Di Nuovo, Elisa, De Paolis, Bianca Maria, Bosco, Cristina & Corino, Elisa (2024). Error identification, correction and tagging: Three inter- annotator agreement experiments in a picture-elicited learner corpus. In Katherine Ackerley & Erik Castello (eds) *Continuing Learner Corpus Research: Challenges and Opportunities*. Corpora and Language in Use Proceedings 7, Louvain-la-Neuve: Presses universitaires de Louvain, 87-120.

Error identification, correction and tagging: Three inter-annotator agreement experiments in a picture-elicited learner corpus

Elisa Di Nuovo[1]*, Bianca Maria De Paolis[1, 2], Cristina Bosco[1], Elisa Corino[1]

University of Turin[1], University of Paris 8[2]

Abstract

Inter-Annotator Agreement (IAA) in learner texts has gained attention in recent years as it brings out both the range of valid interpretations of non-canonical forms and unforeseen problematic areas, but also annotators' biases. In this study, we report on three IAA experiments – error identification, correction and tagging – measured using Cohen's κ. Two annotators were involved in the annotation of VALICO-UD, a sub-corpus of VALICO in Universal Dependency format. The results show that error identification is more consistent across raters in picture-elicited texts than in texts elicited under less controlled conditions. IAA on error correction is lower than on identification. This could be explained by the non-deterministic nature of this task. However, from disagreement analysis it emerged that 40% of disagreement was only apparent (i.e., caused by distraction or format). IAA on error tagging was moderate after a first round of annotation using a tag set of 120 tags. By removing apparent disagreement annotators achieved almost perfect agreement.

* Elisa Di Nuovo as of 1st October 2023 is employed by the Joint Research Centre of the European Commission.

1. Introduction

Error annotation is a complex, time-consuming task requiring specific skills. The complexity can be associated with the definition itself of error. Lennon defines it as "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: 182). Although this definition is not exhaustive as it does not account for the difference between covert and overt errors, or deviations, errors and mistakes (Corder 1982), it is adequate for our purposes as these differences are not applicable in Learner Corpus Research.

Also, to annotate errors, one has to identify their presence – which is not as straightforward as one can think (see for English language the study by Hughes & Lascaratou 1982 in which error identification proved to be a difficult task both for (non-)native teachers and native non-teachers) – and locate them, as in Grammatical Error Detection tasks, where systems are trained to mark errors at token level (Colla et al. 2023; Volodina et al. 2023).

It is also necessary to decide, in analysing errors, whether to adopt an explanatory or a descriptive approach (closely related to the type of taxonomy used), and in tagging errors, whether to choose the error tag according to what the learner wrote or what they should have written, and, finally, what to include in the error span.

Other concepts to be considered when error-tagging a corpus are scope and substance. As defined in Quirk et al. (1985: 85-86), scope "describes the semantic 'influence' that such words have on neighbouring parts of a sentence". Dobrić & Sigott (2014: 114-117) collocate scope in the context of error analysis and define 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-annotated corpora contain in the annotated error only its substance (see Example 2), others also its scope (see Example 1).

(1) **ICLE:** Students have the
(XNCO) possibility to leave $possibility of leaving$

(2) **CLC:** Students have the possibility
<#FV><#RT>to|of</#RT>leave|leaving</#FV>

Example 1 is drawn from the *International Corpus of Learner English* (ICLE) annotation guidelines (Dagneaux et al. 1996, p. 18); in Example 2 the annotation of the same segment applying the error-tagging system of the *Cambridge Learner Corpus* (CLC) is presented. The error marked in 1 indicates erroneous complementation of nouns, and the tag includes not only the substance (i.e., tokens that need to be modified to correct the error, *to leave*), but also its scope (i.e., what triggers the error, the noun *possibility* in the example).

This introduction has given an indication of how complex error annotation is. Because of this complexity, Inter-Annotator Agreement (IAA) – a common practice in Computational Linguistics for comparing the decisions of two or more human judges about the same product (Artstein 2017) – becomes essential, as it is a powerful instrument to validate and improve annotation schemes and guidelines for annotators, bring out the range of valid interpretations of the same non-canonical forms, and identify unforeseen problematic areas, but also annotator biases (Hovy & Prabhumoye 2021).

Despite these advantages and benefits associated with the practice, the first error-annotated learner corpora were usually tagged by one single coder and revised by another one, e.g., CLC (Nicholls 2003), thus not reporting IAA studies. This issue was first raised by Meurers & Müller (2009). Since then, several scholars have started to pay attention to it (Rozovskaya & Roth 2010; Boyd 2012; Lee et al. 2012; Dahlmeier et al. 2013; Rosen et al. 2014; Boyd 2018; Del Río Gayo & Mendes 2018; Köhn & Köhn 2018).

In the present contribution we report on three IAA experiments on error identification, correction and tagging, respectively. The experiments are carried out on the annotation conducted by two annotators on the core section of a novel treebank, i.e., VALICO-UD (Di Nuovo et al. 2022), which has been publicly available in the Universal Dependencies repository since May 2021 and updated in May 2023 (version 2.12).

In the remainder of this contribution, we first review previous IAA studies on learner corpora to contextualise our experiments in relation to previous studies. Then, we describe the rationale of this contribution, present its data and methodology, discuss the results obtained, and finally we conclude it with some remarks on the results and future work perspectives.

1.1. Previous Inter-Annotator Agreement experiments

In this section we report on previous IAA experiments carried out by Dahlmeier et al. (2013), Köhn & Köhn (2018), Boyd (2018) & Del Río Gayo & Mendes (2018) to allow for comparisons with the IAA experiments conducted in this study. However, it should be noted that the experiments mentioned are not fully comparable, since they differ in terms of target languages, types of errors investigated and error annotation systems. Nevertheless, we think that a review of these studies can constitute a valid source of inspiration and, above all, a good argumentative support in favour of the need for uniformity and standardisation of error targets.

Dahlmeier et al. (2013) used the learner English corpus NUCLE, which was in fact created with the Grammatical Error Correction task in mind. It consists of about 1,400 essays (a free writing type of assignment) generated by undergraduate university students, i.e., over one million words. Errors are annotated using a tag set of 27 categories, and corrections are provided for each error.

The authors measured IAA under three different conditions: 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 together.

The experiments were performed on 96 essays, which are not included in the final version of NUCLE. For reasons related to format and annotation tools, the authors had to perform text processing to allow for a comparison of the two annotations (Dahlmeier et al. 2013: 25-26). They reported $\kappa = 0.39$ on error identification, $\kappa = 0.55$ on tag choice and $\kappa = 0.48$ on exact match, which in Landis & Koch's terms (Landis & Koch 1977) can be considered fair (on error identification) and moderate (on tag choice and exact match) agreement.

