Towards error annotation in a learner corpus of Portuguese

Iria del Río[1], Sandra Antunes[1], Amália Mendes[1] and Maarten Janssen[2]

[1] University of Lisbon – CLUL
[2] University of Coimbra – CELGA-ILTEC
iagayo@gmail.com, sandra.antunes@clul.ul.pt,
amalia.mendes@clul.ul.pt, maartenpt@gmail.com

Abstract

In this article, we present COPLE2, a new corpus of Portuguese that encompasses written and spoken data produced by foreign learners of Portuguese as a foreign or second language (FL/L2). Following the trend towards learner corpus research applied to less commonly taught languages, it is our aim to enhance the learning data of Portuguese L2. These data may be useful not only for educational purposes (design of learning materials, curricula, etc.) but also for the development of NLP tools to support students in their learning process. The corpus is available online using TEITOK environment, a web-based framework for corpus treatment that provides several built-in NLP tools and a rich set of functionalities (multiple orthographic transcription layers, lemmatization and POS, normalization of the tokens, error annotation) to automatically process and annotate texts in xml format. A CQP-based search interface allows searching the corpus for different fields, such as words, lemmas, POS tags or error tags. We will describe the work in progress regarding the constitution and linguistic annotation of this corpus, particularly focusing on error annotation.

1 Introduction

The COPLE2 corpus[1] is a written and spoken learner corpus of Portuguese as a foreign or second language (FL/L2) that aims at providing empirical data for the teaching and learning of this language. Several learner corpora have been compiled for English, such as the International Corpus of Learner English (Granger et al., 2009), the Longman Learner's Corpus, or the Cambridge Learner Corpus (Nicholls, 2003). The importance of such empirical data has been increasingly recognized for studies in Second Language Acquisition and language teaching/learning. Recently, we have seen a substantial growth in this area regarding other languages besides English. Concerning Romance languages, there are already some corpora and resources for French (Delais-Roussarie & Yoo, 2010), Spanish (Lozano, 2009) and Italian (Boyd et al., 2014). In the case of the Portuguese language, there are also some initiatives in the compilation of learner corpora. The corpus *Recolha de dados de Aprendizagem do Português Língua Estrangeira*[2], that follows the precursor work developed in Leiria (2001), was compiled at the School of Arts and Humanities of the University of Lisbon, and includes 470 texts and 70,500 tokens. The *Corpus de Produções Escritas de Aprendentes de PL2*[3], compiled at the University of Coimbra, is constituted by 516 texts and 119,381 tokens. Finally, the *Corpus de Aquisição de L2*[4], compiled at the New University of Lisbon, contains 281,301 words, and it includes texts produced by adults and children, as well as a spoken subset. Following these previous projects, we believe that COPLE2 corpus will contribute to broaden this emerging

[1] http://www.clul.ul.pt/en/research-teams/547

[2] http://www.clul.ul.pt/pt/recursos/314-corpora-of-ple
[3] http://www.uc.pt/fluc/rcpl2/
[4] http://cal2.clunl.edu.pt/

Iria del Río, Sandra Antunes, Amália Mendes and Maarten Janssen 2016. Towards error annotation in a learner corpus of Portuguese. *Proceedings of the joint workshop on NLP for Computer Assisted Language Learning and NLP for Language Acquisition at SLTC 2016*. Linköping Electronic Conference Proceedings 130: 8–17.

domain by enhancing the learning data of Portuguese. COPLE2 makes use of a large set of learner texts (from different mother tongues (L1s) and proficiency levels) and, in contrast to the corpora mentioned above, it is linguistically interpreted with information on lemma and POS. Furthermore, it provides rich TEI annotation of the actual writing, the normalization of the orthography and error corrections, as well as a powerful multilayer query options.

We will first introduce the corpus and the interface tool in sections 2 and 3, respectively: section 2 presents the COPLE2 corpus, its design and the transcription process of written and spoken data, while section 3 gives an overview of the visualization and search options provided by the interface tool. In section 4, we introduce the error annotation system, the tagset and the discussion about the distribution of errors.

