Error annotation in the COPLE2 corpus

Iria del Río & Amália Mendes

University of Lisbon, Center of Linguistics - CLUL

Abstract:

We present the general architecture of the error annotation system applied to the COPLE2 corpus, a learner corpus of Portuguese implemented on the TEITOK platform. We give a general overview of the corpus and of the TEITOK functionalities and describe how the error annotation is structured in a two-level system: first, a fully manual token-based and coarse-grained annotation is applied and produces a rough classification of the errors in three categories, paired with multi-level information for POS and lemma; second, a multi-word and fine-grained annotation in standoff is then semi-automatically produced based on the first level of annotation. The token-based level has been applied to 47% of the total corpus. We compare our system with other proposals of error annotation, and discuss the fine-grained tag set and the experiments to validate its applicability. An inter-annotator (IAA) experiment was performed on the two stages of our system using Cohen's kappa and it achieved good results on both levels. We explore the possibilities offered by the token-level error annotation, POS and lemma to automatically generate the fine-grained error tags by applying conversion scripts. The model is planned in such a way as to reduce manual effort and rapidly increase the coverage of the error annotation over the full corpus. As the first learner corpus of Portuguese with error annotation, we expect COPLE2 to support new research in different fields connected with Portuguese as second/foreign language, like Second Language Acquisition/Teaching or Computer Assisted Learning.

Keywords*:* learner corpus, error annotation, second language acquisition, natural language processing

Palavras-chave: corpus de aprendentes, anotação do erro, aquisição de língua segunda, processamento de língua natural

1. Introduction

Error tagging has been proved to be an important aspect in learner corpora research, since it helps to identify problematic areas in the learning process (Granger, 2003) and provides useful data for many areas of study (Díaz-Negrillo & Thompson, 2013). Nevertheless, error tagging is not always present in learner corpora. We can identify at least two important causes for this fact: error tagging is a high time-consuming task that has to be performed manually; there are no standards, and taxonomies are a result of particular projects with specific interests (Díaz-Negrillo & Fernández-Domíguez, 2006). Error tagging techniques have evolved over the past few years from inline annotations with a unique interpretation, to standoff, multi-layer annotations with multiple error hypotheses. On the contrary, the conceptual design of taxonomies shows less development, with fewer changes in the categories and dimensions observed. Finally, the automatization of the annotation process is still a challenge.

We present the error annotation system designed for the COPLE2 corpus, as well as the different layers of annotation and the first results of its implementation. We show that our system takes advantage of the COPLE2 architecture as well as the TEITOK platform possibilities to reduce manual effort and produce a final annotation that follows the actual trends for error tagging. We discuss the methodology to create and evaluate the system and we describe its current structure. Since COPLE2 is the first corpus with error

annotation for Portuguese, we hope that our work will open new possibilities in the study of Portuguese as second/foreign language.

The paper is structured as follows: section 2 shows related work in error annotation; in section 3 we present COPLE2 corpus; section 4 describes the error annotation system error, its evaluation and our first annotation results; finally, section 5 presents the conclusions and future challenges.

2. Error annotation in learner corpora

The analysis of error tagging development leads to three relevant conclusions (among others). First, conceptual aspects related to the design of taxonomies show little variation through the years. Secondly, innovations have affected mainly the technical aspects of the annotation process. Finally, manual annotation is still the most common procedure and implies a high human effort.

Concerning the design of taxonomies, we can verify that most of them are: designed for written text, while schemes for oral data are scarce; grounded on three linguistic areas: spelling, grammar and lexis, leaving out others like phonetics or discourse; POS-centered, so certain linguistic units are undefined and certain levels of analysis are unexamined (Díaz-Negrillo & Fernández-Ramírez, 2006).

Moving to technical aspects, there has been an evolution from in-line and flat architectures to multi-layer standoff systems in all areas of corpus annotation. In first learner corpora with error annotation, like the Cambridge Learner Corpus (CLC) (Nicholls, 2003), or the International Corpus of Learning English (ICLE) at Louvain (Granger *et al.*, 2009), the tags were inserted in the learner text and a unique interpretation was proposed. We can see below an example of this type of annotation from the Louvain corpus:

(1) […] barons that (GVT) lived $had lived$ in those (FS) castels $castles$. (ICLE-Louvain; Dagneaux *et al.*, 1998: 16).

Lüdeling *et al.* (2005) point out two problems of this approach: (i) the number and category of annotation layers must be decided in the corpus design phase; (ii) it is difficult to annotate beyond the token level, that is, to annotate sequences of words. The first problem goes against one of the design principles for error annotation stated by Granger (2003), flexibility. The second problem can be solved if an XML format is used, as in FreeText (Granger, 2003: 470) or CLC. However, as noted again by Lüdeling *et al.* (2005), 'it is not possible to annotate overlapping ranges on different annotation layers since these cannot be mapped on a single ordered tree'. We can add a third problem of this methodology: annotations are mixed with the original learner text, which makes it difficult to manage the different levels of information in the corpus. The FALKO corpus (Lüdeling *et al.*, 2005) introduced a paradigm shift in the area. This system proposed for the first time a multi-layer and standoff design for error (and other types of) annotation in learner corpora. This architecture solved the problems that we mentioned above. On the one hand, the multi-layer design allows for the annotation of different types of information at the same time. For error annotation this means that different hypothesis for a given error can be proposed, where (in general) each layer corresponds to one level of interpretation. Besides this, the multi-layer architecture makes possible to add/remove layers when needed, which makes the system more flexible. On the other hand, standoff annotations make possible to store the different annotations apart from the original text. Finally, they allow for the annotation of sequences of words and also for managing overlapping ranges of text. Most recent learner corpora with error annotation show this type of design. We can find it in FALKO, MERLIN (Boyd *et al.*, 2014) (which uses the same target hypothesis than FALKO) or CzeSL (Rosen *et al.*, 2013).