Köhn & Köhn (2018) used a picture-elicited corpus of learner German in which two Target Hypotheses (THs), form-based TH and meaning-based TH,[1] are annotated following the FALKO's annotation guidelines (Reznicek et al. 2010). The average IAA reported on the two THs is $\kappa = 0.79$ on error identification and $\kappa = 0.64$ on error correction.

Also using a learner German corpus, specifically a reading comprehension corpus, Boyd (2018) reported $\kappa = 0.68$ on the error identification task (mean-

1 Form-based THs correct only grammatical errors, ignoring the context and the learner's intended meaning. Meaning-based THs take the context and the learner's intended meaning into account.

ing-based TH). The author adds that, in cases in which the annotators agreed in the identification, 70% of the time annotators agreed also about the correction.

Del Río Gayo & Mendes (2018) measured IAA on error tagging and correction of a learner Portuguese corpus (COPLE2). They evaluated two tag sets in two different samples. In the first, token-based, only errors affecting single tokens were corrected and tagged as either orthographic, grammatical or lexical. In the second, a fine-grained tag set with 37 tags was tested, and the correction was also requested. The IAA achieved on the token-based is $\kappa = 0.86$ ($\kappa = 0.85$ if considering also the correction) and only 0.01 less on the fine-grained tag set (i.e. $\kappa = 0.85$ and $\kappa = 0.84$ with the correction).

Both the studies by Dahlmeier et al. (2013) & Del Río Gayo & Mendes (2018) make use of learner essays, but the reported IAA is very different. This might be due to various factors, e.g., the L2, the error tag set.

With regard to the correlation of error identification and error types, previous studies reported a higher agreement for orthographic and grammatical errors (Del Río Gayo & Mendes 2018 reported 0.96 and 0.93 respectively), lower for lexical errors (Del Río Gayo & Mendes 2018 reported 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.

None of these studies have carried out more than one annotation round. Instead, a second round of annotation would be necessary to ensure annotation quality, as it allows for intra-annotator agreement checks (Díez-Bedmar 2021: 96).

1.2. The present contribution

Based on this short review, several key issues stand out as possible points to be addressed in further studies. Firstly, the studies mentioned exhibit large variation in their approaches, including differences in targeted languages, types of errors examined, and error annotation systems: this heterogeneity supports the argument for the necessity of uniformity and standardisation in error annotation, a goal toward which the practice of IAA remains a key step. Secondly, the reported IAA values vary considerably among these studies, but the factors that

contribute to these differences in agreement remain unclear: this underlines the need for a clearer methodology description, apart from the need for clear format criteria and annotation guidelines. Finally, none of the studies mentioned conducted multiple annotation rounds, indicating another potential area for improvement.

These findings collectively emphasise the importance of rigorous and standardised IAA in error analysis studies: in the present contribution, we seek to draw the attention of the Learner Corpus Research community to IAA on error annotation, and possibly try to tackle some of the aforementioned issues.

The main aim of this study is to validate and improve our error annotation guidelines and tag set (fully described in Di Nuovo 2023). In addition to this, our work is motivated by a few gaps in the existing literature on IAA and error annotation in learner corpora. First, although some studies report IAA, they are still rare compared to the total number of papers published in this area. Second, studies reporting IAA do not make their data available, not allowing reproducibility and full comparability.[2] Moreover, none of the previous studies focused on Italian as a second or foreign language. Adapting the CLC tag set to Italian can serve the dual purpose of demonstrating its applicability to other languages and assessing whether modifications are necessary: if changes are required, it is essential to understand what adjustments are needed, and the reasons behind them.

In the following sections, we will describe in detail the data and the methods used, trying to establish a link between our data and that used in the above-mentioned studies, to answer the following research questions:

1. Is error identification more reproducible using picture-elicited texts (i.e., within a constrained linguistic and extra-linguistic context)?
2. When different annotators agree on the presence of an error, do they also agree on its correction?
3. Is error tagging more reliable when based on an explicit Target Hypothesis (TH) (i.e., a correct version of what the learner wrote)?

2 The data used in this article is publicly available in the UD repository. The annotated documents for reproducibility and comparability are available on request.

2. Data and methodology

2.1. Corpus

The present study is based on the core section of VALICO-UD (Di Nuovo et al. 2022), publicly available on the Universal Dependencies repository (https://github.com/UniversalDependencies/UD_Italian-Valico/) since May 2021. VALICO-UD, as its name suggests, is a subcorpus of VALICO (Corino & Marello 2017), the biggest learner Italian written corpus publicly available and downloadable. The main characteristic of VALICO – apart from being enriched by a variety of metadata, hence enabling the creation of sub-corpora following precise design criteria – is that texts are elicited from learners based on comic strips. This means that the reconstruction of a possible TH is more reliable, since lexical choices and semantic frames are circumscribed to the comic strip, as has been widely demonstrated in the literature (Corino & Marello 2009; Marello 2011; Meurers 2015).

Figure 1. Comic strip eliciting VALICO-UD core-section texts

The texts included in the core section of VALICO-UD were selected according to three design criteria: learners' L1 (constrained to English, French, German and Spanish), eliciting comic strip (the one shown in Figure 1), and number of

years studying Italian as an L2 (trying to have a balanced set of learners in their first, second and third year of study of Italian).

The chosen comic strip (Figure 1) comprises a sequence of four images devoid of textual content. In the initial image, we see a man engrossed in reading a newspaper, but his concentration is abruptly disrupted by the arrival of another man, who is carrying a distressed woman. The subsequent illustration portrays the first man deciding to take action. In the third image, the same man appears elated, while the other man, who was carrying the woman, is depicted lying on the ground, and the woman displays a mix of astonishment and concern. Lastly, in the fourth and final frame, the enraged woman, with her left-hand index finger pointing downward, appears to engage in a heated argument with the first man. The cause of her anger is suggested by the thought bubble above her head, which contains a heart symbol.

The corpus consists of 36 learner texts, with a total of 398 sentences and 6,784 tokens. Each learner text is paired with a corrected, sentence-aligned text (i.e., 398 THs for a total of 6,832 tokens).

The tag set used for VALICO-UD – described in Di Nuovo (2023: 57-80) and used in this study – is adapted from the error tag set used to annotate the *Cambridge Learner Corpus* (Nicholls 2003). It is made of 120 unique tags ascribable to eight macro-categories, indicated by the first letter in the error code (e.g., IJG stands for Inflection, adJective, Gender).

2.2. Procedure

Each text was annotated by two annotators. We opted not to involve a third annotator for two primary reasons: first, as we will see in the results, the Inter-Annotator Agreement (IAA) was already high between the two, so no resolution techniques were necessary (Díez-Bedmar 2021: 96); second, the literature on LCR and error annotation typically reports about one annotator, and only recently two, as IAA has gained attention: the inclusion of a third annotator is usually recommended for tasks in which agreement tends to be lower, such as sentiment analysis.