2 The COPLE2 corpus

COPLE2 corpus is constituted by written and spoken Portuguese learning data produced by students that attended Portuguese FL/L2 courses (annual or summer) at the School of Arts and Humanities of the University of Lisbon[5], and by applicants to accreditation exams, between 2010 and 2014.

2.1 Corpus Design and Metadata

The written subpart of COPLE2 currently contains 966 free essays, in a total of 156,691 tokens, produced by 424 students that represent 14 different L1s. We only selected L1s that had a minimum of 6 informants in our initial data set (cf. Table 1).

L1	Inf.	Texts	L1	Inf.	Texts
Chinese	129	323	Italian	20	34
English	65	142	Dutch	11	15
Spanish	52	139	Tetum	9	22
German	39	76	Polish	8	22
Russian	25	70	Arabic	8	13
Japanese	23	50	Korean	6	9
French	23	43	Romanian	6	8

Table 1: Informants and texts of the written subcorpus.

Given the heterogeneous nature of the informants, we registered detailed metadata regarding both the learner and the task profiles. Thus, concerning the learner's profile, we established a set of 8 required fields: name, age (18-40 years old), gender, mother tongue, nationality, proficiency based on the Common European Framework of Reference for Languages[6] (A1 (7%), A2 (40%), B1 (31%), B2 (19%), C1 (3%)), knowledge of other foreign languages and period of time studying Portuguese.

The text profile includes fields on: genre (argumentative (35,5%), narrative (17,5%), personal letter (12,5%), formal letter (10,5%), informative (9,6%), dialogue (6,4%), message/e-mail (6,3%), retell a story (1,5%) and literary critic (0,2%)), topic, task description (diagnostic test, mid-term or final test, homework, accreditation exam), timebound or not, with access to reference books or not, number of tokens and date.

Regarding the spoken subpart, the compilation of this subcorpus is still in progress. At the moment, 12 recordings are transcribed. The recordings consist on conversations between 2 or 3 learners of different proficiency levels moderated by the examiner, on topics such as: (i) presentation of the students; (ii) simulation of communicative situations; (iii) discussion of particular subjects, presenting arguments and opinions.

The metadata of the spoken task also encode information on the recording situation, such as: total time of the recording, total time of the segment that is transcribed and the location of the transcribed segment, acoustic quality, hidden or visible recording, involvement of the evaluator (dialogue, monologue or monologue with few interactions), spontaneous or planned, elicitation or non elicitation, social context (family, private, public, controlled environment) and channel (face to face, experimental, media, phone conversations, etc.).

Table 2 shows the current contents of the corpus per level and per modality.

[5] The corpus compilation is funded by Fundação para a Ciência e Tecnologia (UID/LIN/00214/2013), Fundação Calouste Gulbenkian (Proc. nr. 134655) and ADFLUL.

[6] Council of Europe (2001).

Level	Written		Spoken		Total	
	Texts	Tokens	Texts	Tokens	Texts	Tokens
A1	72	6,438	10	18,803	82	25,241
A2	382	49,761	0	0	382	49,761
B1	305	53,042	0	0	305	53,042
B2	181	39,665	1	3,010	182	42,675
C1	26	7,785	1	3,970	27	11,755
Total	**966**	**156,691**	**12**	**25,783**	**978**	**182,474**

Table 2: COPLE2 design.

2.2 Data Transcription

The hand-written essays were first scanned and saved in pdf format, and then manually transcribed. The transcriptions are encoded in TEI compliant XML (Burnard & Bauman, 2013). Each file is composed by a header (with the metadata mentioned above) and the transcription, as illustrated in Figure 1, below.

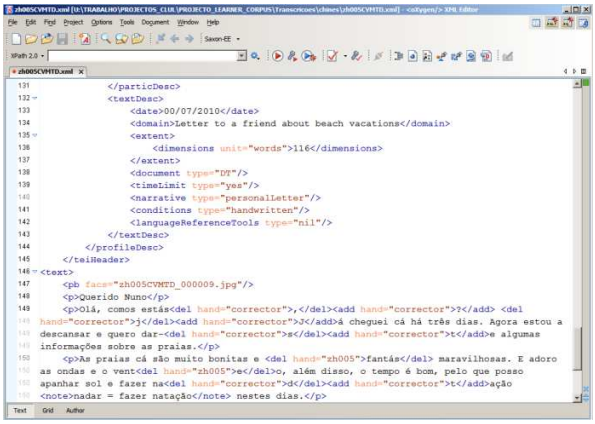

Figure 1: Part of a written transcription following XML.

The written transcriptions are very close to the original document in the sense that all the changes made by the student during the writing process (such as deletions, additions, transposition of segments, etc.) were also encoded. This information is extremely useful to assess, for instance, the difficult areas for the learning process according to the student's L1, the discourse restructuring or errors triggered by homophone words. In addition, all the corrections and comments made by the teacher were also transcribed. Teacher's feedback can be useful for future pedagogical studies and, as we will explain below, it constitutes a valuable support for error identification in the error annotation process. All personal information (such as names, addresses, phone numbers) were anonymized (Hinrichs, 2006).

Regarding the spoken corpus, the recordings were transcribed following CHILDES (MacWhinney, 2000) and C-ORAL-ROM (Cresti and Moneglia, 2005) guidelines, which favours a transcription based on prosody. Thus, instead of punctuation marks, we used symbols that represent intonation. Also, all the speech disfluencies (such as fragmented words, false starts, filled pauses and other non-lexical utterances) were transcribed. All the transcriptions were text-to-sound aligned using the EXMARaLDA editor (Schmidt, 2012).

3 TEITOK Interface Tool

After completion of the transcriptions, all the files were imported into the Tokenized TEI Environment – TEITOK[7] for visualization, linguistic annotation and search functions (Janssen, 2012; 2016). This system makes it easy to display XML files, edit metadata and individual tokens, and perform complex searches through the corpus.

The corpus was firstly automatically tokenized, which means that all lexical words and contracted words (such as prepositions contracted with articles, demonstratives, etc.) were identified (e.g. *naquele* = $em_{\text{preposition}}$ 'in' + $aquele_{\text{demonstrative}}$ 'that one'). The automatic POS annotation and lemmatization were performed, using the Neotag tagger (Janssen, 2012), which was trained over a gold standard subset of the Reference Corpus of Contemporary Portuguese (Mendes et al., 2014). For error tagging purposes, as we will see in the next section, a normalized version (orthographic, lexical or syntactic) may be provided also for each token. Because learner errors affect automatic POS tagging and lemmatization, default POS and lemma are normalized, that is, corrected when needed and stored at the first level of error annotation (orthographic). We will come back to this intersection of POS and error annotation in section 4.

Afterwards, for the written subcorpus, TEITOK interprets the XML encoding (CSS rules define how to display the XML elements) to enable the visualization of different versions of the text: (i) the XML version; (ii) the transcription version

[7] http://alfclul.clul.ul.pt/teitok/site/index.php?action=about

(visualization close to the full information of the XML document); (iii) the student form, which corresponds to the final version intended by the student; (iv) the corrected form, which displays the teacher corrections; (v) the error-annotated form; (vi) the image of the handwritten essay, on request. Each version has a specific separator, and all the changes made to the original student text are displayed in different colours. Figure 2 shows the teacher's correction version, where the corrected words are in red.

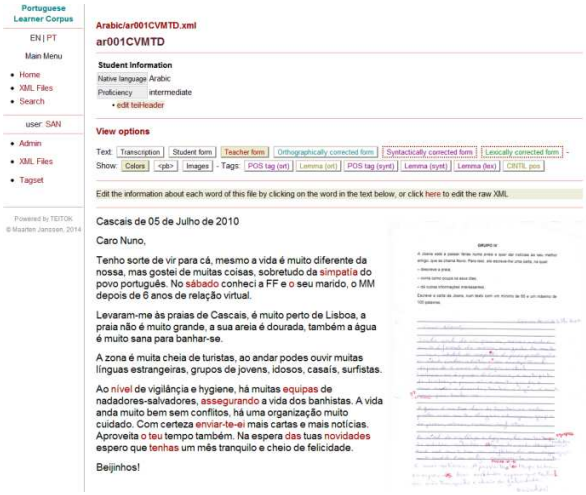

Figure 2: Visualization of the correction of a written essay.

All this information can be also displayed when moving the mouse over the words in the text. Figure 3 shows a misspelled word with the respective correction and all the linguistic information.

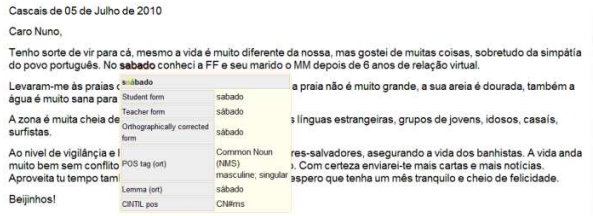

Figure 3: Highlighted word with linguistic information.

Regarding the spoken transcriptions, EXMARaLDA files were converted into TEI format. The spoken transcriptions are visualized as speech turns with a link to the audio sequence (cf. Figure 4).

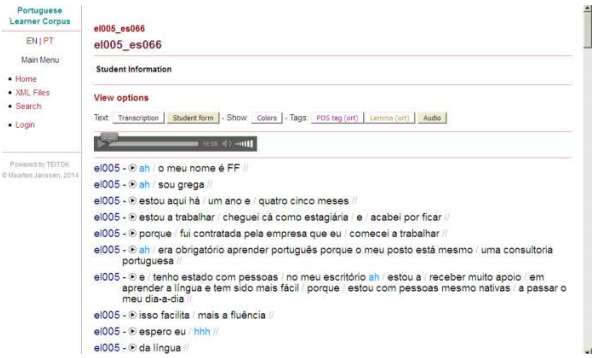

Figure 4: Visualization of a recording transcription.

TEITOK allows for multi-token annotation (POS, lemma, error-annotation) with the possibility of using regular expressions when specific replacements have to be made.

Finally, the TEITOK environment also provides corpus search facilities using CQP (Christ et al., 1999). In the creation of the CQP corpus, various types of encoded information can be exported: metadata, POS, lemma, original orthography, normalized orthography, error annotations and the teacher corrections. This way, searches can combine all these different types of information, making it possible to perform complex and powerful search queries (cf. Figure 5).

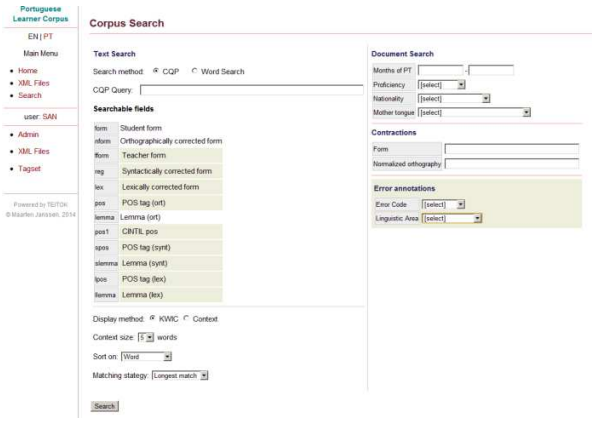

Figure 5: TEITOK query system.

The next step is to label the data following a typological scheme for error annotation (Tono, 2003; Nicholls, 2003; Dagneaux et al., 2005), as we describe in further detail below.

4 Error Annotation

Error tagging is an important step in learner corpora annotation since it helps to identify problematic areas in the learning process (Granger, 2004). Despite this fact, error tagging is not always present in learner corpora. There can be many possible reasons for that, but we can identify at least two important causes:

1. It is a high time-consuming task, that most of the times has to be performed manually.
2. There is no standard for error tagging and, in general, taxonomies are a result of particular projects with specific interests (Díaz-Negrillo & Fernández-Domíguez, 2006). As a consequence, an error taxonomy and an annotation paradigm have to be defined for each learner corpus and this is not a trivial task (Meurers, 2015), since it entails several complex sub-tasks like: define what an error is and what types of errors are considered; decide which is the scope of a given error (one word vs. multiple words); determine if corrections are provided or not; etc.

As we will show, in the case of COPLE2 we have tried to take advantage of the corpus architecture and the possibilities that the TEITOK environment offers to overtake the problems above.

There are examples of learner corpora with error annotation for many languages but, to the best of our knowledge, none of the learner corpora for Portuguese offers error annotation. Therefore, error tagging in COPLE2 constitutes the first attempt of this type of encoding for the Portuguese language.

4.1 Error annotation system in COPLE2

The error annotation paradigm in COPLE2 exploits the possibilities provided by the TEITOK environment. We have already described different levels of annotation that TEITOK allows for each token in the corpus (student form of the token *versus* teacher form of the token). For error tagging, we have defined three linguistic levels of annotation: orthographical, grammatical and lexical. In all the cases, the annotation consists on the addition of the correct word form with its lemma and POS. The three levels can be filled for a given token at the same time.

The first level is used if there is a spelling error in the student production. The orthographically corrected form (*nform*) is introduced, as well as the corresponding POS (*pos*) and lemma (*lemma*). Figure 6 below shows an example of an orthographical error, where the student wrote *novedades* instead of *novidades* ('news').

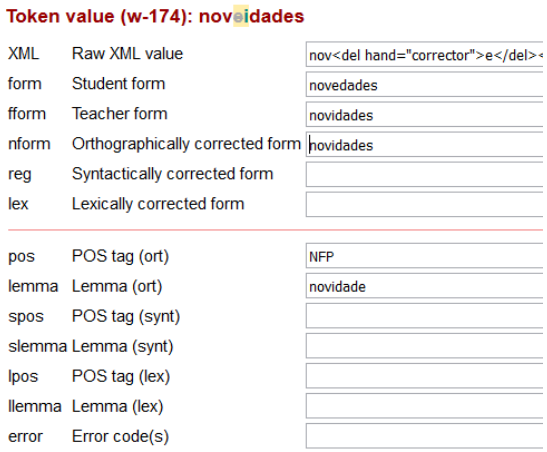

Figure 6: Annotation of an orthographic error.

As we have mentioned above, this first level contains the default POS and lemma for each token, which are corrected (normalized) when needed.

The second level operates if there is a grammatical error, that is: the word used by the student generates an ungrammatical utterance. Figure 7 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*).

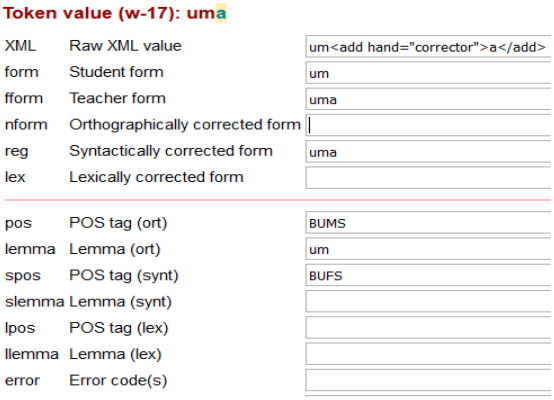

Figure 7: Annotation of a grammatical error.

Note that in this case the field *slemma* is not annotated. The reason is that there is inheritance between levels, from the bottom (orthographic data) to the top (lexical data), that is: *form* > *nform* > *reg*

> *lex*; *pos* > *spos* > *lpos*; *lemma* > *slemma* > *llemma*, and only what is new 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*). If *reg* is filled in and *lex* is empty, the value for the *lex* is the same as for *reg*; and the same for the POS and the lemma. In the example in Figure 7, the value for *slemma* is the same as the value in *lemma*, and therefore *slemma* is empty. 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 level.

Finally, the third level is used if there is a lexical/semantic error in the student form, i.e., the word can be grammatically correct, but it is not the natural word that a native speaker would use. Figure 8 shows an example where the student used the word *tropas* ('troops') in a context where *equipas* ('teams') was more adequate.

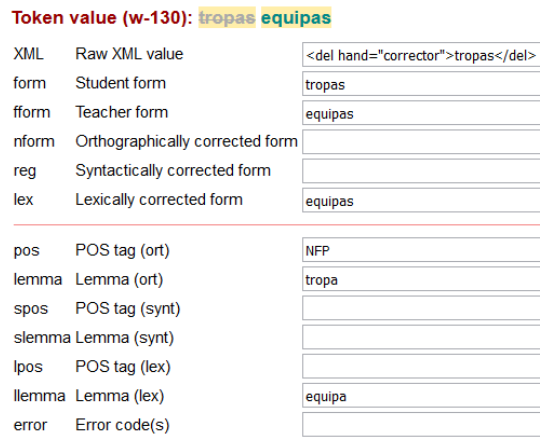

Figure 8: Annotation of a lexical error.

Again, in Figure 8, 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.

The different levels provide also different visualizations of the text, where the introduced corrections replace the student forms. This way, it is possible to visualize the same text corrected at different levels, 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-tier 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 levels 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 level 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 the native language.
- Grammatical level: the operations at this level 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 interpreted form allows for a grammatically correct structure in the learner production.
- Lexical level: the operations allowed at this level 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 8).

Because it works at the level of the token, this annotation system does not work for errors that affect more than one word, like word order errors or errors in multi-word expressions. For those cases, we will use stand-off annotation, which is already implemented in TEITOK (Janssen, 2016).

Currently, we are testing this annotation system, which seems intuitive and fast for the annotators. As part of the testing, we plan to perform inter-annotator agreement evaluation, to check the degree of confidence of the system. Considering the results showed by previous works like Rosen et al. (2013), we expect to find a relation between the annotator agreement and the level of interpretation allowed by the annotation tier (less interpretation at the orthographic and grammatical level, more interpretation at the lexical level). For the identification of errors, we plan to combine automatic and manual strategies, taking advantage of the information already encoded in the corpus, for example, teacher's corrections (always reviewed by a human annotator).

4.2 Distribution of errors: preliminary data

We do not have yet quantitative data about the total number of errors per type in the corpus but we have some indicative numbers from a pilot experiment we performed when we were designing a pilot taxonomy of errors. For this experiment we annotate 36 texts (7,073 tokens), trying to include all the languages in the corpus and, if possible, all the language levels. We found 591 errors (8.35% of total tokens), with the following distribution:

Type of error	Absolute Freq.	Percent Freq.
Orthographical	260	43.99
Grammatical	305	51.61
Lexical	26	4.4
Total	591	100

Table 3: Distribution of errors in a corpus sample.

As we expected, the most common errors are grammatical ones, followed by orthographic errors. This tendency was also showed for French in the FRIDA corpus in Granger 2003. On the other hand, lexical errors seem to be not very frequent, especially if the annotator is not very strict with the lexical choices of the learner.

4.3 Tagset of errors

As a further step, we plan to introduce error codes for each error annotated following the system described above. As we will see, the multi-tier error annotation will provide us automatically with the first level of information in the code, with a coarse-grained error annotation of the token.

We are working on the definition of the tagset that will be used, similar to the taxonomies described in Tono, (2003), Nicholls (2003) or Dagneaux et al. (2005). So far, we have defined a pilot tagset that will be applied to the corpus to test its performance. The current tagset has 37 tags and it is structured in two levels of information:

1 General linguistic area affected.

2 Error category (and subcategories in some cases).

For level 1 we consider the three linguistic areas that we have described above: Orthographic (includes spelling and punctuation errors), Grammatical (includes agreement errors; errors affecting verb tense, mode, etc.) and Lexical (includes lexical choice errors; errors affecting derivational suffixes; etc.). As we will show below, the use of the same general linguistic areas to classify the errors allows for transferring information between the multi-tier system and the code system. For level 2 we have common categories like agreement or wrong POS.

To design the tagset we performed the annotation experiment that we referred above, identifying the errors in those 36 texts and defining the necessary categories to annotate them. Besides the phenomena we observed in the annotated sample, we included also other phenomena that we expect to find in the corpus, considering other tagsets developed for similar projects. When defining the error categories, we decided to be as general as possible, trying to avoid restricting ourselves to specific theoretical frameworks or being too detailed. We think that it is always easier to manage general categories that can be sub-specified in later stages than to apply from the beginning very detailed linguistic categories. The tags we defined are position-based tags, where the first letter corresponds to level 1 and the subsequent letters to level 2. For example, for agreement errors affecting gender, we have the tag "GAG" which stands for "Grammar + Agreement + Gender". Since the error tag is added to the affected token/group of tokens in the xml, which include POS information, we do not include the POS information in the label.

We expect that the tagset will provide a fine-grained classification of errors, which in turn will allow for more specific queries concerning different linguistic phenomena (agreement, word order,

use of incorrect POS, etc.). When possible, we will use all the information encoded for each token in COPLE2 to assign the error code automatically, comparing the original form from the student with the corrections (plus lemma and POS) introduced at the error annotation level. The first letter of the error code will be automatically assigned, taking into account the level where the error was annotated in the multi-tier system (orthographic, grammatical or lexical). The subsequent letters corresponding to the error type will be assigned automatically when possible. For example: if there is an annotated form at *nform* (orthographic tier) that means that there is an orthographic error. This allows for classifying automatically the error at the linguistic level, that is, to assign the first letter of the tag (S, in this case). But we can go further in some cases and assign also the error code letter(s). For example, we have an error type for accentuation marks (also S at second position in the error tag). For this error type, we can compare the student form and the *nform* to check if the difference affects only accentuation marks and, in that case, assign the corresponding letters to the error code (SS). Of course, this automatic comparison cannot be performed for the most complex error types, but in many cases it will save a lot of annotation time. This is a good example of the possibilities that COPLE2 offer to apply Natural Language Processing techniques to the annotation process.

We think that the information encoded at the error level (the three tiers described plus error codes) together with all the information already encoded in the corpus (metadata, student's modifications, teacher's corrections) will allow for complex and rich linguistic queries in COPLE2. Our aim is to encode and provide as much information as possible about different aspects of the learner corpus:

- Writing process of the learner.
- Corrections made by the teacher.
- Error corrections with POS at lemma at different tiers plus error tags.
- Metadata (type of text; age; language level; etc.).

We expect that this information can be useful for researchers of different fields: General Linguistics, Language Acquisition, Foreign Language Teaching and Learning, Computer Assisted Language Learning, etc.

5 Final Remarks

COPLE2 corpus is a new learner corpus for Portuguese that encompasses written and spoken data, with a rich XML encoding. For each text included in the corpus, it contains complete metadata (information about the author and the circumstances where the text was produced) and linguistic annotation concerning POS, lemma and modifications/corrections done by the student and the teacher in the original text. Besides this, it will offer soon error-annotation, being the first learner corpus of Portuguese with this type of encoding. Error tagging is an added-value in learner corpora, since it provides valuable quantitative (error statistics) and qualitative (type of error) data that highlight the learners' difficulties. TEITOK's architecture (where each token contains all the linguistic information, following TEI) facilitates the error annotation process. Furthermore, using the CQP search functionality, error tagging information could be combined with the other linguistic features encoded in the corpus, allowing for complex and rich linguistic searches in learner texts. By combining search queries, we can easily conduct studies based on Contrastive Interlanguage Analysis (Granger, 1996, 2015), which allow for uncovering distinctive features of specific L1 learners, as well as general errors across the learner population. Finally, COPLE2 will provide different visualizations of the learner text: text produced by the student; version orthographically corrected; version grammatically corrected and version lexically corrected.

The TEITOK environment provides POS and lemma automatic annotation, along with a full set of functionalities for manual linguistic annotation, as well as visualization and powerful search options. Since it is a highly customizable tool, with a wide range of user-defined annotations, it has proven a valuable resource for corpus analysis.

We believe that this corpus and tool constitute good resources for pedagogical foreign language learning/teaching analysis, since it provides empirical data to: (i) identify general and specific errors in the learning of Portuguese L2; (ii) develop automatic tools for language learning, textbooks and other material targeting specific groups of students; (iii) implement teacher training materials; (iv) illustrate the writing-speech interaction, which has

not been the subject of much analysis and has been insufficiently evaluated.

References

Boyd, A., J. