discussed a partition plan.

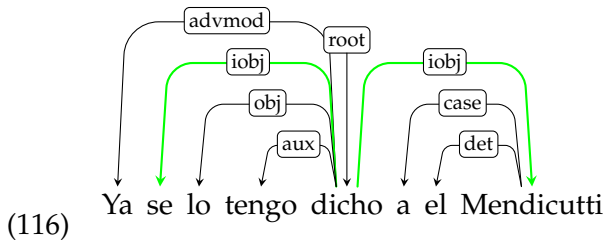
In VALICO-UD we used this relation when learners produce dislocated arguments or also to mark a Spanish structure called *clitic doubling* (Pericchi et al., 2017).¹⁷ The clitic doubling consists of a clause in which two arguments (that can be expressed as a noun phrase a stressed pronoun or a clitic pronoun) refer to the same entity and have the same syntactic function (direct or indirect object). For example, a Spanish learner produced the sentence with two indirect objects (*le* and *al uomo*, one clitic pronoun and a noun phrase, respectively) reported in Example 115. We annotated it using the dislocated relation as shown in the syntactic tree. A similar construction occurs in Example 116 drawn from AnCora treebank. Note that in the Spanish treebanks the relation *dislocated* is not used, and double arguments are used instead.



LS: MI he alzato e **le** detto **al uomo** de lasciarla perchè le faceva danno.

TH: Mi sono alzato e **gli** ho detto di lasciarla perché le faceva male.

*I got up and told **him** to leave her because he was hurting her.*



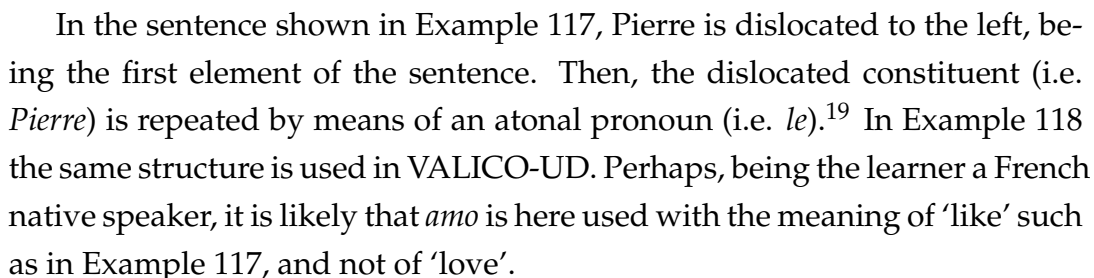
sent_id = train-s953 Ya se lo tengo dicho al Mendicutti [...]

I have already told this to Mendicutti

In Examples 115 and 116 the entities (a man and Mr Mendicutti, respectively) indicated by the clitic pronouns and the noun phrases are the same

¹⁷Clitic doubling is usually used by Italian speakers with some intransitive verbs used impersonally, such as *piacere* in the construction *a me mi piace*. This construction is declared in many grammars as a pleonasm, i.e. one of those fillers or redundancies to which the speaker's emphasis is drawn. However, the first pronoun is stronger than the second and might be considered as the theme of the sentence. For this reason, we annotated these constructions in VALICO-UD using the *dislocated* relation.

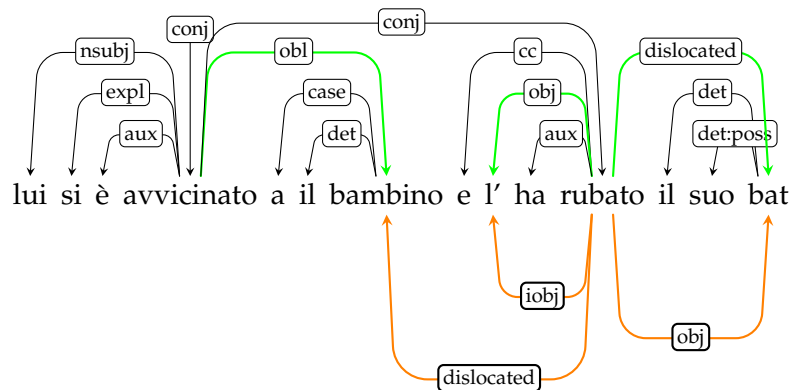
(117) *Pierre je ne l' aime pas beaucoup*
Pierre I don't like him very much



¹⁸French UD page for the relation *dislocated*: <https://universaldependencies.org/fr/dep/dislocated.html>.

¹⁹Is this phenomenon the same of polypersonalism?

the green edges represents what the learner actually produced following the target language rules. For this reason, we decided to annotate it as a direct object (i.e. the green edges) because, following the Italian elision rules (Garapa, 2009, p. 79) the indirect object cannot be elided. In this way, the LS is annotated on its own, and then, when comparing it to its TH, the error of eliding the dative clitic pronoun emerges.



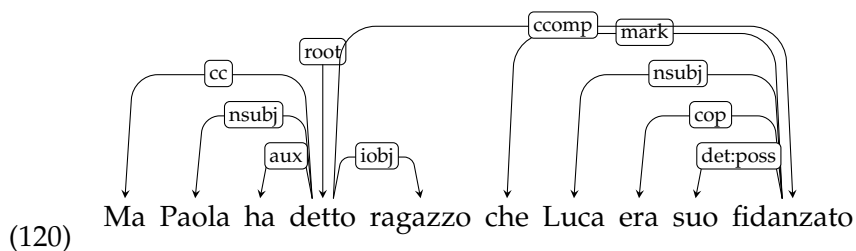
(119)

LS: [...] lui si è avvicinato al bambino e l'ha rubato il suo bat [...]

TH: [...] lui si è avvicinato al bambino e gli ha rubato la sua mazza [...]

[...] *he approached the child and stole his bat* [...]

Following the other principle, i.e. the distributional annotation of LSs, we considered the verb as a guide for the annotation. In Example 120, 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 coteext, annotating it as if the sentence ends at *ragazzo*.



(120)

LS: Ma Paola ha detto ragazzo che Luca era suo fidanzato

TH: Ma Paola ha detto al ragazzo che Luca era il suo fidanzato

But Paola told 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 allows us to be able to carry out these queries without forcing the annotation of the interlanguage using the reconstructed form.

5.1.2.7 Multi-word Expressions

For what concerns multi-word expressions (MWEs), the discussion in the literature is mostly focused on teaching and correlated to proficiency levels, making use of large-scale non annotated corpora making use of quantitative frequencies (e.g. phrase frequency, mutual information),²⁰ whilst it is under-explored how to annotate them in learner language. MWEs in computational linguistics is a hot topic and giving a thorough description of them is out of the objectives of this thesis. However, since treebanking learner corpora means to deal with a wide range of phenomena, in this subsection we describe how we approached them.

MWEs are recognized to be “a pain in the neck for NLP” (Sag et al., 2002). We marked only grammatical MWEs, i.e. fixed expressions (Sag et al., 2002, p. 4) that serve certain grammatical functions (such as adjectival, prepositional, adverbial; these are usually called *locutions* in the Romance linguistics) being consistent with what done in the UD Italian treebanks, as explained in what follow (see Section 5.1.1 to know more about the technical aspects).

In the UD framework, these fixed MWEs are annotated using a specific relation, called *fixed* and the elements composing it usually take the PoS of the grammatical function they serve. However, not all the treebanks in the repository follow this rule, confirming that MWEs are still a pain in the neck. For example, if we look at how it is annotated *due to* in the English treebanks we can notice that EWT²¹, PUD²², and LinES²³ annotate it using the *fixed* relation²⁴ and the PoS of each token composing the fixed MWE resembles the grammatical function that it fulfills, i.e. *fixed*(*due*, *to*): the head is represented by the first token from the left (i.e. *due*), and the PoS of both tokens

²⁰See, for example, the longitudinal study on multi-word expressions in second language writing carried out by Siyanova-Chanturia and Spina, 2020.

²¹EWT UD repository: https://github.com/UniversalDependencies/UD_English-EWT.

²²PUD UD repository: https://github.com/UniversalDependencies/UD_English-PUD.

²³LinES UD repository: https://github.com/UniversalDependencies/UD_English-LinES.

²⁴More information about the *fixed* relation can be found here: <https://universaldependencies.org/u/dep/fixed.html>.

is ADP (i.e. adposition). Others English treebanks, annotate the same fixed MWE differently, e.g. GUM²⁵ links the two tokens using the fixed relation, i.e. `fixed(due, to)`, but annotates *due* as ADJ (i.e. adjective) and *to* as ADP; ParTUT too annotates it as `fixed(due, to)`, but the PoS of both tokens differ from GUM, i.e. *due* is annotated as ADP, *to* as SCNJ. The learner English treebank, ESL, annotates it as in EWT, PUD and LinES. However, it is worth noticing that not all the fixed expressions are annotated using the PoS of the grammatical function they fulfill even in the same treebanks (e.g. *in order to*, *such as* are annotated as fixed but do not share the same PoS also in the treebanks in which there is this tendency).

Things become more complicated if we look for the translation of *due to* in Italian (i.e. *a causa di*) and Spanish (i.e. *debido a*) treebanks, for example. In general, the Italian treebanks are consistent in not annotating *a causa di* as a fixed MWE, even though they do annotate as fixed MWE similar structures having all the features needed to be considered as such just as *a causa di*, e.g. *fino a*. In all the treebanks in which *a causa di* is present (i.e. all except VALICO-UD), it is annotated as `case(causa,a)` and `nmod(noun,di)`. While other structures annotated as fixed MWEs, are usually annotated keeping the PoS related to each token separately (as in GUM and English ParTUT). As far as Spanish treebanks are concerned, *debido a* is found in all three treebanks available and in all it is annotated as fixed MWE but with distinct PoS per token (AnCora annotates *debido* as ADJ, GSD and PUD as VERB; the three annotates *a* as ADP).²⁶

In VALICO-UD we decided to be consistent with the Italian treebanks for what concerns the annotation of PoS, however, we annotated as fixed MWE also non-canonical structures marking also their grammatical function.

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 107

²⁵GUM UD repository: https://github.com/UniversalDependencies/UD_English-GUM.

²⁶It must be noted that the inconsistency in the annotation of *debido* in the three Spanish treebank is not a consequence of how annotating MWEs, but it is related to a common disagreement of the annotation of participial verbs that can act as adjectives.

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 121 in which the learner invented a new multi-word expression (*al invece*) functioning as an adverb.²⁷

(121) **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.

5.2 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 (Manuela Sanguinetti, PhD and myself) annotated independently a 200-sentence sample of VALICO-UD (100 LSs and the 100 corresponding THs) as described in Di Nuovo et al., 2019.²⁸

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).²⁹

The results are reported in Table 5.1, 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.

²⁷This kind of MWEs would be inevitably non included in target language datasets, which are usually used to automatically identify and extract MWEs (see for example the Italian dataset described in Masini et al., 2020).

²⁸I want to thank Manuela for her time and help.

²⁹CoNLL evaluation available here: <http://universaldependencies.org/conll18/evaluation.html>.

Set	UAS	LAS	<i>kappa</i>
LS	92.11%	88.63%	0.8988
TH	92.47%	88.88%	0.9068

TABLE 5.1: Agreement results on the sample set of both LSs and THs.

5.3 Treebank statistics

The gold standard of the treebank as described in Section 3.1.1 is composed of 72 texts: 36 learner texts consisting of 398 sentences (i.e. LSs) and the corresponding 36 target texts made of 398 THs. As far as LSs are concerned, we annotated 6,508 words corresponding to 6,784 syntactic tokens. These tokens correspond to 1,342 unique forms, lemmatized using 968 unique lemmas. In the 398 THs, instead, we annotated 6,569 words corresponding to 6,832 syntactic tokens and 1,188 forms lemmatized using 842 unique lemmas. The different amount of forms and lemmas in the two parts of the parallel treebank, i.e. LS or TH, can inform us about internal variability: in LSs there is more variety. This variability is due to spelling errors and our lemmatization rules (see Section 5.1.2.3).

5.3.1 Distribution of PoS tags

The PoS tags used in VALICO-UD follow the annotation rules of UD formalism. As introduced in Section 5.1.1, in UD word classes are encoded in two columns of the CoNLL-U format (see the fourth and fifth columns of Example 87): the fourth column of the CoNLL-U file contains language-agnostic PoS tags—here described—whilst the fifth contains language-specific tags. Since both LSs and THs are Italian varieties, these language-specific tags are the same for the two sections composing the parallel treebank. Difference can be found in the way in which words are used in the two varieties. This difference can be first identified by the language-agnostic PoS tag assigned to the token according to literal or distributional annotation (see Section 5.1.2.4).

The language-agnostic PoS tags annotated in the core section of VALICO-UD are 15 in THs and 16 in LSs. The difference concerns the use of *X* in the LSs not occurring in the THs, because it indicates wrongly split words (see Example 93 in Section 5.1.2.2).³⁰ The 16 PoS tags are distributed in the two

³⁰In Universal Dependencies the use of *X* is also conceived for foreign words whose PoS is unknown.

sections as shown in Table 5.2.

Set	ADJ	ADP	ADV	AUX	CCONJ	DET	INTJ	NOUN
LS	310	597	390	591	277	1023	12	1026
TH	311	597	390	606	281	1008	12	1030

Set	NUM	PRON	PROPN	PUNCT	SCONJ	SYM	VERB	X
LS	18	509	100	795	157	1	968	6
TH	18	518	100	838	162	1	963	0

TABLE 5.2: PoS tag distribution in the core section of VALICO-UD.

The two varieties can be distinguished also by the different morphological features that can be associated to PoS tags. Whilst, THs strictly follow standard Italian rules (e.g. adverbs cannot have gender features), LSs do not (e.g. adverbs have gender features indeed). For a complete list of features and values please refer to the statistics published in the GitHub devoted to VALICO-UD guidelines.³¹

5.3.2 Distribution of dependency relations

	Nominals	Clauses	Modifier words	Function words
Core arguments	nsubj obj iobj	csbj ccomp xcomp		
Non-core dependents	obl vocative expl dislocated	advcl	advmod discourse	aux cop mark
Nominal dependents	nmod appos nummod	acl	amod	det clf case

TABLE 5.3: Classification of UD relations.

UD relations are distinguished into core arguments, non-core dependents, nominal dependents, in turn distinguished into nominals, clauses, modifier words and function words, as shown in Table 5.3.³² To these, other five groups can be distinguished: coordination (*cc* and *conj*), MWE (*compound*,

³¹Statistics available here: <https://bit.ly/3I9d3J2>. In the UD repository only LS statistics are available.

³²This classification is taken from the official UD page: <https://universaldependencies.org/u/dep/>.

fixed and *flat*), loose (*list* and *parataxis*), special (*orphan*, *goeswith* and *reparandum*) and other (*dep*, *punct*, and *root*).

These relations can be further specified by adding the specification after a colon (e.g. *acl* → *acl:relcl*). UD currently includes 63 different relations.³³

Set	acl	acl:relcl	advcl	advmod	amod	appos
LS	28	83	110	376	179	14
TH	20	88	113	375	175	14
Set	aux	aux:pass	case	cc	ccomp	conj
LS	455	2	494	273	73	302
TH	468	2	486	276	74	318
Set	cop	csubj	det	det:poss	det:predet	discourse
LS	134	12	908	103	7	13
TH	136	12	897	104	7	14
Set	expl	expl:impers	fixed	flat:name	iobj	mark
LS	98	7	30	3	59	240
TH	101	6	29	3	67	253
Set	nmod	nsubj	nsubj:pass	nummod	obj	obl
LS	115	392	1	12	445	396
TH	118	402	1	12	444	383
Set	obl:agent	orphan	parataxis	punct	root	vocative
LS	2	2	89	795	398	5
TH	3	2	72	837	398	5
Set	xcomp					
LS	118					
TH	116					

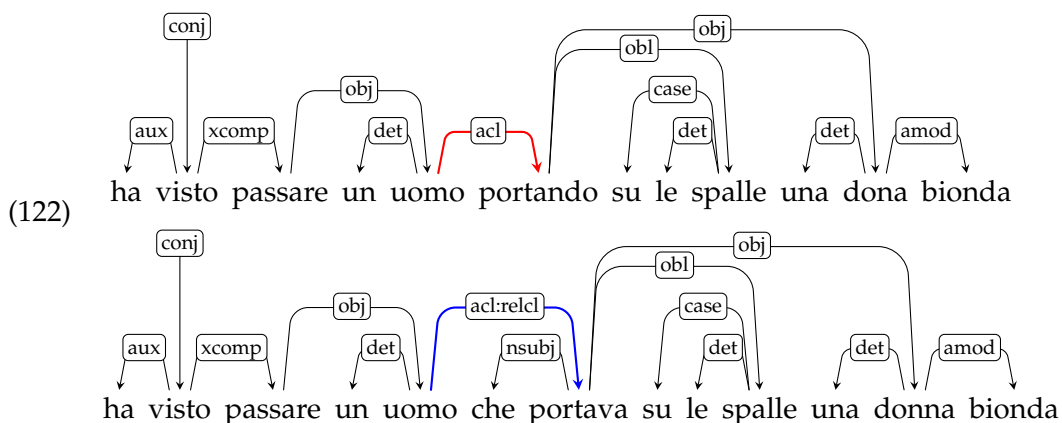
TABLE 5.4: Dependency relation distribution in the core section of VALICO-UD.

The UD relations used in the core section of VALICO-UD are 40 unique in the LSs, 37 in the THs. The difference relies in the use of three dependency relations in the LSs not used in the THs. These three relations are: *dep*, *dislocated* and *goeswith*. *Dep*, occurring twice in the core section, is used only in those cases in which it is not possible to recover the dependency relation existing between two tokens (as shown in Example 111). *Dislocated*, occurring three times in the core section, is used when learners use a marked syntactic structure which includes the dislocation of elements that exceed the usual core grammatical relations of a sentence (see Examples 114–119). Finally, *goeswith*, occurring six times in the core section, is used with wrongly

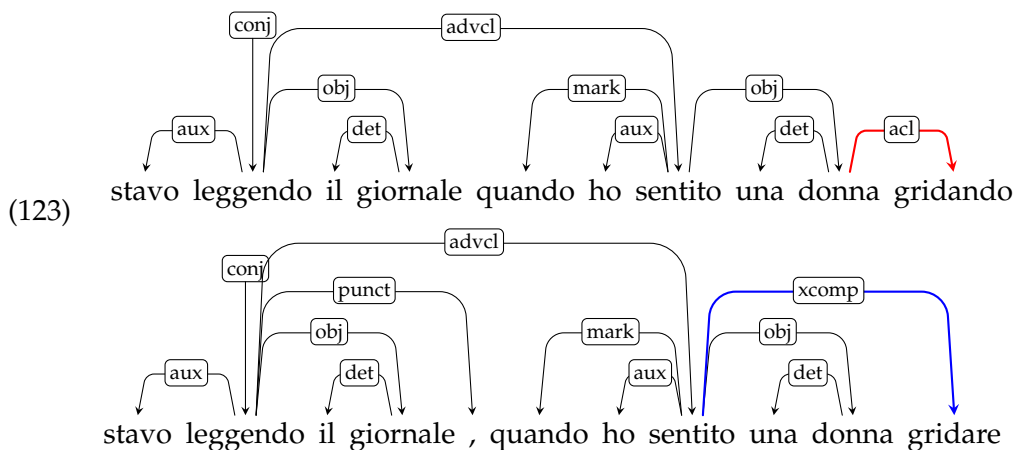
³³Complete list of universal dependency relations available here: <https://universaldependencies.org/u/dep/>.

split tokens (e.g. Example 93). The 37 relations in common are distributed as shown in Table 5.4. Relations that in absolute number are underused by learners are in blue, those overused in red.

Looking at absolute numbers, there are more adnominal clauses (i.e. *acl*) in learner than in normalized texts. This difference is due to the tendency of learners to use the gerund instead of relative clauses (i.e. *acl:relcl*, as shown in Example 122) or open clausal complements (i.e. *xcomp*, as shown in Example 123).



Yesterday in the park a man, wearing glasses, was sitting on a bench, reading the newspaper, when he saw a very strong man pass by, with a tattooed heart on his arm, CARRYING on his shoulders a blond woman who was screaming and seemed to be asking for help.



Yesterday in the park, I was sitting on a bench quietly reading the newspaper, when I heard a woman SHOUTING.

Another trend that can be seen with these absolute numbers is the learners' predilection of using parataxis more than coordinate clause (i.e. *conj*).

The overuse of parataxis is linked also to the underuse of punctuation (i.e. *punct*).

As we saw in the previous sections, the overuse of *case* can be explained by the literal annotation of argument structure (see Example 113, in which the learner used an oblique instead of a direct object, or also Example 100, in which the learner used an adposition instead of a conjunction).

From the numbers reported in the table, we can also notice how learners used very rarely passive forms (observing *aux:pass* and *nsubj:pass*) but always correctly. Differently, impersonal forms (looking at *expl:impers*) have been used slightly more, but once an impersonal form was normalized differently.

However, absolute values do not account for all the differences. In fact, if we add one determiner in a sentence and remove one in another, in absolute values we do not notice this edit. For this reason, to better describe our data, a parallel reading of the trees is required. We developed a python script to read in parallel LSs and THs and extract information about noun phrases, chosen as case study to present a possible application of the core section of the treebank.³⁴

The above mentioned script selects all the sub-trees having a noun as head and confront the lemma to align them.³⁵ Then it loops all the children of the noun, which can be linked through the following relations: *det* (i.e. determiners), *adj* (i.e. adjectives), *nmod* (i.e. nominal modifiers), *acl:relcl* (i.e. relative clauses). Using the parallel sentences of the treebank, it is possible to recover inconsistencies that cannot be automatically retrieved only using PoS tagged, non-parallel texts, e.g.:

- unnecessary inflection of gender in gender invariable adjectives (e.g. *fragila* instead of *fragile*, valid for both masculine and feminine, meaning ‘fragile’);
- wrong lexical gender with correct internal agreement (e.g. *le uccelle* instead of *gli uccelli*, with the determiner correctly agreeing with the feminine for ‘birds’, even though ‘bird’ in Italian is masculine);

³⁴Script comparing noun phrases in LSs and THs in parallel available here: https://github.com/ElisaDiNuovo/VALICO-UD_NP_parallelReading.

³⁵Since we are not using algorithms looking for e.g. longest common sequences to perform alignment, we resorted to lemmas. For this reason, the version of the core section used is one that annotates LS lemmas resorting to the normalized one. This version is provided together with the script.

- different noun phrase structures (e.g. a noun phrase containing a relative clause in turn containing a noun phrase, *l'uomo che aveva la rabbia*, meaning 'the man who had rabies', used instead of a noun phrase with a relative clause containing a copular structure *l'uomo che era arrabbiato* 'the man who was angry').

The script, focuses on determiners and adjectives, checks their position relative to the noun, reads morphological features (such as gender and number) and checks agreement.

Looking at all 36 LS texts together, as far as position is concerned, the totality of determiners are positioned to the left of the noun (994 out of 994), whilst in the THs 1 determiner is positioned to the right of the noun,³⁶ the remaining 981 to the left. Adjectives inside a noun phrase are 171 in the LSs, 173 in the THs. In both sets, adjectives are positioned to the right 120 times, to the left 51 and 53 times on LSs and THs, respectively.

The script counted 554 differences involving adjectives, determiners or nouns having a different form between LS and its TH (i.e. the difference is only in the form, the lemma coincides). In addition, the script counted 286 differences between LSs and THs that involve also the lemma. In particular 150 lemmas were different in the LSs, 136 in the THs. For example, in the LS *mi piace un uomo simpatico*, normalized into *mi piacciono gli uomini simpatici*, meaning 'I like funny men', the noun *uomo* have the same lemma of *uomini*, thus alignment is established. Then the script compares the children of *uomo* with those of *uomini*. It will find that *uomo*, *un*, and *simpatico* have a different form of the TH counterparts. This three nodes will add to the total differences based on form. Whilst only *un* and *gli* will have also a different lemma and will add to the total differences based on lemmas.

The script checks agreement twice. Once inside the LS noun phrase and then comparing it to the correspondent TH noun phrase. For example, in the noun phrase contained in the LS with sentence id 1-01_fr-3 *una donna fragile*, the internal gender and number agreement is correct, but thanks to the comparison with the TH an adjective gender agreement error is spotted, because, being *fragile* a gender invariant adjective, the gender is *None* in the TH, feminine in the LS.

³⁶It is the case of a possessive adjective which was normalised in the TH moving it to the right of the noun. LS: *Mi scusi, era la mia colpa*, TH: *Scusami, è colpa mia* ('Sorry, it's my fault').

In total, 39 agreement errors were found, 34 concerning gender, 5 concerning number. As far as gender agreement is concerned, 21 times issues were found in noun-det pairs, 13 times in noun-det-adj triplets.³⁷

As far as L1s are concerned, number agreement errors were found only in 0.8% of the analysed ES noun phrases and 2.9% of the analysed FR noun phrases. Gender agreement errors were slightly higher in DE and EN L1 texts (14.3% in both sets), than FR L1 texts (13.6%). Finally, ES was the set with the fewest gender agreement errors (7.5%).

5.4 Incremental parsing evaluation

Before using the entire resource (made of the manually corrected plus the automatically annotated LSs and THs), it is necessary to assess the acceptability of the automatic annotation (see Sections 3.1.1 and 3.1.2 for data summary).

As reported in Section 3.1, the resource has been built by applying on a subcorpus of VALICO a UDPipe model trained on ISDT and PoSTWITA UD Italian treebanks. Since we automatically annotated LS and TH sets using this model, in this chapter we incrementally evaluate the output obtained.³⁸ This is possible because part of the annotated output has been manually corrected to obtain a gold standard. Having this manually-checked gold standard, it is possible to compare it with the original output produced by the parser in order to evaluate the quality of the automatic output. This evaluation would not be possible if our guidelines were not in line with the UD treebanks used for training the model. The different decisions explained in the previous sections of this chapter are not, in our opinion, critical enough to prevent the use of models trained with these treebanks.

This evaluation allow us to quantify how much interlanguage affects parser performance. To do so, we incrementally evaluated the parser outputs obtained using different input gold information.

Usually when training a model on a text domain and testing it on another domain, a loss in performance is expected. Indeed, we expect a loss in performance when the model trained on ISDT and PoSTWITA is tested on THs, not

³⁷Note that noun-det-adj triplets are less frequent than noun-det pairs, thus if normalized, gender issues in the triplets are more frequent than in noun-det pairs.

³⁸In Section 6.1, since the objective is comparing with state-of-the-art parsers the results obtained on LSs and THs, we use models trained with the same data used by the state-of-the-art parsers taken into account.

only because it is an out-of-domain test, but also because since THs are written strictly following LSs, as explained in Section 3.2, they are not authentic in the sense defined in the recommendation on corpus and text typology EAGLES, 1996, p. 7³⁹. Subsequently, a bigger loss in performance is expected when the same model is tested on LSs. We started with THs because they are our reference for out-of-domain testing. Then we evaluated parser performance on LSs, thus the challenge is incremented by interlanguage *innovation* (see Section 5.1.2).

For testing, we used as gold standard the manually-checked core section of the treebank (detailed in Section 3.1.1).

For both varieties the evaluation is performed using as input sets of data featured by different degrees of previous validated analysis. First, we started simply applying the annotation pipeline to rough data, where texts are not divided one sentence per line or tokenized, and the annotation pipeline has to segment it into sentences, tokens, and then perform the different levels of annotation. Second, we used data where a gold segmentation has been previously applied, i.e. texts are arranged one sentence per line, hence the annotation pipeline has to tokenize, tag and parse. Third, we add also gold tokenization, i.e. texts are segmented and tokenized, the annotation pipeline has to tag and parse. Fourth, we add also gold tagging, i.e. texts not only are segmented and tokenized, but also morphologically tagged (i.e. lemmas, PoS tags and other morphological information is already given), the annotation pipeline performs only syntactic parsing.

In Table 5.5, 1 stands for the first step without gold information, 2 stands for the second in which gold segmentation is provided, 3 stands for the third step, in which gold tokenization is added, and 4 indicates the step in which also gold morphological tagging is provided. The score reported in the table is F1. The evaluation is obtained using the official evaluation script authored by Milan Straka and Martin Popel and released for the CoNLL 2018 Shared Task.⁴⁰

Observing the results at the first and second step, i.e. without any gold information and with gold segmentation, it is interesting to notice that, even

³⁹It says that all the data that is not “gathered from the genuine communications of people going about their normal business” must be considered a “*special corpus*”.

⁴⁰Downloadable from here: <https://universaldependencies.org/conll18/evaluation.html>. Note that since we are reporting evaluation using this script, also tokens and words have a different meaning than in the rest of this dissertation. In fact, in the evaluation tables, tokens stands for strings divided by spaces, whilst words stand for syntactic word.

	Tokens	Sentences	Words	UPoS	XPoS	UFeats	Lemmas	UAS	LAS
1	99.54	82.34	99.20	96.18	96.20	95.17	87.30	87.10	83.45
2	99.52	100.00	99.18	96.19	96.19	95.22	87.34	88.34	84.69
3	100.00	100.00	100.00	97.01	96.98	96.05	88.11	89.72	86.05
4	100.00	100.00	100.00	100.00	100.00	100.00	100.00	91.88	89.39

TABLE 5.5: Incremental evaluation of THs.

	Tokens	Sentences	Words	UPoS	XPoS	UFeats	Lemmas	UAS	LAS
1	99.42	79.74	98.87	94.20	94.28	93.32	85.42	83.64	78.51
2	99.38	100.00	98.83	94.22	94.31	93.39	85.39	84.75	79.61
3	100.00	100.00	100.00	95.39	95.49	94.55	86.47	86.78	81.54
4	100.00	100.00	100.00	100.00	100.00	100.00	100.00	91.13	87.96

TABLE 5.6: Incremental evaluation of LSs.

tough sentences are harder to identify in LSs than THs, in both cases, their identification does not improve drastically the syntactic annotation, which with gold segmentation improves by only one point in both sets. Note also that token identification and morphological tagging is almost the same in both steps, differing only between the two sets (LS results being lower of almost two points).

The third step, i.e. the one with gold segmentation and tokenization, as expected, improves, however only of one point, morphological tagging. Gold tokenization also has good benefit on syntactic annotation, improving for both sets, UAS and LAS by two points.

Lastly, whilst gold morphological tagging has a benefit on syntactic parsing of maximum three points in the TH set, the parsing in the LS set highly benefits from this information, which is incremented of more than six points. We explain it, because in LSs there are occurrences of words that, even displaying characteristics of some PoS, distributionally they behave like another, as we have seen in various examples of the current chapter.

5.4.1 Error analysis

In order to get an idea of the annotation and correction effort on both parts of the treebank, in this section we report gold, predicted and correct values, and F1 score for each PoS tag and dependency relation used in the core section of the treebank when compared to the automatic output obtained using as input presegmented texts (i.e. the original automatic output we used to obtain our gold).

To obtain these results, we used UDAPI (Popel, Zabokrtský, and Vojtek, 2017). UDAPI is an API and framework, available for Python, Perl and Java, for processing UD that can be used for a wide range of use cases (e.g. tree viewer, format conversion, querying, automatic parsing, evaluation). In particular, we used the function *F1* of the block *eval*—used for computing the similarity between two UD trees with F1—setting attributes and focus in order to obtain F1 for each UPoS and dependency relation individually.⁴¹

5.4.1.1 Evaluation of automatic PoS tagging

In Table 5.7 we report the UDPIPE model evaluation per individual PoS tag.

It can be noted that for almost all PoS the difference between LSs and THs is that on THs the model achieves a higher F1 score of 1–2 points. For these PoS in which the model achieves a F1 similar between LSs and THs, it is possible to state that LSs features do not affect the parser performance.

	ADJ	ADP	ADV	AUX	CCONJ	DET	INTJ	NOUN
LS gold	310	597	390	591	277	1023	12	1030
Predicted	313	617	387	585	267	1011	7	1052
Correct	262	573	348	562	257	989	3	978
F1	84.11%	94.40	89.58%	95.58%	94.49%	97.25%	31.58%	93.95%
TH gold	311	597	390	606	281	1008	12	1026
Predicted	312	619	404	612	267	996	8	1024
Correct	280	586	359	598	265	982	5	981
F1	90.03%	96.38%	90.43%	98.19%	96.72%	98.00%	50.00%	95.71%

	NUM	PRON	PROPN	PUNCT	SCONJ	SYM	VERB	X
LS gold	18	509	100	795	157	1	968	6
Predicted	15	498	120	783	137	1	954	10
Correct	15	464	100	408	124	1	904	0
F1	90.91%	92.15%	90.90%	51.74	84.35%	100.00%	94.07%	0.00%
TH gold	18	518	100	838	162	1	963	0
Predicted	16	510	118	826	150	1	945	9
Correct	15	500	100	447	136	1	916	0
F1	88.24%	97.28%	91.75%	53.76%	87.18%	100.00%	96.02%	0.00%

TABLE 5.7: Evaluation of automatic UPoS tagging.

In both LSs and THs high F1 scores (i.e. F1 equals or higher of 90%) are obtained in 10 out of 16 PoS tags: adpositions (ADP), adverbs (ADV), auxiliaries (AUX), coordinate conjunctions (CCONJ), determiners (DET), nouns (NOUN), numerals (NUM), pronouns (PRON), proper nouns (PROPN), symbols (SYM) and verbs (VERB).

⁴¹We used the same function—but without specifying attributes and focus—in Section 6.2 to quantify the difference between a LS tree and its correspondent TH tree.

In these PoS the difference is, as stated, of only 1–2 points, except for NUM and PRON. As far as NUM is concerned, the system achieves a lower performance on THs than LSs, in which the noun *urla* has been tagged as numeral in the THs, whilst in the LSs it corresponds to *parole forte* which has been annotated correctly as NOUN and ADJ. As far as PRON is concerned, the difference between LSs and THs is greater than 5 points (with better performance in THs). PRON in UD Italian treebanks can be further specified as clitic (PC), demonstrative (PD), personal (PE), indefinite (PI), possessive (PP), interrogative (PQ), and relative pronouns (PR). PQ are the most problematic both in LSs than THs (18.18% and 42.86%, respectively), PE the least (98.80% and 100.00%, respectively). The model achieves a F1 of 88.51% and 92.31% in identifying PR in LSs and THs. For the other pronouns the F1 is above 90% in both sets.⁴²

The PoS tags in which the model achieves the lowest F1 scores are adjectives (ADJ), interjections (INTJ), punctuation (PUNCT), subordinate conjunctions (SCONJ) and X (i.e. indicates wrongly split words). As far as ADJ is concerned, the system achieves an F1 in LSs almost 6 points lower than in THs. In total 24 errors are marked affecting adjectives but they do not seem to correlate with the parser's errors. In fact, nouns and verbs are wrongly mistaken for adjectives also in THs. In particular, the verb *gridare* and its indicative imperfect forms *gridav-a/o* are a problem for the model in both sets. Capital letters, instead, seem to correlate with parser performance. Among the 72 texts of the gold section, there are one LS text and its TH which are written entirely in capitals. Capital letters negatively affect ADJ, NOUN and VERB, they are not an issue for DET, AUX and ADP. In these texts, in the gold data there are 9 adjectives per set,⁴³ of which only 2 are correctly tagged in the LSs (predicted 10), 0 in the THs (predicted 6). The models dealing with capitalized words tends to mistake nouns and verbs (not the participle but the imperfect) for adjectives.⁴⁴ Thus, the model obtains a low F1 also for nouns and verbs (64.00% and 78.79%, respectively on LSs; 68.09% and 86.49% on THs). As far

⁴²PP are absent in both sets, but the model incorrectly predicts one in THs in the sentence, reported in Section 4.2.5.3, *lo faceva per il mio proprio bene—per il mio bene* is more common. *Bene* is annotated as ADV, thus *mio* as PRON. *Bene*, in the Italian UD treebanks used for training, occurs 243 times as adverb, 39 as noun; perhaps in an uncommon structure, the parser opted for the most common output.

⁴³These are: **LS** = **TH**: SPLENDEnte, CONCENTRATO, TRANQUILLO, SCURI, CORTI, NERO (**TH** \approx NERA), NERE, MUSCOLOSO, STRANA.

⁴⁴**LS**: SOLE, CAPELLI, OCCHIALI, SCARPE, UOMO, URLAVA, GRIDAVA, SCENEGGIATURA; **TH**: SOLE, CAPELLI, UOMO, URLAVA, GRIDAVA, SCENEGGIATURA.

as INTJ and PUNCT are concerned, the model achieves a F1 around 50% in both sets, except for INTJ in LSs, with $F1 = 31.58\%$. These poor results are due to our choice of annotating also *mamma mia* or *va bene* and *aiuto* (when used to call for assistance) as interjections, whilst the model is not trained on similar data. As far as PUNCT is concerned, F1 is pulled down by the poor results in both sets in identifying sentence final punctuation (FS). This might be due to our segmentation choices, which, despite are in line with PoSTWITA, they do not coincide with ISDT segmentation rules (see Section 5.1.2.1). As far as X is concerned, generally in all UD treebanks wrongly split tokens are rare phenomena and it is mostly used for tagging foreign words. In particular in ISDT it is usually associated to foreign forms, in PoSTWITA also to truncated tweet-final words. These uses are far from what it is found in VALICO-UD, even though in UD formalism this usage is allowed.

5.4.1.2 Evaluation of automatic parsing

In Tables 5.8 and 5.9 we report the UDPIPE model evaluation per individual dependency relation.

As a general trend, the parser performs better in identifying highly frequent dependency relations. It performs worse in identifying infrequent relations (i.e. *aux:pass*, *nsubj:pass*, *csubj*, *discourse*, *expl:impers*, *obl:agent*, *orphan* and *vocative*). Although infrequent, the system performs better with *det:predet*, *flat:name*, *nummod* because their dependency is less ambiguous than the other relations (e.g. difference between *expl* and *expl:impers* is covert, thus ambiguous).

In both LS and TH sets, the model obtains almost the same F1 scores on identifying *advmod*, *det:poss*, *det:predet*, *nummod*, *orphan*, and *punct*. Similar F1 scores (i.e. about 3–4 point difference) on *aux*, *case*, *cc*, *conj*, *det*, *expl:impers*, *mark*, *nmod*, *obl*, *parataxis*, and *vocative*.

The parser, as expected, achieves a lower F1 score in the LS rather than the TH set, except for *vocative*, and hence it is confirmed that interlanguage interferes with its performance, although the difference is not as big as it could be expected. In fact, parser performance is comparable in both sets. This could be due to our choice to literally annotate learner language, which is an advantage for the parser.

As far as *ac1* is concerned, the low result is in part explained by the original use of the gerund by learners (see Examples 122 and 123). This innovative way—innovative for Italian, not for the other considered L1s—of using the

	acl	acl:relcl	advcl	advmod	amod	appos
LS gold	28	83	110	376	179	14
Predicted	26	90	131	364	204	4
Correct	11	73	80	320	157	3
F1	40.74%	84.39%	66.39%	86.49%	81.98%	33.33%
TH gold	20	88	113	375	175	14
Predicted	23	93	138	374	198	5
Correct	13	82	94	327	164	4
F1	60.47%	90.61%	74.90%	87.32%	87.94%	42.11%

	aux	aux:pass	case	cc	ccomp	conj
LS gold	455	2	494	273	73	302
Predicted	443	14	523	266	90	303
Correct	418	1	472	255	53	264
F1	93.10%	12.50%	92.82%	94.62%	65.03%	87.27%
TH gold	468	2	486	276	74	318
Predicted	462	14	512	266	75	304
Correct	451	2	476	264	61	280
F1	96.99%	25.00%	95.39%	97.42%	81.88%	90.03%

	cop	csbj	det	det:poss	det:predet	discourse
LS gold	134	12	908	103	7	13
Predicted	128	8	900	102	8	8
Correct	111	4	864	100	7	5
F1	84.73%	40.00%	95.58%	97.56%	93.33%	47.62%
TH gold	136	12	897	104	7	14
Predicted	136	5	885	103	8	8
Correct	122	5	869	101	7	6
F1	89.71%	58.82%	97.53%	97.58%	93.33%	54.55%

TABLE 5.8: Evaluation of automatic parsing per dependency relation (Part 1).

gerund in Italian corresponds to the 32.14% of all gold *acl* and none of them have been correctly annotated. Instead, the system annotated them as *advcl* or *acl:relcl*.

In line with higher misattribution of ADJ and INTJ in LSs than THs, also *amod* and *discourse* have a lower F1 score in LSs than THs. In addition, also *fixed* can be influenced by the misattribution of INTJ in the examples discussed in the previous subsection (i.e. *mamma mia* and *va bene*). However, *fixed* relation in LSs is also affected by learners' creativity with non-canonical *fixed* structures (see Section 5.1.2.7).

Similarly, *flat:name* is also directly influenced in misattributions of PROP. In particular, in a sentence beginning with the verb *strappare* in its indicative imperfect third person singular form (i.e. *Strappava*) followed by the name *Marco*, the parser annotated it as a PROP, probably because of the non-canonical usage of the verb (TH: *L'ha strappata da Marco*, meaning 'he tore her

away from Marco'), used as transitive verb (divalent, requiring subject and direct object) instead of ditransitive (trivalent, requiring also an oblique or indirect object).

	expl	expl:impers	fixed	flat:name	iobj	mark
LS gold	98	7	30	3	59	240
Predicted	72	8	22	6	69	229
Correct	63	3	7	3	45	211
F1	74.12%	40.00%	26.92%	66.67%	70.31%	89.98%
TH gold	101	6	29	3	67	253
Predicted	77	8	26	5	75	255
Correct	71	3	10	3	56	239
F1	79.78%	42.86%	36.36%	75.00%	78.87%	94.09%

	nmod	nsubj	nsubj:pass	nummod	obj	obl
LS gold	115	392	1	12	445	396
Predicted	157	408	12	12	456	363
Correct	97	332	0	10	373	311
F1	71.06%	83.00%	0.00%	83.33%	82.80%	81.95%
TH gold	118	402	1	12	444	383
Predicted	143	405	14	12	449	358
Correct	98	361	1	10	395	317
F1	75.10%	89.47%	13.33%	83.33%	88.47%	85.56%

	obl:agent	orphan	parataxis	punct	vocative	xcomp
LS gold	2	2	89	795	5	118
Predicted	3	0	35	783	1	99
Correct	1	0	20	439	1	82
F1	40.00%	0.00%	32.26%	55.68%	33.33%	75.58%
TH gold	3	2	72	837	5	116
Predicted	6	0	42	826	2	101
Correct	3	0	20	470	1	92
F1	66.67%	0.00%	35.09%	56.52%	28.57%	84.79%

	root
LS gold	398
Predicted	398
Correct	340
F1	85.43%
TH gold	398
Predicted	398
Correct	352
F1	88.44%

TABLE 5.9: Evaluation of automatic parsing per dependency relation (Part 2).

As far as appos is concerned, precision is high (75%), recall low (21.43%). As stated in Ahrenberg, 2019, this relation can be easily be mistaken with parataxis or conj, because of its loose definition, according to the author. This could be the cause of the poor parser's performance with appos and parataxis.

As far as *aux:pass* is concerned, recall is lower in LSs than THs, and precision is 100% in THs. In LSs and THs this relation is over-predicted by the system (i.e. only 2 gold instances, predicted 14 times). The difference with THs is only one more correct prediction. *nsubj:pass* is in line with *aux:pass*, being again over-predicted by the system, in both LSs and THs.

As far as *obl:agent* is concerned, it is a case of infrequent dependency relation, over-predicted in THs, but achieving 100% recall. Semantic information would be necessary to solve some ambiguities in identifying this argument.

As far as *csubj* is concerned, recall is lower in LSs than THs, and precision is 100% in THs. Differently than with passive auxiliaries, the system performs better on clausal subjects, even though they are under-predicted (i.e. 12 gold instances, 8 predicted, only 4 correctly in LSs, 5 corrected predictions out of 5 in THs).

As far as *ccomp* and *xcomp* are concerned, their annotation is sometimes interchanged even dealing with standard texts. However, when dealing with interlanguage, one error can be enough to mislead the automatic annotation. For example in *ha cominciato di leggeri il giornale* instead of *ha cominciato a leggere il giornale* ‘he started **reading** the newspaper’, a spelling error plus an adposition replacement caused a parser error. In our opinion the error is due to the wrong spelling of *leggere* because it caused a wrong PoS attribution (NOUN instead of VERB).

As far as *iobj* is concerned, in both sets the parser achieves a F1 below 80%. In both sets *iobj* relation is over-predicted, but in LSs precision is lower. It could also be due to some annotation choices, for example in Example 120 in which we give priority to semantics (and valency grammar) rather than syntax. The system, instead, in the mentioned example, fails annotating it as a direct object. However, as far as *obj* is concerned, in both sets the parser achieves a F1 higher than 80%. For both relations, interlanguage adds difficulty in an already challenging task.

expl is another relation which is under-predicted by the parser, however with high precision in both sets, though higher in THs.

As far as *cop* and *is* is concerned, the system wrongly attributes this relation to verbs that cannot be copulas in Italian, e.g. to *venire* in the sentence *Al provviso venne un tizio con un viso furiosissimo!* and to its TH: *All'improvviso venne un tizio con un viso furiosissimo!*, meaning ‘all of a sudden a guy came with a furious face’, probably because of the use of the *passato remoto*. In case

of dependency relations that can be assigned only to a closed set of words, such as copula, a spelling error can be detrimental. For example *eranno* instead of *erano* in the sentence *Ieri al parco ho visto due ragazzi che non erano molto contenti* ('Yesterday in the park I saw two guys who were not very happy').

Last but not least, also *nsubj* is affected by interlanguage features. Although the system predicts more *nsubj* in LSs than THs (with fewer gold labels in LSs than THs), in LS precision is lower than in THs.

Chapter 6

Quantitative data exploration

In this chapter we report on three quantitative metrics exploited to explore the treebank for assessing the quality of the data and for better understanding the role that this resource can play in the future in the context of computational linguistics.

It is important to notice that the philosophy that guides the construction of a treebank like VALICO-UD is to deepen the study of the learners' interlanguage by applying computational formalisms and tools. In line with this philosophy, data exploration is carried out to obtain quantitative information for describing interlanguage when compared to standard language (i.e. first quantitative metric), or when compared to THs (i.e. second and third quantitative metrics). Therefore, the construction of a gold standard for VALICO-UD does not have the immediate purpose of training a parser that works for learner language. However, this is an interesting side effect, if we think that, for example, it could be useful for conversational systems that more and more often have to deal with sentences produced by non-native speakers or have educational purposes.

Firstly, we resort to the gold standard to compare the performance obtained by two different UDPIPE models on both LSs and THs with state-of-the-art parsers evaluated in-domain. Secondly, we exploit the distance between LSs and THs, relying only on strings. Thirdly, we exploit the similarity of LS and TH trees. These three metrics are described singularly in the next three subsections.

6.1 Quantifying the parser performance

In Section 5.4 we reported an incremental parsing evaluation of the model used to obtain the first automatically annotated version of the treebank—i.e. the UDPIPE model trained on ISDT and PoSTWITA UD treebanks. Also

in this section, we exploit the manually-corrected section of the treebank as gold standard to evaluate the automatic output produced by two different UDPIPE models. This time we use two models trained separately on ISDT and PoSTWITA, in order to be comparable with state-of-the-art parsers that competed to the CoNLL 2018 Shared Task (Zeman et al., 2018).

To be comparable to the models evaluated in the CoNLL 2018 Shared Task, we report the F1 on Labeled Attachment Score (LAS) computed with the official CoNLL-18 evaluation script for evaluating the model when tested on THs and LSs, as done also in the previous chapter.

We tested separately on LSs and THs the model trained on ISDT, then also that trained on PoSTWITA. As input texts we used presegmented sentences. The results are reported on Table 6.1.

Set	ISDT	PoSTWITA
LS	85.34	83.93
TH	88.25	85.46

TABLE 6.1: UDPIPE models' LAS when trained on ISDT and PoSTWITA and tested on LSs and THs.

The two UDPIPE models trained separately on ISDT and PoSTWITA achieved, as expected, an F1 higher on THs (88.25 and 85.46, respectively) than LSs (85.34 and 83.93, respectively). However, the results achieved on LSs are better than expectations. The two results achieved on THs and LSs are, indeed, lower than the best result achieved by the best performing parser, HIT-SCIR (Che et al., 2018), that participated to the CoNLL 2018 Shared Task trained and tested on ISDT obtaining an F1 of 92.00. However, they are higher than the best result achieved by the participating parser when trained and tested on PoSTWITA (HIT-SCIR 79.39).¹ These better results achieved when out-of-domain testing a model trained on PoSTWITA could be explained by the small sample test and the simplified syntactic structures that learners use. From these results, we can draw the conclusion that a model trained on native texts and tested on LSs achieve comparable results than parsers trained and tested in-domain.

Once verified the acceptability of the automatic annotation, we exploited two tools to quantitatively evaluate how much LSs and THs differ (or are similar) and to see if this feature correlates with the learners' stage of inter-language. We exploited two metrics to compute a distance and similarity

¹Note that the size of PoSTWITA is considerably smaller than ISDT, so a lower parsing performance is expected.

metric, respectively, for comparing each LS to its corresponding TH. Then we grouped LSs and THs according to the year of study of Italian in order to classify learners according to two classes, corresponding to two different populations of learners, i.e. initial and advanced learners. In other words, we want to investigate if these two metrics—one based on character edit distance and the other on tree edit distance—might be exploited together to preliminary assess language proficiency. We decided to use these two metrics because they capture different features of LSs when compared to their THs, so we used both of them to observe the difference between the two classes of learners from two different points of view.

6.2 Quantifying learners' proficiency through machine translation metrics

We could consider our LS and corresponding TH as the system output and the reference translation, respectively. In this way we could apply the most suitable metric used in Machine Translation (MT) to our parallel treebank to analyse it. As far as VALICO-UD is concerned, the goal consists in the evaluation of a LS with respect to its reference TH, with a variable degree of acceptability. These machine translation evaluation metrics, then, can be exploited in our parallel treebank to highlight different features of learners' language. One of the possible features, for example, is learner proficiency defined by the year of study of Italian.

The quality of MT can be evaluated using human and automatic metrics (see Chatzikoumi, 2020 for a comprehensive review). As far as automatic metrics are concerned, although BLEU (Bilingual Evaluation Understudy) (Papineni et al., 2002) is the current standard for automatic MT evaluation, it is not the most suitable metric for the comparison of LSs with THs. The reasons are manifold: First, we only have one reference per each output (i.e. one TH per each LS), whilst a key characteristics of BLEU is its direct exploitation of multiple references; Second, being a precision-oriented metric, it only measures what is correct in the output and not to what extent the content of the reference is reproduced; Third, BLEU was designed for large test corpora, and it has been shown to work best when the scores are averaged over many sentences (our treebank is really small compared to the number of parallel sentences that are used to train and evaluate machine translation

systems); Fourth, BLEU scores of individual sentences are not deemed trustworthy (Dorr et al., 2011). In contrast, TER (Translation Error Rate) (Snover et al., 2006), being a variant of WER (Word Error Rate)—i.e. a metrics that compute the Levenshtein distance (Levenshtein 1966) between the words of the system output (i.e. the LS) and the words of the reference translation (i.e. the TH) divided by the length of the reference translation (i.e. the TH)—overcoming the word-ordering limitation of WER, is more suitable to be applied to our treebank (cfr. HTER, Human-mediated Translation Error Rate (Snover et al., 2006)), since we compare two sentences in which the reference is created on the system output to be the closest as possible (see Chapter 3.2).

Furthermore, since we are working with a treebank, we can exploit MT metrics that take the syntactic annotation into account (Chatzikoumi, 2020, p. 8). In particular, we used the tool UDAPI, since it was created specifically to work with UD-annotated texts.

In the next subsections we will report on the results of two different metrics, one exploiting string distance (TER), and the other tree similarity (UDAPI).

6.2.1 Quantifying string distance

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), used in machine translation to measure the number of edits required to change a system output (our LSs) into one of the references (our THs). Its value goes from 0, meaning that the two compared sentences are the same, and 1, meaning that the two compared sentences are completely different. In brief, the lower the score, the better. In Example 124, we show two sentences with a TER value of 0.375.

(124) **LS:** Ieri al parco è successo qualcosa stana.

TH: Ieri al parco è successo qualcosa di strano.

Yesterday in the park something strange happened.

Once computed this metric on the LS-TH-text pair, we compared the results obtained for the two classes of texts/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 6.2. The texts used for the data exploration are in bold.

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

TABLE 6.2: Texts and metadata of the silver standard. Texts selected for the exploration in bold. The question mark indicates that the year of study is *not known*.

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 forth year of study of Italian. Hence, we obtained two classes of texts that we used in order to verify our hypothesis.

For the group of initial learners the mean obtained is 0.29 (standard deviation = 0.10, minimum = 0.11, maximum = 0.50). For the group of advanced learners the mean obtained is 0.20 (standard deviation = 0.10, minimum = 0.04, maximum = 0.39).

Then we wanted to test if the TER values obtained for the two populations were 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 classes, as shown in Figure 6.1. Each column in the histogram indicates the frequency (y axis) of the TER values 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.

Since the data, as shown by the histograms, have a more or less a Gaussian distribution, with no outliers, and the standard deviations are the same in the two classes, 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.

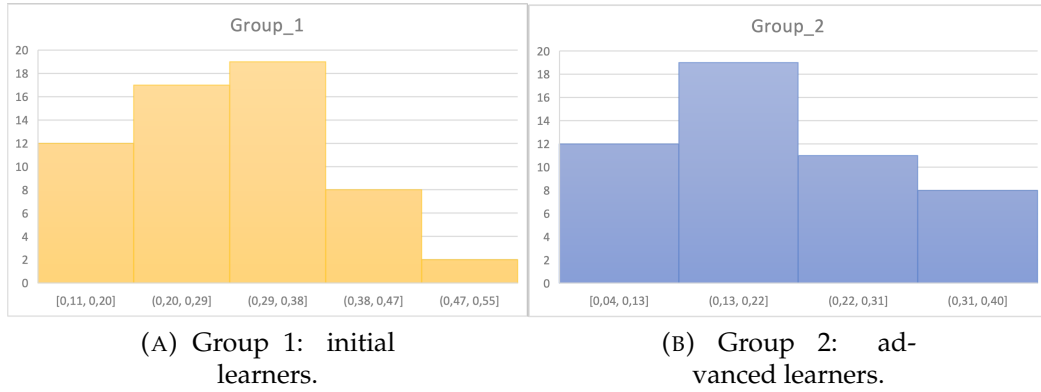
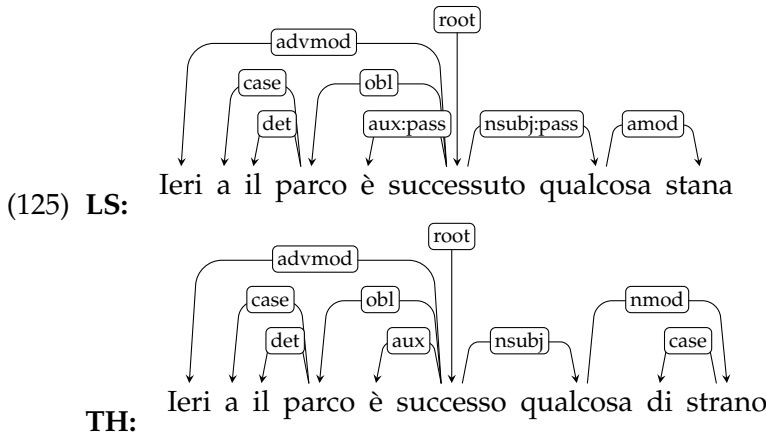


FIGURE 6.1: Histograms of the TER values obtained for each text of the two classes, initial (Group 1) and advanced (Group 2) learners of Italian.

This confirms that there is an extremely statistically significant difference between the initial and advanced groups of learners when described using the TER metric between LSs and THs.

6.2.2 Quantifying tree distance

To assess if the two groups of learners are different also when comparing the syntactic trees, we exploited UDAPI (Popel, Zabokrtský, and Vojtek, 2017). In particular, we used the function *F1* of the block *eval* that is used for computing the similarity between two UD trees with F1. Its value goes from 0% to 100%, with 100% meaning that the two compared sentences are the same. Thus, contrarily to TER, the higher the score, the better. In Example 125 we show the trees of the two sentences reported in Example 124. Note that we are using the automatically parsed trees. The obtained F1 is 73.68%.



For the group of initial learners the mean obtained is 0.80 (standard deviation = 0.07, minimum = 65%, maximum = 93%) For the group of advanced

learners the mean obtained is 0.86 (standard deviation 0.07, minimum = 69%, maximum = 98%).

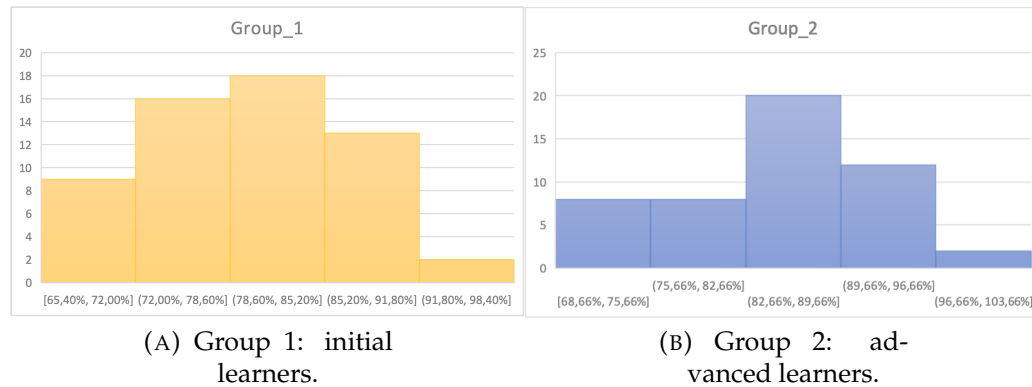


FIGURE 6.2: Histograms of the F1 values obtained for each text of the two classes, initial (Group 1) and advanced (Group 2) learners of Italian.

Again, to test it statistically, we visualized in two histograms the F1 values obtained on each text of the two classes, as shown in Figure 6.2.

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.

In this chapter, we reported on two experiments carried out to show how metrics used for other tasks, MT evaluation in this case, can be used as indicators of language development. Both string distance and tree similarity proved to be suitable indicators to track interlanguage development, as proven by the extremely statistically significant p values obtained. These results are promising also for another reason: they extrinsically assess and validate the consistency of the normalization of the LSs, i.e. the generation of valid THs.

Chapter 7

Conclusions

This thesis has presented the design and development of a novel resource which is an Italian learner treebank, named VALICO-UD. We began by introducing the motivations that led us to create this resource, and presenting an overview of related literature. Then, in Chapter 3 we presented its data (Learner Sentences, LSs) and the principles we followed for creating corresponding normalized sentences (Target Hypotheses, THs) to subsequently build a parallel corpus. We decided to use the VALICO texts because, being elicited by comic strips, error detection and normalisation is more constrained than in other text types. Thus, despite generated by a single annotator, THs have benefited by the constrained text context. In addition, the annotator followed well-defined principles in order to ensure as much as possible an objective normalization which does not rely solely on the single annotator idiolect.

In the following two chapters, we addressed the challenges related to the annotation of the errors occurring in the LSs and to the development of a treebank on the top of this parallel corpus, especially focusing on a core section of the whole resource in which data is balanced according to the learner native language and proficiency in Italian.

Indeed, in Chapter 4 we described in detail the error taxonomy implemented in the core section of the treebank. Inspired by the tagset used for annotating errors in the Cambridge Learner Corpus and implemented using XML-based tags, error annotation exploits surface edit taxonomy plus linguistic category taxonomy methods to describe interlanguage. After presenting error statistics, we reported on three inter annotator agreement experiments aimed at answering three research questions:

- Is error identification more reliable if linguistic and extra-linguistic context is given?

- When different annotators agree on the presence of an error, do they agree also on its normalization?
- Is error annotation more reliable with explicit target hypotheses provided?

The first question can be answered affirmatively. Looking at error identification we obtained a κ value which equals to 0.82. A similar result was achieved by Köhn and Köhn, 2018 who also used a picture-elicited corpus. As also highlighted by previous research, disagreement emerged for lexical but also grammatical issues (Rosen et al., 2014; Del Río Gayo and Mendes, 2018b).

Concerning the second question, the answer is positive, but agreement between annotators was lower ($\kappa = 0.69$) than error identification agreement. This is due to the non-deterministic nature of this task. However, analysing the sources of disagreement, what emerged is that more than 40% of disagreement was caused by mistakes (due to distraction or to format). Perhaps, a second round of annotation is necessary in tasks like this which require high level of concentration and specific skills.

Finally, error annotation with provided THs is more reliable than without them, but they are not enough. In fact, as happened for the two previous experiments, a high percentage of apparent disagreement was present and annotators initially reached only a moderate agreement ($\kappa = 0.50$). This was caused both by human distraction than by the complexity of the tagset used. However, after agreement revision, they achieved perfect agreement ($\kappa = 0.95$) confirming that THs indeed ensure reliability. The remaining disagreement, as emerged from the qualitative analysis, is due to error nature and to the need to have highly specific guidelines.

In Chapter 5 we described the challenges in treebanking a learner corpus using a framework developed with multilingualism in mind, but not specifically for learner language. We focused on each layer of annotation, starting from sentence segmentation, arriving to syntactic parsing, passing through tokenization, lemmatization, PoS tagging, and the annotation of morphological features. Differently from other projects in which new annotation schemes not comparable to other language varieties are developed, in VALICO-UD we decided to use Universal Dependencies (UD) formalism to have a resource that can be analysed contrastively with the other treebanks

available in the UD repository (not only standard language but also interlanguage varieties). In this chapter we answered to two research questions:

- Is Universal Dependencies (UD) formalism adaptable to L2 Italian? Considering that UD has been successfully applied on standard varieties of Italian (e.g. legal texts, newspapers and Wikipedia) and on social media texts (i.e. Twitter), we want to see if the format can address also the challenges of interlanguage, by testing its repertoire of labels against VALICO data.
- What is the performance loss of a parser trained on standard texts when applied to learner language? That is, how much interlanguage features affect parser performance? Systematic differences between native and learner language can emerge when we apply on the latter automatic annotation tools usually developed for the former.

Thanks to the versatility of UD, this formalism proved to be perfectly adaptable to L2 Italian, allowing in depth description of interlanguage phenomena. We did not need to create new labels for dependency relations but, when necessary, we resorted to the tenth column of the CoNLL-U file (i.e. MISC) to add information which is usually not annotated in UD treebanks (e.g. the type of fixed multi-word expression).

As far as performance loss is concerned, we incrementally evaluated the model we used to obtain the first automatically annotated draft of the treebank and to quantitatively describe interlanguage impact on parser performance. This proved to be mostly in line with an out-of-domain evaluation (there are not major differences between learner sentences and target hypotheses), although slightly more challenging. We believe that these high results are due to our choice to literally annotate learner sentences prioritizing syntax over meaning to deal with the majority of learner features. Performing experiments where gold segmentation and tokenization were provided, we have seen that they did not improve significantly parsing, whilst gold morphological tagging did.

For further investigating the topics related to parsing, we also presented a quantitative error analysis. It provides in detail quantitative considerations on each PoS tag and dependency relation. What emerges is that automatic analysis of both learner and out-of-domain standard language is not reliable

to describe infrequent parts of speech or structures, but taken as a whole automatic annotation of learner language can be used just like the way that out-of-domain applications are exploited in Natural Language Processing tasks.

The final experiments of this thesis—described in Chapter 6—explored some possible applications of the treebank, using both gold and silver standard parts of the resource. We used the gold standard data to compare two UDPIPE model, trained separately on ISDT and PoSTWITA, with state-of-the-art parser results trained on the same resources. Then, using the silver standard, we exploited two quantitative machine translation metrics—measuring string and tree distance—to answer the following question:

- Can the similarity between LSs and THs—expressed using quantitative machine translation evaluation metrics such as Translation Error Rate (TER)—be exploited as an indicator of language development/proficiency?

We performed two unpaired t-tests. We created two populations grouping our data per year of study of Italian. We selected texts written by learners at their first and second year of study of Italian in the first group, and texts written by learners at their fourth year of study of Italian and upwards in the second group. We computed TER and tree distance in F1 per each text measuring the difference—at string level and at tree level, respectively—between LSs and THs. In both experiments the P values were less than 0.0001, i.e. extremely statistically significant. These results are promising for two reasons: first, they quantitatively confirm that learners at two different stages of the learning path can be recognized both by looking at string and tree distance between their production and its normalized version, and second, they extrinsically assess and validate the consistency of the normalization of the LSs, i.e. the generation of valid THs.

7.1 Future work

Given the multi-faceted nature of the work presented here, there are many directions for future work. One possible direction would be that of correlating error annotation and parser performance loss. This correlation would be interesting to explain the differences emerged from the quantitative analysis

presented in Chapter 5. With the same vein, it would be interesting to qualitatively investigate what learner language (socio-linguistic and linguistic) features affect the parser performance at what level.

Another direction would be that of profiling learners' proficiency through errors, considering error type and frequency setting experiments similar to those described in Chapter 6. Since only 36 learner texts are currently annotated, in order to overcome this limitation, at least two paths are available: manually annotating texts focusing the annotation on some error tags (e.g. spelling errors); artificially create errors using the error patterns encountered in the manually annotated texts.

Another important future work is to annotate more data and expand VALICO-UD. Size is indeed an issue for Natural Language Processing tasks, such as Native Language Identification (NLI) and Grammatical Error Identification and Correction (GEI and GEC). VALICO-UD, as it currently is, could be used for a simplified NLI experiment—simplified because only four different L1 backgrounds are considered in the data whilst other datasets contain 11 L1s (Blanchard et al., 2013)—which could be confronted to the human baseline presented in Di Nuovo, Bosco, and Corino, 2020.

In order to provide a better resource for NLI, VALICO-UD can be expanded to include more L1 backgrounds. In addition, for GEI and GEC, other THs should be added, so that attention is given, as far as possible, to different correct versions of the same original text. These improvements would be beneficial not only for NLI, GEI and GEC tasks but also for interlanguage research.

The results described in Chapter 5 about the incremental evaluation of the parser suggest that to obtain an improved automatically annotated resource (+ 8 points in the Labeled Attachment Score of the LSs), PoS tagging should be as accurate as possible. Thus, one possible future work is to annotate gold tokenization and PoS tags also in the silver data. This would be a faster way to obtain more reliable automatically-annotated data to be used in computational tasks or interlanguage research.

Some of the current limitations of the resource—i.e. THs written by only one annotator, except for the three inter-annotator agreement experiments based on the core section of the treebank; linguistic annotation performed mostly by only one annotator, except for the inter-annotator agreement study based on 200 sentences—could be overcome by involving students, who are

able to acquire the knowledge necessary to perform annotation. In a non perfectly repeatable task (the same annotator can change their decision in the same experimental setup) and non deterministic task, the number of annotators is crucial to avoid errors stemmed from manual coding and annotation of linguistic features (Larsson, Paquot, and Plonsky, 2020).

However, although students are able to acquire the required skills, it is still challenging to find reasonably motivated students willing to do it.

Bibliography

- Aarts, Jan and Sylviane Granger (2014). "Tag sequences in learner corpora: A key to interlanguage grammar and discourse". In: *Learner English on computer*. Routledge, pp. 132–141.
- Abe, Mariko and Yukio Tono (2005). "Variations in L2 spoken and written English: Investigating patterns of grammatical errors across proficiency levels". In: *Proceedings from The Corpus Linguistics Conference*. URL: <https://bit.ly/3HAkedR>.
- Ahrenberg, Lars (2019). "Towards an adequate account of parataxis in Universal Dependencies". In: *Proceedings of the Third Workshop on Universal Dependencies (UDW, SyntaxFest 2019)*, pp. 94–100.
- Aijmer, Karin (2002). "Modality in advanced Swedish learners' written interlanguage". In: *Computer learner corpora, second language acquisition and foreign language teaching*, pp. 55–76.
- Alfieri, Linda and Fabio Tamburini (Dec. 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)*. Naples, Italy, pp. 19–23.
- Andorno, Cecilia Maria and Stefano Rastelli (2009). "Un'annotazione orientata alla ricerca acquisizionale". In: *Corpora di italiano L2: tecnologie, metodi, spunti teorici*. Ed. by Cecilia Maria Andorno and Stefano Rastelli. Guerra, pp. 49–70.
- Artstein, Ron (2017). "Inter-annotator agreement". In: *Handbook of linguistic annotation*. Springer, pp. 297–313.
- Artstein, Ron and Massimo Poesio (2008). "Inter-coder agreement for computational linguistics". In: *Computational Linguistics* 34.4, pp. 555–596.
- Astaneh, Sadegh and Francesca Frontini (2009). "L'adattamento di un parser di italiano L1: problemi e prospettive". In: *Corpora di italiano L2: tecnologie, metodi, spunti teorici* 2, pp. 199–216.

- Barni, Monica and Francesca Gallina (2009). "Il corpus LIPS (Lessico dell'Italiano parlato da Stranieri): problemi di trattamento delle forme e di lemmatizzazione". In: *Corpora di italiano L2: tecnologie, metodi, spunti teorici*. Ed. by Cecilia Andorno and Stefano Rastelli. Guerra Edizioni, pp. 139–151.
- Bell, Roger T. (1974). "Error analysis: a recent pseudoprocedure in applied linguistics". In: *International Review of Applied Linguistics* 25–26, pp. 35–49.
- Berzak, Yevgeni, Jessica Kenney, Carolyn Spadine, Jing Xian Wang, Lucia Lam, Keiko Sophie Mori, Sebastian Garza, and Boris Katz (Aug. 2016). "Universal Dependencies for Learner English". In: *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics*. Berlin, Germany, pp. 737–746.
- Blanchard, Daniel, Joel Tetreault, Derrick Higgins, Aoife Cahill, and Martin Chodorow (2013). "TOEFL11: A corpus of non-native English". In: *ETS Research Report Series* 2013.2, pp. i–15.
- Bley-Vroman, Robert (1983). "The comparative fallacy in interlanguage studies: The case of systematicity". In: *Language learning* 33.1, pp. 1–17.
- Boyd, Adriane (2012). "Detecting and diagnosing grammatical errors for beginning learners of German: From learner corpus annotation to constraint satisfaction problems". PhD thesis. The Ohio State University.
- (Nov. 2018). "Normalization in Context: Inter-Annotator Agreement for Meaning-Based Target Hypothesis Annotation". In: *Proceedings of the 7th workshop on NLP for Computer Assisted Language Learning*. Stockholm, Sweden: LiU Electronic Press, pp. 10–22. URL: <https://aclanthology.org/W18-7102>.
- Boyd, Adriane, Jirka Hana, Lionel Nicolas, Detmar Meurers, Katrin Wisniewski, Andrea Abel, Karin Schöne, Barbora Stindlová, and Chiara Vettori (May 2014). "The MERLIN corpus: Learner Language and the CEFR." In: *Conference on Language Resources and Evaluation*. Reykjavik, Iceland, pp. 1281–1288.
- Brunato, Dominique and Felice Dell'Orletta (2016). "ISACCO: a corpus for investigating spoken and written language development in Italian school-age children". In: *IJCoL. Italian Journal of Computational Linguistics* 2.2-1, pp. 63–76.
- Bryant, Christopher, Mariano Felice, Øistein E. Andersen, and Ted Briscoe (Aug. 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*. Florence, Italy: Association for Computational Linguistics, pp. 52–75. DOI: 10.18653/v1/W19-4406. URL: <https://www.aclweb.org/anthology/W19-4406>.
- Buchholz, Sabine and Erwin Marsi (2006). “CoNLL-X shared task on multilingual dependency parsing”. In: *Proceedings of the tenth conference on computational natural language learning (CoNLL-X)*, pp. 149–164.
- Bunt, Harry, Paola Merlo, and Joakim Nivre (2010). *Trends in parsing technology: Dependency parsing, domain adaptation, and deep parsing*. Vol. 43. Springer Science & Business Media.
- Burt, Marina and Carol Kiparsky (1972). *The Gooficon: A Repair Manual for English*. Newbury House.
- Carnie, Andrew, Dan Siddiqi, and Yosuke Sato (2014). *The Routledge handbook of syntax*. Routledge.
- Castilho, Richard Eckart de, Éva Mújdricza-Maydt, Seid Muhie Yimam, Silvana Hartmann, Iryna Gurevych, Anette Frank, and Chris Biemann (Dec. 2016). “A Web-based Tool for the Integrated Annotation of Semantic and Syntactic Structures”. In: *Proceedings of the Workshop on Language Technology Resources and Tools for Digital Humanities (LT4DH)*. Osaka, Japan: The COLING 2016 Organizing Committee, pp. 76–84. URL: <https://www.aclweb.org/anthology/W16-4011>.
- Chatzikoumi, Eirini (2020). “How to evaluate machine translation: A review of automated and human metrics”. In: *Natural Language Engineering* 26.2, pp. 137–161.
- Che, Wanxiang, Yijia Liu, Yuxuan Wang, Bo Zheng, and Ting Liu (Oct. 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*. Brussels, Belgium: Association for Computational Linguistics, pp. 55–64. DOI: 10.18653/v1/K18-2005. URL: <https://aclanthology.org/K18-2005>.
- Chomsky, Noam (1956). “Three models for the description of language”. In: *IRE Transactions on information theory* 2.3, pp. 113–124.
- (1965). *Aspects of the Theory of Syntax*. 2015th ed. MIT Press.
- Cignarella, Alessandra Teresa, Valerio Basile, Manuela Sanguinetti, Cristina Bosco, Paolo Rosso, and Farah Benamara (2020). “Multilingual Irony Detection with Dependency Syntax and Neural Models”. In: *arXiv preprint arXiv:2011.05706*.

- Cignarella, Alessandra Teresa, Cristina Bosco, and Paolo Rosso (Aug. 2019). "Presenting TWITTIRÒ-UD: An Italian Twitter Treebank in Universal Dependencies". In: *Proceedings of the Fifth International Conference on Dependency Linguistics (Depling, SyntaxFest 2019)*. Paris, France: Association for Computational Linguistics, pp. 190–197. DOI: 10.18653/v1/W19-7723. URL: <https://aclanthology.org/W19-7723>.
- Cohen, Jacob (1960). "A Coefficient of Agreement for Nominal Scales". In: *Educational and Psychological Measurement* 20.1, pp. 37–46.
- Conte, Giorgia, Cristina Bosco, and Alessandro Mazzei (2017). "Dealing with Italian adjectives in noun phrase: a study oriented to natural language generation". In: *4th Italian Conference on Computational Linguistics, CLiC-it 2017*. Vol. 2006. CEUR-WS, pp. 1–6.
- Cook, Vivian James (1993). *Linguistics and Second Language Acquisition*. Macmillan.
- Corder, Pit (1973). *Introducing Applied Linguistics*. Penguin.
- Corder, Stephen Pit (1967). "The significance of learner's errors". In: *International Review of Applied Linguistics* 5.4, pp. 161–170.
- (1971). "Idiosyncratic dialects and error analysis". In: *International Review of Applied Linguistics* 9.2, pp. 147–160.
- Corino, Elisa and Carla Marengo (2009). "Elicitare scritti a partire da storie diseguate: il corpus di apprendenti VALICO". In: *Corpora di italiano L2: tecnologie, metodi, spunti teorici*. Ed. by Cecilia Andorno and Stefano Rastelli. Guerra.
- (2017). *Italiano di stranieri. I corpora VALICO e VINCA*. Guerra.
- Corino, Elisa and Claudio Russo (Dec. 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*. Naples, Italy, pp. 105–110.
- Costantino, Mauro (2009). "Transcript-o'-matic: la trascrizione dei testi per VALICO". In: *VALICO. Studi di linguistica e didattica*. Ed. by Elisa Corino and Carla Marengo. Guerra.
- Dagneaux, Estelle, Sharon Denness, and Sylviane Granger (1998). "Computer-aided error analysis". In: *System* 26.2, pp. 163–174.
- Dagneaux, Estelle, Sharon Denness, Sylviane Granger, and Fanny Meunier (1996). *Error Tagging Manual. Version 1.1*. Centre for English Corpus Linguistics, Université Catholique de Louvain, Louvain-la-Neuve.

- Dahlmeier, Daniel, Hwee Tou Ng, and Siew Mei Wu (June 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*. Atlanta, Georgia, pp. 22–31.
- Davidson, Sam, Aaron Yamada, Paloma Fernandez Mira, Agustina Carando, Claudia H. Sanchez Gutierrez, and Kenji Sagae (May 2020). "Developing NLP tools with a new corpus of learner Spanish". In: *Proceedings of the 12th Language Resources and Evaluation Conference*. Marseille, France, pp. 7238–7243.
- De Marneffe, Marie-Catherine, Timothy Dozat, Natalia Silveira, Katri Haverinen, Filip Ginter, Joakim Nivre, and Christopher D Manning (2014). "Universal Stanford dependencies: A cross-linguistic typology." In: *LREC*. Vol. 14, pp. 4585–4592.
- De Mauro, Tullio (2016). *Il Nuovo Vocabolario di Base della Lingua Italiana*. Internazionale. URL: <https://bit.ly/3FqWN1k>.
- Del Río Gayo, Iria and Amália Mendes (May 2018a). "Error annotation in a Learner Corpus of Portuguese". In: *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*. Reykjavik, Iceland, pp. 7–11.
- (2018b). "Error annotation in the COPLE2 corpus". In: *Revista Da Associação Portuguesa De Linguística* 4, pp. 225–239.
- Del Río Gayo, Iria, Marcos Zampieri, and Shervin Malmasi (June 2018). "A Portuguese native language identification dataset". In: *Proceedings of the thirteenth workshop on innovative use of NLP for building educational applications*. New Orleans, USA, pp. 291–296.
- Di Nuovo, Elisa, Cristina Bosco, and Elisa Corino (2020). "How good are humans at Native Language Identification? A case study on Italian L2 writings". In: *CLiC-it 2020 Italian Conference on Computational Linguistics 2020*. CEUR, pp. 1–7.
- Di Nuovo, Elisa, Cristina Bosco, Alessandro Mazzei, and Manuela Sanguinetti (Nov. 2019). "Towards an Italian learner treebank in universal dependencies". In: *6th Italian Conference on Computational Linguistics, CLiC-it 2019*. Vol. 2481. CEUR-WS. Bari, Italy, pp. 1–6.
- Díaz-Negrillo, Ana, Nicolas Ballier, and Paul Thompson (2013). *Automatic treatment and analysis of learner corpus data*. Vol. 59. John Benjamins Publishing Company.

- Díaz-Negrillo, Ana and Jesús Fernández Domínguez (2006). "Error tagging systems for learner corpora". In: *Revista española de lingüística aplicada* 19, pp. 83–102.
- Díaz-Negrillo, Ana, Detmar Meurers, Salvador Valera, and Holger Wunsch (2010). "Towards Interlanguage POS Annotation for Effective Learner Corpora in SLA and FLT". In: *Language Forum* 36.1-2, pp. 139–154.
- Díaz-Negrillo, Ana and Paul Thompson (2013). "Learner Corpora". In: *Automatic treatment and analysis of learner corpus data*. Ed. by Ana Díaz-Negrillo, Nicolas Ballier, and Paul Thompson. Vol. 59. John Benjamins Publishing Company, pp. 9–29.
- Dickinson, Markus and Marwa Ragheb (Dec. 2009). "Dependency annotation for learner corpora". In: *Proceedings of the Eighth Workshop on Treebanks and Linguistic Theories (TLT-8)*. Milan, Italy, pp. 59–70.
- (2013). *Annotation for learner English guidelines*. Tech. rep. v. 0.1. Technical report. Indiana University.
- Dobrić, Nikola and Guenther Sigott (2014). "Towards an error taxonomy for student writing". In: *Zeitschrift für interkulturellen Fremdsprachenunterricht* 19.2, pp. 111–118.
- Dorr, Bonnie, Joseph Olive, John McCary, and Caitlin Christianson (2011). "Machine translation evaluation and optimization". In: *Handbook of natural language processing and machine translation*. Springer, pp. 745–843.
- Doval, Irene and M. Teresa Sánchez Nieto (2019). *Parallel corpora for contrastive and translation studies: New resources and applications*. Vol. 90. John Benjamins Publishing Company.
- Dulay, Heidi, Marina Burt, and Stephen Krashen (1982). *Language two*. Oxford University Press.
- EAGLES (1996). *Preliminary recommendations on Corpus Typology*. Expert Advisory Group on Language Engineering Standards. URL: <https://bit.ly/3oIy0Hx>.
- Ellis, Rod (1995). "Interpretation tasks for grammar teaching". In: *Tesol Quarterly* 29.1, pp. 87–105.
- (2015). *Understanding second language acquisition*. 2nd ed. Vol. 31. Oxford University Press, Oxford.
- Garrapa, Luigia (2009). "Vowel elision in spoken Italian". In: *Romance Languages and Linguistic Theory 2006*. John Benjamins, pp. 73–88.
- Garside, Roger, Geoffrey N. Leech, and Tony McEnery (1997). *Corpus Annotation: Linguistic Information from Computer Text Corpora*. Taylor & Francis.

- Giacalone Ramat, Anna (2003). *Verso l'italiano. Percorsi e strategie di acquisizione*. Roma, Carocci.
- Granger, Sylviane (1996). "From CA to CIA and back: An integrated approach to computerized bilingual and learner corpora". In: *Languages in contrast*. Ed. by K. Aijmer, B. Altenberg, and M. Johansson. Lund University Press, pp. 37–51.
- (2002). "A bird's-eye view of learner corpus research". In: *Computer Learner Corpora, Second Language Acquisition and Foreign Language Teaching*. Ed. by S. Granger, J. Hung, and S. Petch-Tyson. John Benjamins, pp. 3–33.
- (2003). "Error-tagged learner corpora and CALL: A promising synergy". In: *CALICO journal* 20.3, pp. 465–480.
- (2004). "Computer learner corpus research: Current status and future prospects". In: *Applied Corpus Linguistics: A Multidimensional Perspective*. Ed. by U. Connor and T. Upton. Rodopi, pp. 123–145.
- (2009). "The contribution of learner corpora to second language acquisition and foreign language teaching". In: *Corpora and language teaching* 33, pp. 13–32.
- (2015). "Contrastive interlanguage analysis: A reappraisal". In: *International Journal of Learner Corpus Research* 1.1, pp. 7–24.
- Granger, Sylviane, Estelle Dagneaux, Fanny Meunier, and Magali Paquot (2002). *International Corpus of Learner English. Handbook and CD-ROM*. Louvain-la-Neuve: Presses Universitaires de Louvain.
- (2009). *International Corpus of Learner English. Version 2 (Handbook and CD-ROM)*. Louvain-la-Neuve: Presses Universitaires de Louvain.
- Hana, Jirka, Alexandr Rosen, Barbora Stindlová, and Petr Jäger (2012). "Building a learner corpus". In: *Eighth International Conference on Language Resources and Evaluation (LREC 2012)*, pp. 3228–3232.
- Hawkins, John A. (1978). *Definiteness and Indefiniteness: A study in reference and grammaticality prediction*. London: Croom Helm.
- Hovy, Dirk and Shrimai Prabhumoye (2021). "Five sources of bias in natural language processing". In: *Language and Linguistics Compass* 15.8, e12432.
- Hudson, Richard A. (1984). *Word grammar*. Blackwell, Oxford.
- Hughes, Arthur and Chrysoula Lascaratou (1982). "Competing criteria for error gravity". In: *English Language Teaching Journal* 36.3, pp. 175–182.
- James, Carl (1998). *Errors in language learning and use*. Pearson Educational Limited.

- Köhn, Christine and Arne Köhn (Aug. 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)*. Santa Fe, New Mexico, USA, pp. 121–132.
- Lalande, John F. (1981). "Systematic marking of German compositions". In: *Die Unterrichtspraxis/Teaching German* 14.2, pp. 236–245.
- Landis, J Richard and Gary G Koch (1977). "An application of hierarchical kappa-type statistics in the assessment of majority agreement among multiple observers". In: *Biometrics*, pp. 363–374.
- Larsson, Tove, Magali Paquot, and Luke Plonsky (2020). "Inter-rater reliability in learner corpus research: Insights from a collaborative study on adverb placement". In: *International Journal of Learner Corpus Research* 6.2, pp. 237–251.
- Lee, John, Herman Leung, and Keying Li (May 2017). "Towards Universal Dependencies for Learner Chinese". In: *Proceedings of the NoDaLiDa 2017 Workshop on Universal Dependencies (UDW 2017)*. Gothenburg, Sweden, pp. 67–71.
- Lee, John, Keying Li, and Herman Leung (Sept. 2017). "L1-L2 parallel dependency treebank as learner corpus". In: *Proceedings of the 15th International Conference on Parsing Technologies*. Pisa, Italy, pp. 44–49.
- Lee, Sun-Hee, Markus Dickinson, and Ross Israel (2012). "Developing learner corpus annotation for Korean particle errors". In: *Proceedings of the Sixth Linguistic Annotation Workshop*, pp. 129–133.
- Lennon, Paul (1991). "Error: Some problems of definition, identification, and distinction". In: *Applied linguistics* 12.2, pp. 180–196.
- Li, Keying and John Lee (May 2018). "L1-L2 Parallel Treebank of Learner Chinese: Overused and Underused Syntactic Structures". In: *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*. Miyazaki, Japan, pp. 4106–4110.
- Little, David (2006). "The Common European Framework of Reference for Languages: Content, purpose, origin, reception and impact". In: *Language Teaching* 39.3, pp. 167–190.
- Lozano, Cristóbal (2009). "CEDEL2: Corpus escrito del español L2". In: *Applied linguistics now: Understanding language and mind*, pp. 197–212.
- Lozano, Cristóbal and Amaya Mendikoetxea (2013). "Learner corpora and second language acquisition". In: *Automatic treatment and analysis of learner*

- corpus data*. Ed. by Ana Díaz-Negrillo, Nicolas Ballier, and Paul Thompson. Vol. 59. John Benjamins Publishing Company.
- Lu, Xiaofei (2010). "Automatic analysis of syntactic complexity in second language writing". In: *International journal of corpus linguistics* 15.4, pp. 474–496.
- Lüdeling, Anke (2008). "Mehrdeutigkeiten und Kategorisierung: Probleme bei der Annotation von Lernerkorpora". In: *Fortgeschrittene Lernervarietäten*, pp. 119–140.
- Lüdeling, Anke and Hagen Hirschmann (2015). "Error annotation systems". In: *The Cambridge handbook of learner corpus research*. Ed. by Sylviane Granger, Gaëtanelle Gilquin, and Fanny Meunier. Cambridge: Cambridge University Press, pp. 135–157.
- Lüdeling, Anke, Maik Walter, Emil Kroymann, and Peter Adolphs (2005). "Multi-level error annotation in learner corpora". In: *Proceedings of Corpus Linguistics 2005*. Vol. 1, pp. 14–17.
- Lynn, Teresa (2016). "Irish Dependency Treebanking and Parsing". PhD thesis. Dublin City University, Ireland and Macquarie University, Sydney, Australia.
- Lyons, Christopher (1999). *Definiteness*. Cambridge University Press.
- Malchukov, Andrej L. and Andrew Spencer (2008). *The Oxford Handbook of Case*. Oxford University Press.
- Marello, Carla (2011). "Interpretare testi scritti composti a partire da storie diseguate". In: *Dimensionen der Analyse von Texten und Diskursen. Dimensionen dell'analisi di testi e discorsi*. Ed. by Klaus Hölker and Carla Marello. Vol. 1. LIT Berlin, pp. 283–304.
- Masini, Francesca, M. Silvia Micheli, Andrea Zaninello, Sara Castagnoli, and Malvina Nissim (2020). "Multiword Expressions We Live by: A Validated Usage-based Dataset from Corpora of Written Italian". In: *Proceedings of the Seventh Italian Conference on Computational Linguistics, CLiC-it 2020, Bologna, Italy, March 1-3, 2021*. Ed. by Johanna Monti, Felice Dell'Orletta, and Fabio Tamburini. Vol. 2769. CEUR Workshop Proceedings. CEUR-WS.org. URL: http://ceur-ws.org/Vol-2769/paper_33.pdf.
- McEnery, Tony and Nazareth Amselom Kifle (2002). "Epistemic modality in argumentative essays of second-language writers". In: *Academic Discourse*. Ed. by J. Flowerder. Longman, pp. 182–195.
- McEnery, Tony and Andrew Wilson (2001). *Corpus Linguistics: An Introduction (Second Edition)*. Edinburgh: Edinburgh University Press.

- Meara, Paul (1984). "The study of lexis in interlanguage". In: *Interlanguage*, pp. 225–235.
- Mendes, Amália, Sandra Antunes, Maarten Janssen, and Anabela Gonçalves (May 2016). "The COPLE2 corpus: a learner corpus for Portuguese". In: *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*. Portorož, Slovenia: European Language Resources Association (ELRA), pp. 3207–3214. URL: <https://aclanthology.org/L16-1511>.
- Meunier, Fanny and Céline Gouverneur (2009). "New types of corpora for new educational challenges: Collecting, annotating and exploiting a corpus of textbook material". In: *Corpora and Language Teaching*, pp. 179–201.
- Meurers, Detmar (2015). "Learner corpora and natural language processing". In: *The Cambridge handbook of learner corpus research*. Ed. by Sylviane Granger, Gaëtanelle Gilquin, and Fanny Meunier. Cambridge University Press Cambridge, UK, pp. 537–566.
- Meurers, Walt Detmar and Stefan Müller (2009). "Corpora and Syntax (Article 42)". In: *Corpus linguistics*. Ed. by Anke Lüdeling and Merja Kytö. Vol. 2. Berlin: Mouton de Gruyter, pp. 920–933. URL: <http://purl.org/dm/papers/meurers-mueller-09.html>.
- Nemser, William (1971). "Approximative systems of foreign language learners". In: *International Review of Applied Linguistics* 9.2, pp. 115–123.
- Nesselhauf, Nadja (2004). "Learner corpora and their potential for language teaching". In: *How to use corpora in language teaching* 12, pp. 125–156.
- Nicholls, Diane (2003). "The Cambridge Learner Corpus: Error coding and analysis for lexicography and ELT". In: *Proceedings of the Corpus Linguistics 2003 Conference*. Vol. 16, pp. 572–581.
- Osborne, Timothy (2014). "Dependency grammar". In: *The Routledge handbook of syntax*. Routledge, pp. 604–626.
- Papineni, Kishore, Salim Roukos, Todd Ward, and Wei-Jing Zhu (July 2002). "Bleu: a method for automatic evaluation of machine translation". In: *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*. Philadelphia, Pennsylvania, USA: Association for Computational Linguistics, pp. 311–318.
- Pericchi, Natalia, Bert Cornillie, Freek Van de Velde, and Kristin Davidse (2017). "La duplicación de clíticos en español como estrategia de marcación inversa". In: *Revue Romane. Langue et littérature. International Journal of Romance Languages and Literatures* 52.2, pp. 190–206.

- Petrov, Slav, Dipanjan Das, and Ryan McDonald (2011). "A universal part-of-speech tagset". In: *arXiv preprint arXiv:1104.2086*.
- Popel, Martin, Zdenek Zabokrtský, and Martin Vojtek (May 2017). "Udapi: Universal API for Universal Dependencies". In: *Proceedings of the NoDaLiDa Workshop on Universal Dependencies, UDW@NoDaLiDa 2017*. Ed. by Marie-Catherine de Marneffe, Joakim Nivre, and Sebastian Schuster. Gothenburg, Sweden: Association for Computational Linguistics, pp. 96–101. URL: <https://www.aclweb.org/anthology/W17-0412/>.
- Quirk, Randolph, Sidney Greenbaum, Geoffrey Leech, and Jan Svartvik (1985). *A Comprehensive Grammar of the English Language*. Longman.
- Radford, Andrew (1988). *Transformational grammar*. Cambridge University Press.
- Ragheb, Marwa (2014). "Building a Syntactically-Annotated Corpus of Learner English". PhD thesis. Indiana University.
- Ragheb, Marwa and Markus Dickinson (2011). "Avoiding the comparative fallacy in the annotation of learner corpora". In: *Selected proceedings of the 2010 Second Language Research Forum: Reconsidering SLA research, dimensions, and directions*. Cascadia Proceedings Project Somerville, MA, pp. 114–124.
- (Dec. 2012). "Defining syntax for learner language annotation". In: *Proceedings of COLING 2012: Posters*. Mumbai, India, pp. 965–974.
- (Dec. 2014a). "Developing a corpus of syntactically-annotated learner language for English". In: *Proceedings of the 13th International Workshop on Treebanks and Linguistic Theories (TLT13)*. Tübingen, Germany, pp. 137–148.
- (2014b). "The effect of annotation scheme decisions on parsing learner data". In: *CLARIN-D*, pp. 137–148.
- Renzi, Lorenzo, Giampaolo Salvi, and Anna Cardinaletti (2001). *Grande grammatica italiana di consultazione*. Vol. 1–3. Il mulino.
- Reznicek, Marc, Anke Lüdeling, and Hagen Hirschmann (2013). "Competing target hypotheses in the Falko corpus". In: *Automatic treatment and analysis of learner corpus data 59*, pp. 101–123.
- Reznicek, Marc, Maik Walter, Karin Schmidt, Anke Lüdeling, Hagen Hirschmann, Cedric Krummes, and Torsten Andreas (2010). "Das Falko-Handbuch: Korpusaufbau und Annotationen". In: *Institut für deutsche Sprache und Linguistik*.

- Rosen, Alexandr, Jirka Hana, Barbora Štindlová, and Anna Feldman (2014). "Evaluating and automating the annotation of a learner corpus". In: *Language Resources and Evaluation* 48.1, pp. 65–92.
- Rosén, Victoria and Koenraad De Smedt (2010). "Syntactic annotation of learner corpora". In: *Systematisk, varieret, men ikke tilfældig*, pp. 120–132.
- 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". In: *A rainbow of corpora: Corpus linguistics and the languages of the world*, pp. 27–38.
- Rozovskaya, Alla and Dan Roth (2010). "Annotating ESL errors: Challenges and rewards". In: *Proceedings of the NAACL HLT 2010 fifth workshop on innovative use of NLP for building educational applications*, pp. 28–36.
- Russell, Bertrand (1905). "On denoting". In: *Mind* 14.56, pp. 479–493.
- Ruzaitė, Jūratė, Sigita Dereškevičiūtė, Viktorija Kavaliauskaitė-Vilvinienė, and Eglė Krivickaitė-Leišienė (2020). "Error Tagging in the Lithuanian Learner Corpus". In: *Human Language Technologies–The Baltic Perspective*. Ed. by Andrius Utkas, Jurgita Vaičenonienė, and Jolanta Kovalevskaitė. IOS Press, pp. 253–260.
- Sag, Ivan A, Timothy Baldwin, Francis Bond, Ann Copestake, and Dan Flickinger (2002). "Multiword expressions: A pain in the neck for NLP". In: *International conference on intelligent text processing and computational linguistics*. Springer, pp. 1–15.
- Sagae, Kenji, Eric Davis, Alon Lavie, Brian MacWhinney, and Shuly Wintner (2010). "Morphosyntactic annotation of CHILDES transcripts". In: *Journal of child language* 37.3, pp. 705–729.
- Salem, Ilana (July 2007). "The lexico-grammatical continuum viewed through student error". In: *ELT Journal* 61.3, pp. 211–219. ISSN: 0951-0893. DOI: 10.1093/elt/ccm028. eprint: <https://academic.oup.com/eltj/article-pdf/61/3/211/1118374/ccm028.pdf>. URL: <https://doi.org/10.1093/elt/ccm028>.
- Sampson, Geoffrey (1995). *English for the Computer: The SUSANNE Corpus and Analytic Scheme*. Clarendon Press.
- 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, pp. 51–69.

- Sanguinetti, Manuela, Cristina Bosco, Alberto Lavelli, Alessandro Mazzei, Oronzo Antonelli, and Fabio Tamburini (May 2018). "PoSTWITA-UD: an Italian Twitter Treebank in Universal Dependencies". In: *Proceedings of the Eleventh International Conference on Language Resources and Evaluation*. Miyazaki, Japan, pp. 1768–1775.
- Schmitt, Norbert (1977). "Judgments of error gravities". In: *English Language Teaching Journal* 31, pp. 116–124.
- (1993). "Comparing native and nonnative teachers' evaluations of error seriousness". In: *Japanese Association of Language Teachers: Journal* 15.2, pp. 181–191.
- Selinker, Larry (1972). "Interlanguage". In: *International Review of Applied Linguistics* 10.3, pp. 209–231.
- Sgall, Petr, Eva Hajicová, Eva Hajicová, Jarmila Panevová, and Jarmila Panevova (1986). *The meaning of the sentence in its semantic and pragmatic aspects*. Springer Science & Business Media.
- Siemen, Peter, Anke Lüdeling, and Frank Henrik Müller (2006). "FALKO—ein fehlerannotiertes Lernerkorpus des Deutschen". In: *Proceedings of Konvens*. Vol. 2006. CiteSeer, p. 107.
- Silveira, Natalia, Timothy Dozat, Marie Catherine De Marneffe, Samuel R. Bowman, Miriam Connor, John Bauer, and Christopher D. Manning (May 2014). "A Gold Standard Dependency Corpus for English". In: *Proceedings of the 9th International Conference on Language Resources and Evaluation, LREC 2014*. Reykjavik, Iceland: ELRA, pp. 2897–2904.
- Simi, Maria, Cristina Bosco, and Simonetta Montemagni (May 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*. European Language Resources Association (ELRA). Reykjavik, Iceland, pp. 83–90.
- Simone, Raffaele (2008). *Fondamenti di linguistica*. Laterza.
- Sinclair, John (2005). "Corpus and Text—Basic Principles". In: *Developing Linguistic Corpora: a Guide to Good Practice*. AHDS. URL: http://icar.cnrs.fr/ecole_thematique/contaci/documents/Baude/wynne.pdf.
- Siyanova-Chanturia, Anna and Stefania Spina (2020). "Multi-word expressions in second language writing: A large-scale longitudinal learner corpus study". In: *Language Learning* 70.2, pp. 420–463.

- Snover, Matthew, Bonnie Dorr, Richard Schwartz, Linnea Micciulla, and John Makhoul (Aug. 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*. Cambridge, Massachusetts, USA, pp. 223–231.
- Sparrow, Wendy (2014). "Unconventional word segmentation in emerging bilingual students' writing: A longitudinal analysis". In: *Applied linguistics* 35.3, pp. 263–282.
- Stemberger, Joseph Paul (1982). "Syntactic errors in speech". In: *Journal of Psycholinguistic Research* 11.4, pp. 313–345.
- Straka, Milan (Oct. 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*. Brussels, Belgium, pp. 197–207.
- Tenfjord, Kari, Jon Erik Hagen, and Hilde Johansen (2006). "The hows and whys of coding categories in a learner corpus (or How and why an error-tagged learner corpus is not ipso facto one big comparative fallacy)". In: *Rivista di Psicolinguistica Applicata (RiPLA). Special Issue on "Interlanguage: current thoughts and practices"* 6.3, pp. 93–108.
- Tenfjord, Kari, Paul Meurer, and Knut Hofland (2006). "The ASK Corpus - a Language Learner Corpus of Norwegian as a Second Language". In: *Proceedings of the Fifth International Conference on Language Resources and Evaluation (LREC'06)*. Genoa, Italy: European Language Resources Association (ELRA). URL: http://www.lrec-conf.org/proceedings/lrec2006/pdf/573_pdf.pdf.
- Tesnière, Lucien (1959). *Elments de syntaxe structurale*. Klincksieck.
- Tetreault, Joel and Martin Chodorow (2008a). "Native judgments of non-native usage: Experiments in preposition error detection". In: *Coling 2008: Proceedings of the workshop on human judgements in computational linguistics*, pp. 24–32.
- (2008b). "The ups and downs of preposition error detection in ESL writing". In: *Proceedings of the 22nd International Conference on Computational Linguistics (Coling 2008)*, pp. 865–872.
- Tono, Yukio (Mar. 2003). "Learner corpora: design, development and applications". In: *Proceedings of the Corpus Linguistics 2003 conference*. Lancaster, United Kingdom, pp. 800–809.
- (2013). "Criterial feature extraction using parallel learner corpora and machine learning". In: *Automatic treatment and analysis of learner corpus data*.

- Ed. by Ana Díaz-Negrillo, Nicolas Ballier, and Paul Thompson, pp. 169–203.
- Vann, Roberta J., Daisy E. Meyer, and Frederick O. Lorenz (1984). “Error gravity: A Study of Faculty Opinion of ESL Errors”. In: *TESOL Quarterly* 18.3, pp. 427–440.
- Whorf, Benjamin Lee (1945). “Grammatical categories”. In: *Language*, pp. 1–11.
- Wisniewski, Katrin, Karin Schöne, Lionel Nicolas, Chiara Vettori, Adriane Boyd, Detmar Meurers, Andrea Abel, and Jirka Hana (2013). “MERLIN: An online trilingual learner corpus empirically grounding the European Reference Levels in authentic learner data”. In: *ICT for Language Learning 2013, Conference Proceedings, Florence, Italy. Libreriauniversitaria. it Edizioni*.
- Yannakoudakis, Helen, 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*, pp. 180–189.
- Zeman, Daniel (2008). “Reusable Tagset Conversion Using Tagset Drivers”. In: *LREC*. Vol. 2008, pp. 28–30.
- Zeman, Daniel, Jan Hajič, Martin Popel, Martin Potthast, Milan Straka, Filip Ginter, Joakim Nivre, and Slav Petrov (Oct. 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*. Brussels, Belgium: Association for Computational Linguistics, pp. 1–21. DOI: 10.18653/v1/K18-2001. URL: <https://aclanthology.org/K18-2001>.