The two annotators involved meet three out of four requirements stated by Díez-Bedmar (2021: 94-95): (1) experience in error-tagging, (2) good command of the target variety and the learners' L1 variety, (3) familiarity with applied linguistics and corpus linguistics. In fact, when they performed this task, the two annotators were both PhD students in Digital Humanities at the

University of Turin, with a background in Foreign Languages and a master's degree thesis in Applied Linguistics. They have expertise in applied linguistics, corpus linguistics, second language acquisition and error-tagging. They are both native speakers of Italian (one from Sicily and one from Piedmont) and proficient (certified C1 level in CEFR terms) in English plus another language, one knows Spanish and the other French, both at C2 (not certified); in addition, they have a basic knowledge (approximately A1 level) of German. The fourth requirement is about being part of a multilingual team covering target and learners' L1 varieties. Aware that we could not meet this requirement, we decided to exclude from annotation any features potentially related to crosslinguistic influence. In this way, we circumvented any bias related to the annotators' knowledge, or lack of knowledge, of the learners' L1. Some concise guidance was given to ensure that both annotators shared a common understanding of the study's methodology, theoretical framework, and acquaintance with the dataset, labels, and annotation guidelines. Guidelines and label sets have been provided in a detailed version, accompanied by ship logs, with extensive comments and examples. The guidelines and all materials containing the instructions can be retrieved in Di Nuovo (2023) and in the related online repository (https://github.com/ElisaDiNuovo/VALICO-UD_guidelines). This choice is motivated by both annotators' skills with similar tasks and their willingness to evaluate the quality of the guidelines.

One of the issues in reporting IAA concerns the decision of the best-suited measure for the particular kind of task. A thorough survey of methods by Artstein & 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 tag sets and set-value interpretations. They also attest the use of Krippendorff's α and Cohen's κ in the vast majority of studies and restate their appropriateness. To date, Cohen's κ is the most used measure for IAA in the learner corpus community, hence the measure we have chosen to report in this study. To calculate Cohen's κ, we used the script written by Lippincott (https://cswww.essex.ac.uk/Research/nle/arrau/Lippincott/) (Boyd 2018), adapting it to make it runnable in Python 3.[3]

The first and second experiments were carried out in one continuous session, synchronously, by the two annotators. For the third experiment, which included error tagging, a more extended timeframe was allocated: it spanned two weeks

3 Please note that the mentioned script is no longer working as of June 2023, probably due to last changes in python version. We recommend the Cohen's kappa score calculation available in sklearn.metrics (https://scikit-learn.org/stable/modules/generated/sklearn.metrics.cohen_kappa_score.html).

and was carried out asynchronously by the two annotators. After the first annotation round and first IAA calculation and disagreement analysis, a second round of annotations was performed by the same annotators, meant to identify, comment on, and correct annotators' mistakes.[4] A week elapsed between the end of the first round and the beginning of the second one.

2.2.1. *Error identification and correction*

Error identification and correction have been performed at token level. This requires that texts are first tokenised. We tokenised them using spaces, as in the studies described in Section 1.2. This means that orthographic words made of more than one syntactic word which are not divided by spaces, such as *al* made of *a* ('to') plus *il* ('the_MASC-SING') are kept in a single line. Each annotator also had access to the entire text: this approach ensured that when annotating errors, they always had the entire context at their disposal, allowing them to identify and correct larger-span errors.

We created a tab-separated file with one token in each line and an empty line to divide different sentences. Each line contained three columns: in the first one we have the token; in the second the binary annotation to indicate the presence of an error (i.e., 1) or its absence (i.e., 0); in the third the correction of the identified error (nothing if no error is identified). The annotators used a simple text editor (specifically, one annotator used *Sublime text* the other *Notepad*) to carry out the annotation.

Missing token errors are marked in the following token, as done in the other studies on token-level error identification and correction tasks. Let us consider the following invented learner sentence **You are best,* which contains two errors: the absence of *the* and the sentence-final punctuation. In error identification and correction experiments, the sentence would be tokenised as shown in Example 3.

(3)
You	0	
are	0	
best	1	the best

4 We use the term 'mistakes' referring to annotators to mean anything that includes oversights, format issues, and inconsistencies regarding the established annotation guidelines. Please note that our use is different from the term 'mistakes' in the sense of Corder (1982), which refers to performance errors in learners' productions not due to competence.

In the example, in the first column we have the tokens composing the learner sentence (*You are best*), in the second column the annotation at token-level indicating with 0 no error and with 1 the presence of an error, and finally, in the third column the correction of the identified error. Two corrections have been made: the addition of *the*, which being a missing token error has been added to the following token, *best*, and the addition of the sentence-final period. Please notice that, if the missing token is at the end of a sentence, it is corrected in the previous and sentence-final token actually written by the learner, *best* in the example. Also notice that, error identification being treated as a binary task, it does not provide information about the number of errors occurring in a token. This could be overcome by marking the number of errors at token-level and not merely the presence or absence of an error (cf. Example 3 and Example 4). This might be interesting to investigate in future studies.

(4) You 0
are 0
best 2 the best

Errors arising from the correction of another error (i.e., cascade errors) are also marked and corrected, as shown in Example 5.

(5) seduto 0
sul 1 sulla
banco 1 panchina

In the example both the tokens *sul* and *banco* are marked as to-be-edited, because the articled preposition *sul* agrees with *banco*, masculine singular, but not with the hypothesised correction, *panchina*, which is instead feminine singular. As a result, by correcting cascade errors we can obtain complete THs. However, cascade error correction makes error identification even more challenging, because it is influenced by correction. Marking a token as "to-be-edited" did not imply the presence of only one error in that token, nor did it mean that the same error was limited to a single token. The one-token-per-line annotation was not designed for error counting; instead, it facilitated the identification of tokens that needed correction to match the THs. Error counting was performed using inline error annotation, extensively described in Di Nuovo (2023: 57-80). A short description of the error tag set is outlined in the Appendix.

2.2.2. *Error tagging*

The error tag set used for this experiment, the same used to annotate VALICO-UD, is complex, since it is potentially expandable *ad infinitum*. One of the purposes of this third experiment is precisely to validate the tag set.

Since the objective was to evaluate error tags and not error identification and correction (as in the previous two experiments), this time we provided learner sentences (LSs) and associated THs to both annotators. In this way, the 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 tag set 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.

The annotators carried out the annotation using two different tools: first Transcript'o-matic (Costantino 2009) and then WebAnno (de Castilho et al. 2016), since manual error annotation in XML format using only Transcript'o-matic proved to be slow and error-prone. The output of WebAnno is moreover not an XML file but a special type of TSV file.

To be able to compare the two produced annotations, a conversion was necessary. We created a script that aligns and compares pairs of annotations, adding a zero per each additional tag annotated only by one annotator.

3. Quantitative analysis

3.1. Error identification

Using the space-based tokenisation we obtained 5,602 tokens. The annotators worked independently. Both annotators marked 950 tokens as to-be-edited; in addition to these 950 tokens, the first annotator also marked 159 tokens not marked by the second annotator, for a total of 1,109 tokens marked as to-be-edited by the first annotator. The second annotator instead marked as to-be-edited 1,148 tokens, including 198 not marked by the first annotator. These figures are reported in Table 1. Their agreement, expressed in Cohen's κ, is then 0.82.

	Annotator 1	Annotator 2	Error	No Error
Annotator 1	159	950	1,109	4,493
Annotator 2	950	198	1,148	4,454

Table 1. Number of tokens marked as to-be-edited (Error), or not (No Error) per annotator

Although it is not directly comparable for the reasons mentioned in Section 1.1, it is interesting to notice that a similar result ($\kappa = 0.79$) is reported in Köhn & Köhn (2018). This similar result could be explained by the fact that Köhn & Köhn (2018) also used a picture-elicited corpus.

Lower results are instead reported by Boyd (2018) ($\kappa = 0.68$) & Dahlmeier et al. (2013) ($\kappa = 0.39$). In these two studies, texts are not elicited by comic strips, so the context is not circumscribed, and errors are harder to identify. For instance, in Example 5 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 1) 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 sit') and without the comic strip nor the defined context (i.e., a man in a park) it would not be identified as error.

3.2. Error correction

The error correction experiment is meant to verify if the two annotators agree on the correction of a token providing the same edited version. To do so, the annotators were asked to mark the correction next to the tokens that, in their opinion, needed to be edited, as shown in the third column of Examples 3-5.

As in the studies mentioned in Section 1.1, we measured the agreement of the correction 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 corrections differing (out of 950 tokens marked as to-be-edited by both annotators).

Also in this task, Köhn & Köhn (2018) reported a similar result ($\kappa = 0.64$). We do not report here the results in Dahlmeier et al. (2013) because they calculated the agreement of the correction considering also the associated error tag. Boyd

(2018), instead, reported only the percentage of cases in which both annotators provided the same correction (i.e. 70%).

3.3. Error tagging

IAA on error tagging was moderate ($\kappa = 0.50$), in Landis & Koch's (1977) terms. Since we wanted to avoid disagreement due to format conversion of other mistakes (considering also the complexity of this task compared to the previous two experiments), we carried out a second round of annotation and computed IAA after the revisions of annotators' work. The two annotators revised together the disagreement, solving apparent disagreement (please see Section 4.1.1 for the definition of apparent disagreement). In total 1,203 error tags were marked by both annotators (more than one tag can be applied to one token, and a tag can span across different tokens). Annotators were given student's sentences with the pre-identified errors. They were also equipped with the TH and they had to tag the errors as belonging to one or more error types using the error tag set. 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 ($\kappa = 0.95$). The remaining disagreement can be due to different tags selected for the same error (as identified by the difference between learner sentence and TH), or the annotation of multiple tags to the same error solely on the part of one annotator (see Section 4.2).

4. Qualitative analysis

4.1. Error identification and correction

To analyse IAA in relation to error category we empirically classified disagreement relying on the corrected texts, since we do not have information about error category in error identification and correction experiments.

As far as error identification is concerned, we identified 17 sources of disagreement. We put them in order of frequency and not of relevance. We identified punctuation errors as a major source of disagreement. 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 a 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 correct these instances more than the other annotator.

Also in error correction, when the annotators disagreed on the corrected 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 quotation marks). In particular, looking at commas – which make up 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 corrected more missing commas in error identification.

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 & Chodorow (2008), 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 generalise the results. Then, if this is confirmed, it would be interesting to verify how this subject is treated in the textbooks that learners use, or if it is covered at all (see McEnery & Kifle 2002 for a similar study on strong modality markers). This study can be performed using complementary corpora, referred to by Meunier & Gouverneur (2009) as *pedagogical corpora*. They 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 error identification, after punctuation errors (34.76%), lexical issues are the second most common sources of disagreement (16.67%), then mistakes (clearly due to annotator distraction or format) are the third (12.80%), and different correction involving tense, mood and aspect are the fourth (10.98%). The fifth most frequent sources of disagreement in *ex aequo* are due to prepositions and to the annotator having 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 2.

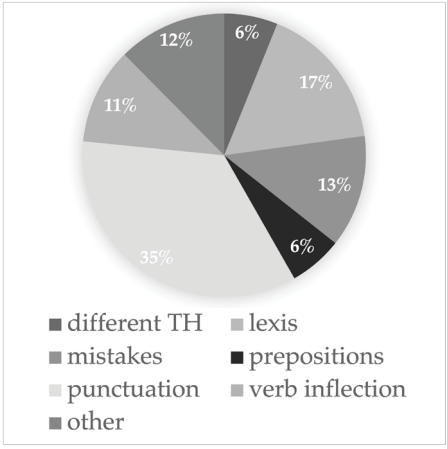

Figure 2. Most common sources of disagreement in error identification

Disagreement on lexical errors can occur when one of the two annotators identified a lexical error and the other did not (see Examples 6 and 8) – i.e., disagreement on error identification – or when both annotators identified a lexical error but they provided different solutions (see Example 7) – i.e., disagreement on error correction.

(6) LS: Matteo non poteva **vederla** senza fare niente. [*see her*]
Annotator 1: Matteo non poteva **stare a guardare** senza fare niente. [*stand by*]
Annotator 2: *No changes.*
*Matteo could not **see it** without doing anything.*

(7) LS: Mi alzai e **battai** questo tizio arrabbiato.
Annotator 1: Mi alzai e **colpii** questo tizio arrabbiato.
Annotator 2: Mi alzai e **battei** questo tizio arrabbiato.
*I got up and **beat** this angry guy.*

(8) LS: Ha **battuto** l'uomo che è caduto.
Annotator 1: Ha **colpito** l'uomo che è caduto.
Annotator 2: *No changes.*
*He **defeated** the man who fell.*

Example 6 shows different sensitivity to lexical usage in which *vederla* could be referred to the woman of the comic strip (feminine in Italian). Examples 7 and 8 show how different annotators' background can influence annotation.

In fact, the annotator knowing French identified *battere* as a semantic calque, from the French *Il a* ***battu*** *l'homme qui est tombé.* The French verb *battre* is a false friend of *battere*, because it means 'to hit' and not 'to defeat', so in Italian it should be rendered as *picchiare* or better *colpire* in this sentence. However, these sentences also work with the meaning of defeat, so we should ask the learners' intention to solve this disagreement.

As introduced above, disagreement can also be due to mistakes: unintentional annotators' errors induced by human distraction or by methods. Distraction largely affected the identification of spelling mistakes (Example 9) because, for cognitive reasons, they are especially easy to miss: in fact, it is generally recognised that the brain does not process what is read grapheme by grapheme, but processes the orthographic word as a whole, thus leading the reader to neglect spelling accuracy (Coltheart et al. 2001). Disagreement induced by format is illustrated in Example 10, in which both annotators marked the same tokens as to-be-edited, but in correcting the word-boundary error, they wrote the corrected word in a different slot.

(9)	Annotator 1	quando	0	
		all'improviso	1	all'improvviso
	Annotator 2	quando	0	
		all'improviso	0	

(10)	Annotator 1	ha	1	è
		suceso	1	successo
		un	0	
		distrasto	1	disastro
		per	1	perché
		che	1	
	Annotator 2	ha	1	è
		suceso	1	successo
		un	0	
		distrasto	1	disastro
		per	1	
		che	1	perché

Other errors induced by format are caused by the difficulty of consistently following the annotation rules (e.g., missing tokens to be edited in the following one when a linguistically-motivated correction would have foreseen the annotation in the previous token).

Format induced disagreement can also be closely linked to the nature of errors. In Example 11, in order to correct an error affecting two tokens, one annotator opted for a replacement of the auxiliary preserving the verb, the other changed the verb instead.

(11)				
	Annotator 1	Anche	1	Comunque
		ho	1	sono
		rimasto	0	
		de	1	a
		leggere	0	
	Annotator 2	Anche	1	Comunque
		ho	0	
		rimasto	1	continuato
		de	1	a
		leggere	0	

As a result, this counts as error identification disagreement, but the disagreement does not lie in the presence or absence of the error but rather in the way it is corrected and encoded in the experiment.

The fourth most frequent source of disagreement concerned verb inflection, in particular tense, mood and aspect (10.98%). In this case, the disagreement can be due to different perceptions 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 correct 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.

(12) 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.*

In the sentence reported in Example 12, one of the two annotators corrected the mood of the verb using the subjunctive in a subordinate headed by the verb *pensare* ('to think'). In Italian, with some verbs it is possible to use both subjunctive and indicative mood, slightly changing its semantics: using the former, it means 'to suppose'; using the latter, it means 'to be sure'. This is a case in which annotators' subjectivity influences error identification.

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 corrected it (see Example 13);

2. Both annotators identified an error but corrected it providing a different solution for the same text span (see Example 14);

3. The number of errors corrected by the two annotators does not coincide (see Example 15);

4. The learner's sentence is not clear enough and leaves considerable room for interpretation (see Example 16).

(13) 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*

*this man with big muscles **who was unconscious on the** ground was her boyfriend*

(14) 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.

(15) LS: Il brutto uomo dormendo era l'amico della donna. [sleep-GER-UND]

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.*

(16) 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.

*Another man sitting on a park bench **saw her spilling the surplus equips the shoulder** and went to conserve it.*

Annotator 1: Un altro uomo che stava seduto su una panca del parco **ha visto la donna portata sulla spalla** e è andato a salvarla.

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

Annotator 2: Un altro uomo che era seduto su una panchina del parco **ha visto che la donna si divincolava sopra la spalla** e è andato a salvarla.

*Another man who was sitting on a park bench **saw that the woman was crawling over his shoulder** and went to rescue her.*

As far as error correction is concerned, we analysed disagreement when both annotators agreed that at least one error was present in the token (i.e., 950 tokens with 240 differing). As a subsection of the disagreement resulted from the error identification experiment, we observed the same categories, but with different distributions. In this case, from the most common sources of disagreement to the least, we have: (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 3.

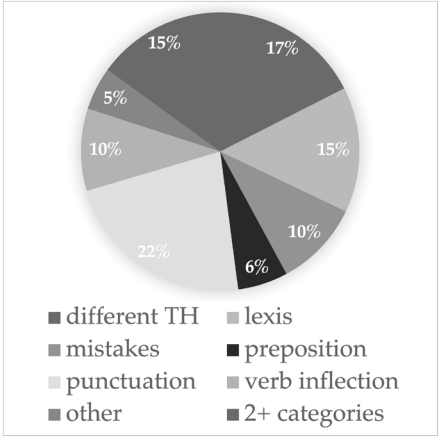

Figure 3. Most common sources of disagreement in error correction

4.1.1. *Apparent disagreement*

Revising the annotations, it emerged that apparent disagreement was higher than those classified as mistakes in the qualitative analysis (i.e., 12.80% and 10.98% in error identification and correction, respectively).

In fact, when the two annotators reviewed the disagreement together, they found that 40.12% of disagreement in the first annotation round was apparent. Apart from the mistakes (i.e., distraction and format issues) which were straightforwardly identified as apparent disagreement, 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 was recognised as "apparent". This apparent disagreement emerged during the revision of disagreement due to human errors. In Figure 4, the distribution of these categories of apparent disagreement is shown.

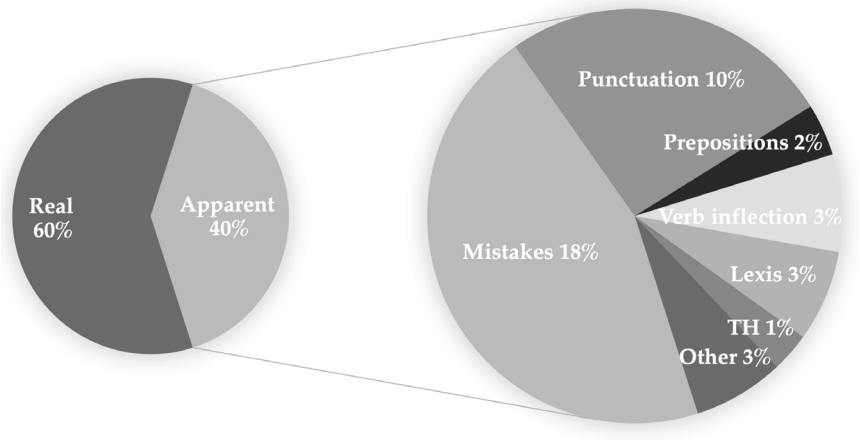

Figure 4: Distribution of categories involved in apparent disagreement

During the revision of the disagreement, which was conducted by the two annotators together, it became clear 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 both the identification of spelling, textual (such as verb tenses), and lexical errors. In fact, spelling errors require careful reading, while correcting textual errors, such as tense errors, requires a long-term working memory to consider textual consistency over a considerably wide span. A second reading would also be necessary in order to pay attention to consistency. Lexical errors are those that are most affected by annotators' inconsistency. This can be noticed in Examples 17 and 18, in which the same annotator corrected the word *suolo* to *terra* in Example 17, as adjunct of the verb *cadere* ('to fall'), but left it unchanged in Example 18 (again as argument of *cadere*, even though in this case the verb was corrected to *cadere* by the annotator).

(17) LS: L'uomo **cade** al **suolo** con la sua lingua fuori e senza spirito.
Annotator 1: L'uomo **cade** a **terra** con la lingua fuori e senza sensi.
*The man **falls** to the **ground** with his tongue out and unconscious.*

(18) LS: L'uomo è stato sull'**suolo** e Giacomo si ha sentito molto bene.
Annotator 1: 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 error identification and correction, apparent disagreement is caused by distraction (see Example 9) and format-induced (see Example 10) errors. Dis-

traction 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 19, where distraction caused one of the two annotators correcting the spelling of *gritava* to forget to rewrite the comma.

(19)	Annotator 1	La	0	
		donna	0	
		grivata,	1	gridava
		così	0	
		Io	1	io
		trovava	1	ho provato
		rescatarla	1	a liberarla
	Annotator 2	La	0	
		donna	0	
		grivata,	1	gridava,
		così	0	
		Io	1	io
		trovava	1	cercavo
		rescatarla	1	di salvarla

In Example 19, the verb *trovava* (meaning 'to find') is corrected by both annotators into 'to try', 'attempt to', but one of the two annotators also changed the verb tense from *imperfetto* into *passato remoto*, locating the action in a punctual past, and not in an imperfective past which does not specify start, end or duration. Eventually, *rescatarla*, a non-existent word in Italian, was corrected into two valid alternatives.

(20)	Annotator 1	portando	0	
		una	0	
		donna,	0	
		sulla	1	sulle
		sua	1	sue
		spalle	0	
	Annotator 2	portando	0	
		una	0	

donna	0	
sulla	0	
sua	1	
spalle	1	spalla

Format issues can create apparent disagreement when the annotators decide to correct the same error providing the correction in a different place (see Example 10). In addition, agreement errors corrected on the head or on the dependents, as in the case reported in Example 20, produce apparent disagreement for error identification, but real disagreement for error correction.

To reduce apparent disagreement due to distraction and format, some solutions could be provided. As far as distraction issues are concerned, one possible solution could be that of reducing the number of texts to be annotated 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.

4.1.2. Real disagreement

In both error identification and correction, we considered real disagreement when only one of the two annotators corrected an error – in error correction also if the provided correction is different and there is more than a valid solution, as happens with prepositions or lexis – or have a different TH in mind.

Register was always a source of real disagreement because the two annotators, correcting punctuation but also lexical issues, had different degrees of tolerance. See Examples 21-23.

(21) 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!!

(22) 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**.*

(23) 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 correction of the issues reported in Examples 21-23. In fact, the presence of the error can be detected only if register is taken into account. In Example 21, the correction of the two exclamation marks into one is mandatory in formal register, whilst *suolo pavimentato* in Example 22 might be used in user manuals or legal texts. Also the correction (or not) of *levato* in Example 23 depends on register. The verb *levarsi*, although marked as a term of common use in De Mauro (2016), was corrected into *alzato* by one annotator due to register issues (*levarsi* is considered 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 16) or requires specific language skills (e.g., borrowing from L1/L2 as reported in Examples 7, 8, 24 and 25). In Example 24 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, easily recognised this semantic calque and corrected it with *sono andato via*. Conversely, in Example 25, one of the two annotators, not knowing the Spanish verb plus clitic pronoun *derribarle*, corrected it using a distributional valid verb (considering also *salvare* after it). The other annotator, knowing Spanish, corrected *derribarle* using *batterlo*, meaning 'to take him down', preserving the meaning of the Spanish verb plus clitic but adding a further error (the missing coordinating conjunction *e*, 'and').

(24) LS: Ho chiesto scusa e **ho lasciato**.
*I apologised and **left**.*

(25)	Annotator 1	ma	0	
		Io	0	io
		può	1	potevo
		derribarle	0	batterlo
		salvare	1	e salvare
		a	1	
		la	1	
		donna	0	
	Annotator 2	ma	0	

Io	1	io
può	0	potevo
derribarle	1	cercare di
salvare	1	salvare
a	1	
la		
donna		

Disagreement caused by different language skills can also reveal itself where there is apparently no linguistic borrowing, as shown in Example 8.

By providing more clear guidelines, some of the above-mentioned real disagreements could be avoided. Disagreement due to a different way of correcting a sentence in the least invasive way possible, perhaps, could be avoided in some cases. Guidelines could be clearer in the indication of what *not to correct*, such as cross-lingual interference if the resulting sentence is acceptable as in Example 8 or register errors as in Example 23. In addition, some disagreements could be avoided if there are specifications in the guidelines on how to correct a sentence in the least invasive way possible. For example, the guidelines could be enriched by adding some answers to questions such as: (1) Is changing the verb to favour the syntactic dependency less invasive than leaving the verb and correcting the argument? (2) When dealing with disagreement errors, is it less invasive to correct the head or its dependents (Example 20)? However, this type of disagreement is very rare and since the majority of real disagreement is due to plausible alternatives, guidelines will never be exhaustive.

4.2. Error tagging

After the comparison of the second round of annotations, we identified as sources of real disagreement:

1. Annotator sensitivity in deepening the error annotation by marking each step necessary to arrive at the final corrected token (see Example 26);

2. Annotator identification of foreign borrowings (see Example 27);

3. Different interpretation of the error (see Example 28);

4. Identification of cascade errors (see Example 29).

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

TH: Si sentiva come un buon carabiniere 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: Derivation Adjective: forzo → forzoso;
Replacement adjective: forzoso → violento.

Annotator 2: Replacement Adjetive: forzo → violento.

(27) LS: **tratava** di ricordare dove la ha lasciato ma non ricordava niente si è desisperata moltissomo, 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: Spelling Double Consonants: tratava → trattava; Replace Verb Loanword: trattava → cercava.

Annotator 2: Spelling Double Consonants: tratava → trattava; Replace Verb: trattava → cercava.

(28) 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: Unnecessary Auxiliary: l'ho → l'.

Annotator 2: Spelling Pronoun: l'ho → l'.

(29) 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: Inflection Determiner Gender: la → il;
Inflection Determiner Gender: mia → mio.
Annotator 2: Inflection Determiner Gender cascade: la → il;
Inflection Determiner Gender: mia → mio.

The disagreement reported in Examples 26-28 cannot be solved by improving the guidelines, because it cannot be comprehensively defined.

In Example 26, 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.

Guidelines cannot comprehend lists of possible foreign borrowings/cross-linguistic interferences to inform annotators, thus disagreement in this case is inevitable, as the disagreement arising in Example 27, where the word *tratava* is probably a borrowing from Spanish *tratar*, meaning 'to try'. Depending on annotators' personal knowledge to identify these errors, it might be solvable by underspecifying the error tag using the two-letter tag (indicating only replacement verb) instead of the three-letter tag (indicating also that there is a cross-lingual interference). Moreover, the appropriateness of marking cross-linguistic influence is still debated in the context of descriptive taxonomies (such as the one used in VALICO-UD), as it would seem to be an attempt to explain an error rather than describe it.

Disagreement in Example 28 is due to a different way of classifying the same error. 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 Introduction). Annotator 1 considered only the auxiliary *ho* as substance, whilst Annotator 2 must have considered *l'ho* and classified it as a spelling error involving the pronoun (i.e., a spelling confusion error, such as the confusion between *where* and *were* in English).

Finally, in Example 29, 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 on 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 by clarifying in the guidelines how to deal with agreement errors.

5. Conclusion

In the present study, we reported on three inter-annotator agreement experiments to answer to three research questions referring to the experience gained in the development of the Valico-UD treebank:

1. Is error identification more reproducible using picture-elicited texts (i.e. within a constrained linguistic and extra-linguistic context)?

2. When different annotators agree on the presence of an error, do they agree also on its correction?

3. Is error tagging more reliable when based on an explicit TH (i.e. the corresponding corrected version of the learner sentence)?

The first question can be answered affirmatively. Looking at error identification, we obtained a κ value equal to 0.82 (Section 3.1). A similar result was achieved by Köhn & Köhn (2018), who also used a picture-elicited corpus. Lower results were obtained in experiments using texts elicited under less controlled environments (e.g., essays).

As also highlighted by previous research, disagreement emerged for lexical but also grammatical issues (Rosen et al. 2014; Del Río Gayo & Mendes 2018). In addition, in our case studies, punctuation was the most common source of disagreement. This suggests that punctuation might be a fuzzy area in foreign language teaching and L2 writing and this study might pave the way for future in depth investigations and experiments, involving more annotators (Section 4.1).

Concerning the second question, the answer is again positive, but agreement between annotators was lower ($\kappa = 0.69$) than on error identification. This is plausibly due to the non-deterministic nature of this task. However, analysing the sources of disagreement, it emerged that 40% of disagreement was apparent (Section 4.1.1). This suggests the usefulness of a second round of annotation in the methodology in tasks like this one, which require high level of concentration and specific skills.

Finally, error annotation with provided THs achieved a moderate agreement, despite the use of a very complex error tag set. The results obtained using the κ statistic confirm that the guidelines are suitable for the annotation task and clear enough to ensure a reasonable, objective interpretation of their content by two different annotators.

A lesson we learned from the three experiments is that a high percentage of apparent disagreement can occur and that the initially moderate agreement reached by the annotators can be meaningfully improved by considering the avoidance of human distractions and the increasing awareness of the complexity of the tag set used. In fact, after a second round of annotation, in the third experiment dealing with error tagging, the two annotators achieved perfect agreement ($\kappa = 0.95$, from an initial moderate agreement: $\kappa = 0.50$) confirming that THs indeed ensure reliability (Section 3.3). The remaining disagreement, as emerged from the qualitative analysis (Section 4.2), is due to error nature and to the need for highly specific guidelines.

References

Artstein, R. (2017). Inter-annotator agreement. In N. Ide & J. Pustejovsky (eds) *Handbook of Linguistic Annotation*. New York: Springer, 297-313.

Artstein, R. & Poesio, M. (2008). Inter-coder Agreement for computational linguistics. *Computational Linguistics – Association for Computational Linguistics* 34(4), 555-596.

Boyd, A. (2012). *Detecting & Diagnosing Grammatical Errors for Beginning Learners of German: From Learner Corpus Annotation to Constraint Satisfaction Problems*. PhD thesis. The Ohio State University.

Boyd, A. (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 (NLP4CALL)*, Stockholm, Sweden: LiU Electronic Press, pp. 10-22. URL: https://aclanthology.org/W18-7102 (last accessed on 5 May, 2023).

Colla, D., Delsanto, M. & Di Nuovo, E. (2023). ELICODE at MultiGED2023: fine-tuning XLM-RoBERTa for multilingual grammatical error detection. In *Proceedings of the 12th Workshop on NLP for Computer Assisted Language Learning (NLP4CALL)*, Faroe Islands, 22 May, 2023.

Coltheart, M., Rastle, K., Perry, C., Langdon, R. & Ziegler, J. (2001). DRC: a dual route cascaded model of visual word recognition and reading aloud. *Psychological Review* 108, 204-256.

Corder, S. P. (1982 [1981]). *Error Analysis & Interlanguage*. Oxford: Oxford University Press.

Corino, E. & Marello, C. (2009). Elicitare scritti a partire da storie disegnate: il corpus di apprendenti VALICO. In C. Andorno & S. Rastelli (eds) *Corpora di italiano L2: tecnologie, metodi, spunti teorici*. Perugia: Guerra, 113-138.

Corino, E. & Marello, C. (2017). *Italiano di Stranieri. I Corpora VALICO e VINCA*. Perugia: Guerra.

Costantino, M. (2009). Transcript-o'-matic: la tarscrizione dei testi per VALICO. In E. Corino & C. Marello (eds) *VALICO. Studi di linguistica e didattica*. Perugia: Guerra.

Dagneaux, E., Denness, Sh., Granger, S. & Meunier, F. (1996). *Error Tagging Manual. Version 1.1*. Louvaine-la-Neuve: Centre for English Corpus Linguistics, Université catholique de Louvain.

Dahlmeier, D., Ng, H. T. & Wu S. M. (2013). Building a large annotated corpus of learner English: The NUS corpus of learner English. *Proceedings of the Eighth Workshop on Innovative Use of NLP for Building Educational Applications*. Georgia, 22-31.

de Castilho, R. E., Mújdricza-Maydt, E. Yimam, S. M., Hartmann, S., Gurevych, I., Frank, A. & Biemann, C. (2016). A Web-based tool for the integrated annotation of semantic and syntactic structures. *Proceedings of the Workshop on Language Technology Resources & Tools for Digital Humanities (LT4DH)*. Osaka, Japan: The COLING 2016 Organizing Committee, 76-84. URL: https://aclanthology.org/W16-4011/ (last accessed on 5 May, 2023).

Del Río Gayo, I. & Mendes A. (2018). Error annotation in the COPLE2 corpus. *Revista da Associação Portuguesa de Linguística* 4, 225-239.

De Mauro, T. (2016). Il Nuovo Vocabolario di Base della Lingua Italiana. *Internazionale*. URL: https://bit.ly/3FqWNlk (last accessed on 5 May, 2023).

Díez-Bedmar, M. B. (2021). Error Analysis. In N. Tracy-Ventura & M. Paquot (eds) *The Routledge Handbook of Second Language Acquisition and Corpora*. New York: Routledge, 90-104.

Di Nuovo, E. (2023). *Introducing VALICO-UD. A Parallel, Learner Italian Treebank for Language Learning Research.* Bologna: Pàtron Editore.

Di Nuovo, E., Sanguinetti, M., Mazzei, A., Corino, E. & Bosco, C. (2022). VALICO-UD: Treebanking an Italian Learner Corpus in Universal Dependencies. *IJCoL* 8(1). URL: http://journals.openedition.org/ijcol/1007 (last accessed on 5 May, 2023).

Dobrić, N. & Sigott, G. (2014). Towards an Error Taxonomy for Student Writing. *Zeitschrift für interkulturellen Fremdsprachenunterricht* 19(2), 111-118.

Hovy, D. & Prabhumoye, S. (2021). Five sources of bias in natural language processing. *Language & Linguistics Compass* 15(8). https://compass.onlinelibrary.wiley.com/doi/epdf/10.1111/lnc3.12432 (last accessed on 5 May, 2023).

Hughes, A. & Lascaratou C. (1982). Competing criteria for error gravity. *English Language Teaching Journal* 36(3), 175-182.

Köhn, C. & Köhn A. (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)*, New Mexico, 121-132.

Landis, J. R. & Koch G. G. (1977). An application of hierarchical kappa-type statistics in the assessment of majority agreement among multiple observers. *Biometrics*, 363-374.

Lee, S. H., Dickinson, M. & Israel, R. (2012). Developing learner corpus annotation for Korean particle errors. In *Proceedings of the Sixth Linguistic Annotation Workshop*, 129-133.

Lennon, P. (1991). Error: Some problems of definition, identification & distinction. *Applied Linguistics* 12(2), 180-196.

Marello, C. (2011). Interpretare testi scritti composti a partire da storie disegnate. In K. Hölker & C. Marello (eds) *Dimensionen der Analyse von Texten und Diskursen. Dimensioni dell'analisi di testi e discorsi*. Vol. 1. LIT Berlin, 283-304.

McEnery, T. & Kifle, N.A. (2002). Epistemic modality in argumentative essays of second-language writers. In J. Flowerder (ed.) *Academic Discourse*. London: Longman, 182-195.

Meunier, F. & Gouverneur, C. (2009). New types of corpora for new educational challenges: Collecting, annotating and exploiting a corpus of textbook material. *Corpora and Language Teaching*, 179-201.

Meurers, D. & Müller, S. (2009). Corpora and syntax. In A. Lüdeling & M. Kytö (eds) *Corpus Linguistics* Vol. 2. Berlin: Mouton de Gruyter, 920-933. URL: http://purl.org/dm/papers/meurers-mueller-09.html (last accessed on 5 May, 2023).

Meurers, D. (2015). Learner corpora and natural language processing. In S. Granger, G. Gilquin & F. Meunier (eds) *The Cambridge Handbook of Learner Corpus Research*. Cambridge: Cambridge University Press, 537-566.

Nicholls, D. (2003). The Cambridge Learner Corpus: Error coding & analysis for lexicography and ELT. *Proceedings of the Corpus Linguistics 2003 Conference* Vol. 16, 572-581.

Quirk, R., Greenbaum, S., Leech, G. & Svartvik, J. (1985). *A Comprehensive Grammar of the English Language*. London: Longman.

Reznicek, M., Walter, M., Schmidt, K., Lüdeling, A., Hirschmann, H., Krummes, C. & Torsten, A. (2010). *Das Falko-Handbuch: Korpusaufbau und Annotationen* v.2. *Institut für deutsche Sprache und Linguistik*. URL: https://www.linguistik.hu-berlin.de/de/institut/professuren/korpuslinguistik/forschung/falko/Falko-Handbuch_Korpusaufbau%20und%20Annotationen.pdf (last accessed on 5 May, 2023).

Rosen, A., Hana, J., Štindlová, B. & Feldman A. (2014). Evaluating and automating the annotation of a learner corpus. *Language Resources and Evaluation* 48(1), 65-92.

Rozovskaya, A. & Roth D. (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*, 28-36.

Tetreault, J. & Chodorow, M. (2008). The ups & downs of preposition error detection in ESL writing. In *Proceedings of the 22nd International Conference on Computational Linguistics (Coling 2008)*, 865-872.

Volodina, E., Bryant, C., Caines, A., De Clercq, O., Frey, J. C., Ershova, E., Rosen, A. & Vinogradova, O. (2023). MultiGED-2023 shared task at NLP4CALL: Multilingual grammatical error detection. *Proceedings of the 12th Workshop on Natural Language Processing for Computer Assisted Language Learning (NLP4CALL)*, Faroe Islands, 22 May, 2023.

Appendix

Error tag set description

Unlike the two-letter tags employed in the CLC, the VALICO-UD tag set features a three-letter format. In accordance with Nicholls' explanation (2003: 573) "the first letter signifies the broad error category (e.g., wrong form, omission), while the second letter denotes the word class of the required word."

Additionally, we introduced a third letter, which, as outlined by Simone (2008: 303-341), represents the involved grammatical category, allowing for a more nuanced error description.

Adhering to this guiding principle, all errors are encoded using a predetermined set of letters that may occupy the first, second, or third position.

The **first letter** indicates the general error category or the type of correction needed to transition from the marked LS sequence (which could be a single token, a phrase, a clause, or a sentence, depending on the specific error) to the TH. The error types include D (derivation), F (form), I (inflection), M (missing), R (replace), S (spelling/mechanical), U (unnecessary), and W (word order).

The **second letter** specifies the orthographic, grammatical, or syntactic category of the required word. The permissible letters for 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).

The **third letter**, which is optional, provides further details about the error category indicated by the first letter. Depending on the first letter, the third position may include the following letters to denote specific features: A (aspect with I in the first position), B (co-occurring tense and mood or double letters, depending on whether the first position contains I or S, respectively), G (gender-related errors, distinguished with F or I in the first position), L (initially indicating a cross-linguistic influence, after IAA experiments assigned only to identifiable foreign words, distinguished with F or R in the first position), M (mood with I in the first position), N (number with I in the first position), O (collocation error or gerund error if the first letter is R or I, respectively), P (person with I in the first position), S (capitalization with S in the first position), T (tense or tokenization3 with I or S in the first position, respectively), W (multi-word expression with F, M, R, S, or U in the first position), and X (existential construction with F, M, R, S, or U in the first position).