Finally, one of the main problems of error tagging is that annotation is performed manually, being automatization one of the pending tasks. Different strategies have been tested to solve this drawback. Kutuzov & Kuzmenko (2015) explore the option of pre-processing learner texts with a spell-checker to identify

potential errors. Rosen *et al.* (2013) apply different tools designed for native language to the learner texts and compare their output with manual error annotation. They conclude that this strategy helps to identify potential errors and may even replace manual annotation in large-scale projects. Andersen (2011) explores the possibility of developing automatic rules for error detection and correction derived from manually error-annotated text. Unfortunately, those approaches solve only partial aspects of the problem and, until now and to the best of our knowledge, all learner corpora have mainly used human annotators for error tagging.

3. The COPLE2 corpus

COPLE2 (Mendes *et al*., 2016) is a learner corpus of Portuguese as a second/foreign language developed at the University of Lisbon. It contains written and oral productions of Portuguese learners with different L1s and proficiencies (15 languages, A1 to C1 levels), and provides rich TEI annotation through the TEITOK environment (Janssen, 2016).

The corpus contains complete metadata related to the learner (age, native language/s, years studying Portuguese, etc.), the topic of the text or the circumstances where the text was produced. The original hand-written texts and oral productions (audios) are accessible in the platform. All the changes made by the students (additions, deletions, transpositions of segments, etc.) are annotated, as well as the corrections suggested by the Portuguese teachers. The texts are tokenized, lemmatized and POS tagged using the Neotag tagger. All the information is stored together with the original texts in XML files that can be searched through the CQP query language.

4. Error annotation in the COPLE2 corpus

For error annotation in COPLE2 (del Río *et al*., 2016) we take advantage of the corpus architecture, the information already annotated, and the TEITOK possibilities to build an annotation system that: (i) deals with the challenges of error annotation; (ii) follows the current trends in the field; (iii) reduces and simplifies the manual annotation as much as possible and tries to automatize it.

Error annotation in COPLE2 is performed through two complementary systems: a flat, token-based system with three error categories that is applied inside the XML files, and a multi-word, fine-grained, standoff system that uses error tags. The token-based system makes possible a quick and simple annotation, supports complex queries using CQP and the visualization of the corrected text. But, what is more important: it allows for the automatic generation of the fine-grained annotation system's tags using all the information annotated in the corpus and the possibilities of the TEITOK platform. Next, we will describe both systems in detail and the relation between them.

4.1. Token-based coarse-grained annotation

In the token-based annotation, errors may be classified into three linguistic areas: orthographic, grammatical and lexical. Each area contains three fields of annotation: word form, lemma and POS. Depending on the problem/s affecting the original student form, the annotator has to select the affected linguistic area/s and introduce the required correct form/s (word form, POS, lemma). Multiple linguistic areas can be filled for a given token at the same time, for example, when a student form shows an orthographical problem, a grammatical problem and a lexical problem.

The orthographic layer is used if there is a spelling error in the student production, as illustrated in Figure 1: the student wrote *novedades* instead of *novidades* ('news'). The orthographically corrected form (nform) is introduced, as well as the corresponding POS (pos) and lemma (lemma), if necessary.

Nº 4 – 09/2018 | 225-239 | https://doi.org/10.26334/2183-9077/rapln4ano2018a42

Revista da Associação Portuguesa de Linguística

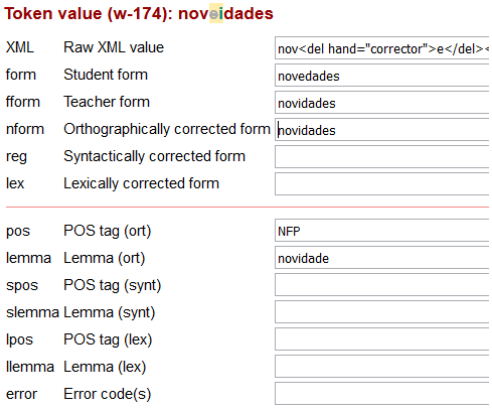

Figure 1. Annotation of an orthographic error

The grammatical layer operates if there is a grammatical error, that is: the word used by the student generates an ungrammatical utterance. Figure 2 shows an example: the student wrote *um cidade* ('a_MASC city') instead of *uma cidade* ('a_FEM city'), therefore, there is an agreement error which is annotated in the token corresponding to *um*. The syntactically corrected form is introduced (reg) as well as the corresponding POS (spos).

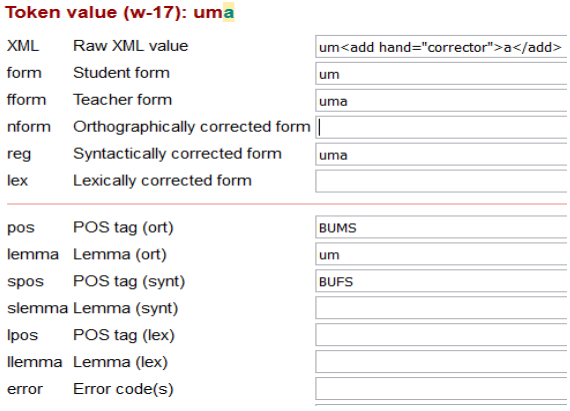

Figure 2. Annotation of a grammatical error

Note that in this case the field slemma is not annotated because the value for slemma is the same as the one indicated in the orthographic layer for lemma. The reason is that there is inheritance between layers, from the bottom (orthographic data) to the top (lexical data), and **only what is different from the previous level has to be annotated**. Therefore, if nform is empty, the system reads that its value is the same as form (there is no inheritance from the teacher's correction, fform). This is another advantage of the annotation system provided by TEITOK: the annotator only needs to annotate what is different, and not all the fields at each layer.

Nº 4 – 09/2018 | 225-239 | https://doi.org/10.26334/2183-9077/rapln4ano2018a42

Finally, the lexical layer is used if there is a lexical error in the student form, i.e., the word is grammatically correct, but it is not the natural word that a native speaker would use. Figure 3 shows an example where the student used the word *tropas* ('troops') in a context where *equipas* ('teams') was more accurate. In Figure 3, only llemma is annotated, because its value is different from the one in lemma; lpos has the same value as pos and, therefore, it remains empty.

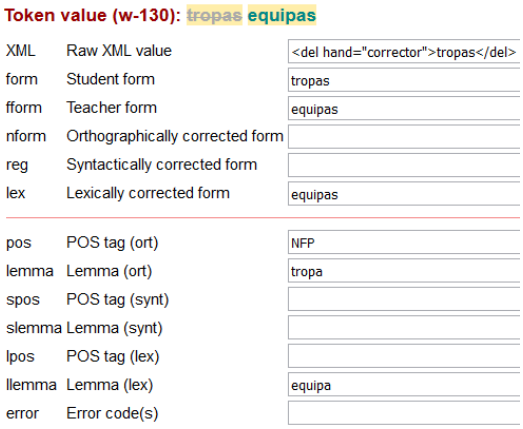

Figure 3. Annotation of a lexical error

The different layers are associated to different visualizations of the text that show the student's original text or the different corrections introduced. This way, it is possible to visualize the same text corrected at different layers, from the closer version to the original (only orthographic corrections) to the most modified one (orthographical, grammatical and lexical corrections).

The system described is a multi-layer annotation system, similar to the one presented in Rosen et al. (2013). Like in the Corpus of Czech as a Second Language, we define different tiers of annotation that work bottom-up, where different representations of the learner form take place. As we can see, there is a hierarchy in the level of interpretation assumed by the annotator at each tier, from errors with clear boundaries (orthographical and grammatical) to errors more open to interpretation (lexical ones), where it is sometimes hard to determine the "naturalness" of a given utterance. In our system, we assume a target hypothesis (Meurers, 2015) where the reference linguistic system is the target native language. At each tier, different transformations are applied to produce the equivalent native language form:

Orthographical level: the operations at this level are restricted to the word form and to punctuation marks. Punctuation, spelling and word boundaries problems are fixed, trying to generate the closest native form to the learner form. We include at this layer problems in inflectional or derivational suffixes, like in the learner form *estabilitamos*, instead of *estabelecemos* ((we) 'establish'). The final interpreted form is a valid word in standard European Portuguese.

Grammatical level: the operations at this layer are related to grammatical problems, that is, errors that go beyond the word and affect syntactic structures. Therefore, the annotator has to take into account the context surrounding the error. Examples are agreement problems (subject-verb, determiner-noun, noun-modifier, etc.), problems in the verb form (incorrect verbal tense, mode, etc.), subcategorization problems or problems in the POS selection. The final corrected form allows for a grammatically correct structure in the learner production.

Lexical level: the operations allowed at this layer affect mainly meaning. The word used by the learner is orthographically and grammatically correct, but it is not the most natural choice for a native speaker (see above the example of *tropas* in Figure 3).

All these annotations are integrated in the XML files with the students' texts and the other annotations mentioned in section 3.1. For errors that go beyond the token and do not fit into this schema, the first token of the wrong sequence is annotated with a special code that stands for "multi-token". This way, we ensure that all the errors are identified and classified.

Because of its simplicity and its integration in the TEITOK architecture, this system shows several advantages. First, from the taxonomical point of view, it is simple and general. The annotator decides between a limited number of possibilities (three types of errors with three possible corrections: word form, POS and lemma). There are no fined-grained error types with linguistic details to judge. Moreover, it is intuitive because the annotator decides on the error type by recovering the expected form in that particular context, i.e., the corrected form determines the error type. Furthermore, it allows for three different target hypotheses for a given error. Besides this, the system is perfectly integrated in the TEITOK environment: it allows for complex queries at the token level using all the information stored in the corpus through CQP; it makes possible a visual representation of the learner text corrected at three different levels (orthographic, grammatical and lexical).

4.2. Multi-word fine-grained annotation

The token-based annotation is simple and intuitive and well integrated with the TEITOK functionalities, but consists of a limited error tag set. It is therefore complemented with a fine-grained, standoff, multi-word system that uses error tags plus corrected forms. The annotations are stored standoff in XML files, can be applied to sequences of words and to overlapping fragments of text and provide detailed error categories. As we will see, one of the main advantages of this architecture is that most of these tags can be automatically generated (at least partially) from the token-based level of annotation.

The fine-grained tag set system was designed to complement the token-level annotation. As we explained in the previous section, from the technical point of view, error tags can be applied to multiple tokens and to overlapping fragments of texts, and are stored standoff in XML files. These aspects are in line with the current trends for error annotation (cf. section 2). From the theoretical point of view, the tag set makes possible a fine-grained classification of errors, which in turn allows for more specific queries concerning the different linguistic phenomena involved in error annotation (agreement, word order, use of incorrect POS, etc.). Finally, we can generate most of the tags automatically (at least partially) from the token-level annotations, making the annotation process at this fine-grained level quick and simple.

The tag set designed for this system is similar to the taxonomies described in Tono (2003), Nicholls (2003) or Dagneaux *et al*. (2005). To define its categories, we performed in the first place a systematic review of the state-of-the art in the subject. As we explained in section 2, we found that most of the error tagging systems are similar from the theoretical point of view: they are designed for written texts and they use roughly the same linguistics areas: orthography, grammar and lexis. The main difference comes when we look at the error categories considered by the different systems (for a detailed comparison of different tag sets, see Díaz-Negrillo & Fernández-Domíguez, 2006).

We tried to follow some theoretical principles when designing our tags. First of all, we tried to be as general as possible, that is, we avoided creating very specific tags. The reason for this decision is that a general tag can be always specified if necessary, creating several sub-tags, while the inverse path is much more complicated. Besides this, the annotation process very often reveals issues that were not previously planned. Working with a flexible schema, as noted by Granger (2003), is crucial in those scenarios. Secondly, we tried to be as neutral as possible concerning the theoretical framework of the schema. This aspect makes

our schema accessible for researches with different backgrounds. Finally, being aware of the distance between designing a theoretical schema for annotation and applying that schema to real data, we performed an annotation experiment to develop a pilot error taxonomy.

In the pilot experiment, we annotated 36 texts (7,073 tokens) from COPLE2. We tried to create a representative sample of the native languages and proficiency levels present in the corpus. We first identified the errors in those texts; secondly, we defined the necessary categories to classify them. Considering the small amount of texts we used in the experiment, we decided to complete the pilot taxonomy with categories that account for linguistic phenomena that we expected to find in COPLE2, considering the results of similar annotation projects like (Granger, 2009). We ended up with a pilot taxonomy containing 37 tags. The next step was to evaluate the performance of this pilot schema. We describe the process and its particularities in section 4.3.

4.2.1. Description of the tag set

The tag set is structured in two levels of information: (i) general linguistic area affected; (ii) error category (and subcategories in some cases) affected. Level 1 includes (for the moment) the same three linguistic areas as the token-based system: Orthographic (includes spelling and punctuation errors), Grammatical (includes agreement errors; errors affecting verb tense, mode, etc.) and Lexical (lexical choice errors). Level 2 accounts for common error categories like agreement or wrong POS. The tags are position-based, that is, each position in the tag corresponds to a specific level of information. The first letter corresponds to the general linguistic area affected and the subsequent letters to error category and subcategories (if applicable). For example, for agreement errors affecting gender, the tag is "GAG" that is:

Level 1- Linguistic level affected= Grammar= G +

Level 2 - Error category = Agreement = A + Error subcategory = Gender = G

Final tag = GAG

Currently, the tag set contains **38 tags**, with the following distribution:

- Orthographic tags= 11.
- Grammatical tags = 25.
- Lexical tags = 2[1].

Due to the flexible structure of the schema, it is possible to modify the number of tags if required. In fact, our preliminary results on error annotation using the token-based system (see section 4.4) suggest that we need to consider the inclusion of new linguistic levels. The main reason is that there are some phenomena in the corpus that cannot be precisely described using the three linguistic levels above. It can be the case, for example, of errors affecting the discourse structure.

As a schema designed for error annotation in learner corpora, our tag set follows the principles stated in Granger (2003). That is, COPLE2 tag set is:

1 **Consistent**: we have evaluated the annotation system, obtaining a general value of inter-rater agreement $\kappa = 0.84$ (see section 4.3).

2 **Informative**: each tag accounts for a clearly defined linguistic issue and is defined in the guidelines with examples. The number of tags (38) is reduced and manageable.

1 Examples of the tags with examples from the corpus can be found in the final Appendix.

Nº 4 – 09/2018 | 225-239 | https://doi.org/10.26334/2183-9077/rapln4ano2018a42

Revista da Associação Portuguesa de Linguística

3 **Flexible**: the schema uses hierarchical categories, and it is structured in two levels. On the other hand, it is easily adjustable.

4 **Reusable**: it accounts for general categories that describe common errors in three linguistic areas. It can be adaptable to close languages like Spanish.

The tag set is described in detail in the guidelines of the project. As explained in section 4.3, the evaluation of the error tagging system revealed weaknesses of the schema and allowed for a crucial reconfiguration of the guidelines. Besides this and thanks to the flexible nature of our schema, the Guidelines are constantly enriched and detailed during the annotation process (see section 4.4).

4.2.2. Automatic generation of tags

We have explained that one of the main problems of error tagging is that it is a high-time consuming task because it is a process that has to be performed manually. On the other hand, error tagging is also a highly interpretable task (see section 4.3) where, in some cases, the object to be annotated (the error) can be linguistically interpreted in different ways, making possible to apply different tags to the same error. This fact can lead to divergences in error annotation, causing low rates of IAA (see section 4.3). The use of automatic techniques to perform the annotation could help to solve the two problems described: on the one hand, it will certainly reduce the annotation time; on the other hand, it will allow for a systematic annotation of the same phenomena. We have not arrived to a fully automatic annotation system in COPLE2, but we have designed an architecture that allows for the automatic generation (at least partially) of most of the tags in the multi-word fine-grained level. This is far from ideal but it reduces considerably the annotation time and ensures a robust and coherent annotation.

The automatic generation of tags is performed (at least partially, as we will see) comparing the original form of the student with the corrections introduced at the token level. The first letter of the tag can be always generated just checking the linguistic area where the corrections were added (remember that the linguistic areas considered are the same at the token level and at the multi-word level). This fact allows for an unambiguous assignation of the first letter of the tag. The subsequent letters can be inferred in most of the cases using the other linguistic information annotated in the corpus. Let's see an example. One of the most common problems in the corpus involves the wrong use of accentuation marks. For example, there is a case in the corpus where the student wrote *simpátía* instead of *simpatia* ('simpathy').

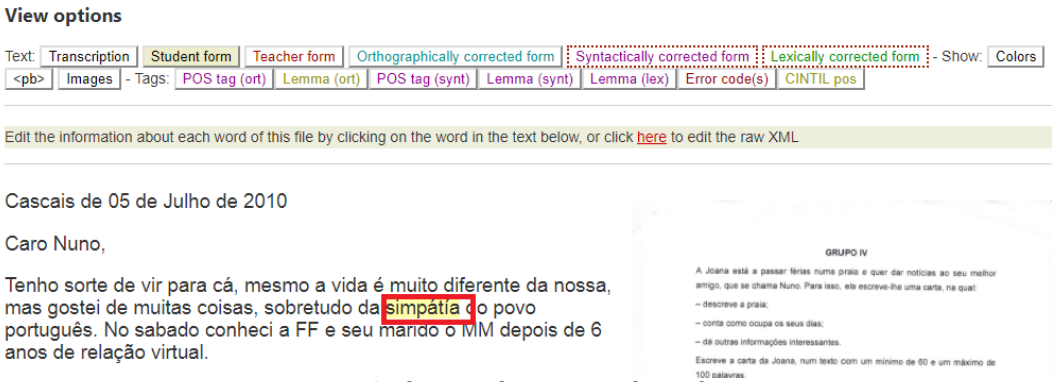

Figure 4. Student text showing an orthographic error

Nº 4 – 09/2018 | 225-239 | https://doi.org/10.26334/2183-9077/rapln4ano2018a42

This error is annotated at the token level as:

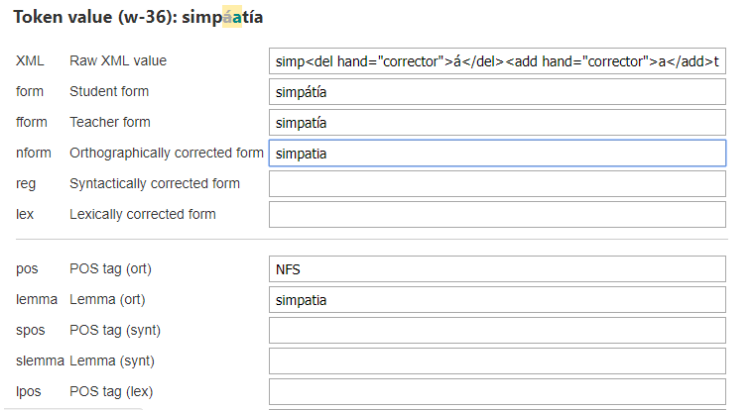

Figure 5. Token-based annotation for the error *simpátía*

As we can see, the annotator introduced the correct word form in the field nform, which is part of the orthographic layer.

In the tag set, "Spelling Stress Mark", SS, covers this type of spelling errors. To annotate *simpátia* with the tag SS plus the correct form *simpatia*, we follow this process: first, we check the linguistic area that contains annotations at the token-based level. As we can see in the previous figure, the token level annotation corresponds to the orthographical layer, therefore, we know that the first letter of our tag corresponds to the linguistic area "Orthography", that is, the first letter of the tag is "S". To determine the error category, we check the three possible fields that can be annotated for each linguistic layer: word form, pos or lemma (see section 4.1). In this case, we have an annotation at the nform field. Comparing this annotation with the original student form we can see that the difference between the two word forms affects accentuation marks, therefore, we know that the error category corresponds to "Stress Mark" and we have the second letter of our tag, "S". This way we have inferred that the error tag is SS. With this strategy, we take advantage of the TEITOK and COPLE2 possibilities to automatically produce a detailed error annotation with low manual effort. This is a good example of the possibilities that COPLE2 offer to apply Natural Language Processing techniques to the annotation process.

We will perform this inference through conversion scripts that take as input all the token-based XML annotations and generate as output a new XML with the corresponding standoff annotations (tag + correction suggested). We have done the calculations and it is possible to generate (fully or partially) 29 of the 38 tags. From the remaining 9 tags, 6 go beyond the token, affect mainly the verbal phrase and correspond to rare errors. One example is the tag GVH, for errors affecting verbal periphrasis, like in:

(2) *Espero que não va acontecer* > va a acontecer ('I hope it is not going happen > going to happen').

The other 3 tags are token-based but require human interpretation.

4.3. Evaluation of the error annotation system

In general, no information is provided about the number of annotators or the performance of the error tagging systems used in learner corpora. This gap can constitute a problem since, as it has been showed (Tetreault & Chodorow, 2008; Rozovskaya & Roth, 2010), even native speakers may differ considerably with respect to what constitutes acceptable or correct usage. Meurers (2009) discusses the issue of verification of

Nº 4 – 09/2018 | 225-239 | https://doi.org/10.26334/2183-9077/rapln4ano2018a42

error annotation validity, pointing out the lack of studies that analyze inter-annotator agreement (IAA) in the manual annotation of learner corpora, and considering this fact a serious impediment for the development of annotation tools. On the other hand, there has not been yet an in-depth discussion about the metrics to be used for this kind of evaluations. Rosen *et al.* (2013) claim that: "There is no widely accepted metric evaluating the consistency of annotation of learner corpora." The common metric used is Cohen's kappa (κ) for IAA, although there is some discussion in the literature concerning the adequacy of this measure for certain linguistic classification tasks (Arstein & Poesio, 2008).

In order to test our error annotation system, we performed an annotation experiment. Our goal was to test the reliability of the system as well as to identify possible adjustments and refinements required. Two experienced annotators, native speakers of Portuguese, annotated two samples of texts extracted from COPLE2 using the GATE annotation tool (Cunningham *et al.*, 2011). Simple guidelines describing the main issues of the annotation system described in del Río *et al*. (2016) were provided to the annotators.

Each sample of texts covered most of the languages and proficiencies included in the corpus. The token-based system was tested in one sample and the fine-grained tag set in another one. Table 1 shows the size of each sample[2].

	Texts	Tokens
Token-based sample	14	2,385
Fine-grained tag set sample	10	2,118

Table 1. Description of the two samples used in the IAA experiment

In the token-based annotation sample, the annotators had to identify errors affecting only one token. Each error had to be classified into one of the categories that we described in section 4.1: orthographical, grammatical or lexical, and a correction had to be provided. The same token could have more than one error. In the fine-grained tag set sample, the annotators could identify errors beyond the token. They had to classify each error into one of the 37 error categories of the pilot tag set (cf. section 4.2), and add a correction. A single text span could contain multiple errors.

In our evaluation, we measured error classification using Cohen's kappa. We chose κ because it is the common metric used in the learner corpora field, although we are aware of the discussion concerning the adequacy of this measure for certain linguistic classification tasks. The results we obtained are presented below.

κ without correction	κ with correction
0.86	0.85

Table 2. IAA for the token-based sample

We can see in table 2 that general kappa values were good for the token-based annotation, with and without correction. As expected, considering or not considering the correction as a variable had some impact on the general results.

κ without correction	κ with correction
0.85	0.84

Table 3. IAA for the fine-grained tag set sample

[2] We know that the size of the samples was limited and far from ideal, but we were limited by the costs of manual annotation and by the fact of testing two different systems with many variables involved.

The general kappa value was also good for the fine-grained tag set sample, with and without correction. The negative effect of corrections is visible here too.

We also evaluated the classification of errors considering only the linguistic category (Orthography, Grammar and Lexis). What we found is that, for both annotation systems, agreement is higher for orthographic errors, lower for grammatical errors and much lower for lexical errors. Table 4 below presents the total cases of agreement and disagreement, as well as the observed agreement by linguistic area for the token-based sample:

Linguistic Area	**Agreement**	**Disagreement**	**Total**	**Observed Agreement**
Orthography	280	14	294	0.96
Grammar	242	19	261	0.93
Lexis	36	15	51	0.7
Total	558	48	606	0.92

Table 4. Error classification: raw numbers and observed agreement per linguistic area (without correction)

The analysis of the disagreements in both samples showed some tendencies. For example, in the evaluation of the fine-grained system, we found that *Grapheme Substitution*, an orthographical tag (SGS), was usually confused with different grammatical categories like *Wrong Category* (GC), *Verb Tense-Mode* (GFM) or *Agreement Gender* (GAG).

(4) *há 3 anos e meia* (‘three years_masc-sing and a half_fem-sing ago’).

Annotator A: tag: SGS; correction: *meio* (‘half_masc-sing’).

Annotator B: tag: GAG; correction: *meio* (‘half_masc-sing’).

In this example, both annotators introduced the same correction, *meio* (‘half_masc-sing’), but the interpretation of the error was different: for Annotator A *meia>meio* is a spelling problem, but for Annotator B *meia>meio* is a grammatical problem affecting gender agreement.

The results of our experiment were in general positive. First of all, we proved that, even with general guidelines, two experienced annotators reached good IAA κ values when applying the COPLE2 error annotation system. Although a more extensive evaluation would be ideal, we think that this first evaluation indicates that the system is reliable. Moving to error classification, we confirmed that the level of disagreement increases from Orthography errors (low) to Lexical errors (high), as predicted in del Río *et al.* (2016). Besides this, it has been proved that considering correction as a variable decreases the performance of the error annotation system. Finally, and more important, the evaluation showed which tags could be commonly confused, allowing for an improvement of the annotation guidelines.

4.4. Results of the error annotation at the token level

We have started the annotation of the corpus at the token level. So far, we have annotated 442 texts (47% of the total files), corresponding to 72,858 tokens (42.5% of the total tokens in the corpus). We have added 14,984 annotations. Of these, 13,581 are token-based (91%) and 1,403 are labeled as going beyond the token (9%). The token-based annotations have the following distribution: 6,432 orthographical errors; 5,881 grammatical errors; 1,268 lexical errors.

For the moment, our results indicate that the token-based representation may account for most of the errors found. However, these results may be biased by the fact that the annotator has tried to adjust the annotation to the token-based representation and we think that a deeper analysis is necessary to draw precise conclusions. For example: we have annotated predicative adjectives with disagreement problems at the token level, as in:

(5) *As praias são muito <u>lindos</u>*, […] > *lindas* ('The beaches_FEM are very beautiful_MASC > beautiful_FEM').

In this case, the error is visible on the adjective although the error goes beyond the token, affecting a grammatical structure (the sentence, in this case). Technically it is possible to annotate at the token level, but conceptually maybe this is not the ideal representation of the error. One simple example of an error that cannot be annotated at the token level is the following, where two tokens have to be corrected into one:

(6) *Foi uma expêriencia que eu nunca <u>tenho esquecido</u>* > esqueci ('It was an experience that I haven't forgotten > forgot').

Our next step will be to automatically generate the tags of the fine-grained tag set from the token-based annotations, as described in section 4.2.2.

5. Conclusions and future work

We have implemented a system for error annotation in COPLE2 that attempts to reduce manual effort by taking advantage of the corpus information and the possibilities of the TEITOK environment. We have started to apply the system, and we have already annotated 47% of the corpus at the token level, being COPLE2 the first Portuguese learner corpus with error annotation. From in-line, token-based and flat annotations we will generate automatically standoff, multi-word annotations, which will contain position-based tags covering 38 error types. Most of the tags will be fully generated using this automatic approach, although some of them will require manual work.

Currently, we continue annotating at the token level and developing the scripts for the automatic generation of tags. Besides this, we have identified some future lines of work. First of all, we need to explore how to transform the multi-token in-line annotations into tags, reducing as much as possible the manual effort. One way could be to identify error patterns (using information concerning the word form, POS, word order, etc.) in multi-token structures that correspond to a certain tag, automatizing the generation. A second line of work is related to the addition of new linguistic areas for error annotation, like semantics or discourse. In fact, some annotation cases at the token level suggest the need of higher linguistic levels of abstraction in the scheme.

We believe that error annotations (token-based plus error tags) together with all the information already stored in the corpus (metadata, student's modifications, teacher's corrections) will allow for complex and rich linguistic queries in COPLE2. We expect that this information can be useful for researchers of different fields like Second Language Acquisition, Foreign Language Teaching and Learning or Computer Assisted Language Learning.

Acknowledgements

This work was partially supported by Fundação Calouste Gulbenkian (Proc. nr. 134655), Fundação para a Ciência e a Tecnologia (project PEst-OE/LIN/UI0214/2013; postdoctoral research grant SFRH/BPD/109914/2015) and Associação para o Desenvolvimento da Faculdade de Letras da Universidade de Lisboa (ADFLUL).

References

Andersen, Ø. (2011) Semi-automatic ESOL error annotation. English Profile Journal, 2. Christ, O., Schulze, B., Hofmann, A. and Koenig, E. (1999). *The IMS Corpus Workbench: Corpus Query Processor (CQP): User's Manual*. Institute for Natural Language Processing. University of Stuttgart. (CQP V2.2).

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

Boyd, A., Hana, J., Nicolas, L., Meurers, D., Wisniewski, K., Abel, A., Schöne, K., Štindlová, B. & Vettori, C. (2014) The MERLIN corpus: Learner Language and the CEFR. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation* (LREC'14), pp. 1281-1288.

Cunningham, H. *et al.* (2011) Text Processing with GATE (Version 6). University of Sheffield Department of Computer Science. 15 April 2011.

Dagneaux, E., Denness, S. & Granger, S. (1998) Computer-aided Error Analysis. *System*, 26, pp. 163-174.

Dagneaux, E., Denness, S., Granger, S., Meunier, F., Neff, J. & Thewissen, J. (Eds.) (2005) *Error Tagging Manual. Version 1.2.* Centre for English Corpus Linguistics. Université Catholique de Louvain.

del Río, I., Antunes, S., Mendes, A. & Janssen, M. (2016) Towards error annotation in a learner corpus of Portuguese. In *Proceedings of the 5th NLP4CALL and 1st NLP4LA workshop in Sixth Swedish Language Technology Conference (SLTC).* Umeå University, Sweden, 17-18 November.

Díaz-Negrillo, A. & Fernández-Domíguez, J. (2006) Error Tagging Systems for Learner Corpora. *RESLA*, 19, pp. 83-102.

Granger, S. (2004) Computer learner corpus research: current status and future prospects. In U. Connor & T. Upton (Eds.), *Applied Corpus Linguistics: A Multidimensional Perspective*. Amsterdam & Atlanta: Rodopi.

Díaz-Negrillo, A. & Thompson, P. (2013) Learner corpora: Looking towards the future. In N. Ballier, A. Diaz-Negrillo, & P. Thompson (Eds.) *Automatic treatment and analysis of learner corpus data*. Amsterdam & Philadelphia: John Benjamins, pp. 249–264.

Granger, S. (2003) Error-tagged Learner Corpora and CALL: A Promising Synergy. *CALICO Journal*, 20 (3). Special issue on error analysis and error correction in computer-assisted language learning, pp. 465-480.

Granger, S., Dagneaux, E., Meunier, F., Paquot, M. (Eds.) (2009) *International Corpus of Learner English. Version 2*. UCL: Presses Universitaires de Louvain.

Janssen, M. (2016) TEITOK: Text-Faithful Annotated Corpora. In *Proceedings of LREC 2016*, Portorož, Slovenia.

Kutuzov, A. & Kuzmenko, E. (2015) Semi-automated typical error annotation for learner English essays: Integrating frameworks. In *Proceedings of the 4th workshop on NLP for Computer Assisted Language Learning at NODALIDA 2015*, Vilnius, 11th May, 2015, Volume , Issue 114, 2015-05-06, pp. 35-41.

Lüdeling, A., Walter, M., Kroymann, E. & Adolphs, P. (2005) Multi-level annotation error annotation in a learner corpora. In *Proceedings of Corpus Linguistics 2005* 1, Birmingham (England), July 2005, pp. 14-17.

Mendes, A., Antunes, S., Janssen, M. & Gonçalves, A. (2016) The COPLE2 Corpus: a Learner Corpus for Portuguese. In *Proceedings of LREC 2016*, Portorož, Slovenia.

Meurers, D. (2009) On the automatic analysis of learner language: Introduction to the special issue. *CALICO Journal* 26(3), pp. 469-473.

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 University Press, pp. 537-566.

Nicholls, D. (2003) The Cambridge Learner Corpus – error coding and analysis for lexicography and ELT. In D. Archer, P. Rayson, A. Wilson & T. McEnery (Eds.) *Proceedings of the Corpus Linguistics 2003 Conference*. Lancaster University, pp. 572-581.

Rosen, A., J. Hana, B. Štindlová & A. Feldman (2013) Evaluating and automating the annotation of a learner corpus. *Language Resources and Evaluation,* pp. 1-28.

Rozovskaya, A., & Roth, D. (2010) Annotating ESL errors: Challenges and rewards. In *Proceedings of NAACL'10 workshop on innovative use of NLP for building educational applications*. University of Illinois at Urbana-Champ.

Tetreault, J., & Chodorow, M. (2008) Native judgements of non-native usage: Experiments in preposition error detection. In *COLING workshop on human judgements in computational linguistics*. Manchester.

Tono, Y. (2003) Learner corpora: Design, development and applications. In D. Archer, P. Rayson, A. Wilson & T. McEnery (Eds.) *Proceedings of the Corpus Linguistics 2003 Conference*. Lancaster University, pp. 800-809.

Appendix

Examples of orthographic tags

LINGUISTIC CATEGORIES	POSITION BASED	DESCRIPTION OF THE TAG	ERROR EXAMPLES
Spelling_StressMark	S+S	It is used when there is an extra or a missing accent mark.	*diferentes* ***paises*** *e povos*
Spelling_Grapheme_Addition	S + G+A	One or more graphemes is/are erroneously added. This includes the doubling of consonants and vowels (but not in inflectional/derivational suffixes)	***practicamente***
Spelling_Grapheme_Deletion	S+G+D	This tag includes all errors concerning the choice of the correct grapheme, and it is used if one or more graphemes is/are missing at the beginning or the middle of the word (but not in inflectional/derivational suffixes). This includes the simplification of consonant groups.	***qerem***
Spelling_Grapheme__Substitution	S+G+S	This tag includes all errors concerning the choice of the correct grapheme, and it is used if a grapheme is wrongly used instead of another grapheme (again, this does not apply to inflectional/derivational suffixes).	***spportei***
Spelling_Grapheme_Transposition	S+G+T	Two graphemes have exchanged positions.	***apíses***
Spelling_Capitalization	S + C	The word is written in lower case and should be capitalized or the opposite.	*a liberdade de que fala* ***pessoa***
Spelling_WordBoundarySplit	S+B+S	One word incorrectly split.	***última mente*** *não falamos*

Examples of grammatical tags

LINGUISTIC CATEGORIES	POSITION BASED	DESCRIPTION OF THE TAG	ERROR EXAMPLES
Grammar_UnnecessaryWord	G + U	The written word is unnecessary.	*eu vou* ***a*** *organizar uma festa*
Grammar_OmittedWord	G + E	Omission of a necessary word.	*[a] fala do dia a dia do cidadão*
Grammar_WrongWord	G + W	Cases where the lemma (not the POS) selected by the learner is not correct, according to the grammatical surrounding context.	*ninguém sem tempo* ***por*** *nada*
Grammar_WrongCategory	G + C	Wrong POS selection.	*não vive nas* ***selvagens*** *com tantos riscos*
Grammar_Agreement_Gender	G + A + G	Agreement error afecting gender.	*os ideáis* ***humanitarias***
Grammar_Agreement_Number	G + A + N	Agreement error afecting number.	*tem paisagens* ***lindíssima***
Grammar_Agreement_Gender&Number	G + A + B	Agreement affecting gender and number.	*pode ser palavras* ***bo***
Grammar_WordOrder	G + O	The error affects the order of constituents.	*não lêem* ***livros muitos***
Grammar_Verb_Tense	G + F + T	Incorrect tense.	*Sempre* ***havia*** *e sempre havrá*
Grammar_Verb_Mode	G + F + M	Incorrect mode.	*uma battaria nova* ***deva*** *durar*
Grammar_Verb_Tense&Mode	G + F + Z	Incorrect tense and mode.	*senhor prometeu-me que* ***irão*** *funcionar*
Grammar_Verb_FiniteNoFinite	G + F + F	Confusion between finite and no finite.	*cada vez mais* ***melhorar***
Grammar_VerbalConstruction_Periphrasis	G + V + H	Error in periphrasis.	*E espero que não* ***va acontecer***
Grammar_VerbalConstruction_Clitization	G + V + K	Error in clitized forms.	***descobrimos*** *-as*
Grammar_PronounClitic_Case	G + F + C	Error in case (pronoun).	*visitou-* ***lhe*** *ontem*
Grammar_Noun_Number	G + F + N	Error in number (noun) when the noun has to be singular or plural (no for agreement structures).	*Minha* ***última féria*** *ésteve*

Examples of lexical tags

LINGUISTIC CATEGORIES	POSITION BASED	DESCRIPTION OF THE TAG	ERROR EXAMPLES
Lexical_LexicalChoice	L + C	Used word exists in the language and the POS is correct, but the lemma is not right since it is not semantically correct in the given context.	*Se não* ***tiver*** *medidas de proteção*

Nº 4 – 09/2018 | 225-239 | https://doi.org/10.26334/2183-9077/rapln4ano2018a42

Revista da Associação Portuguesa de Linguística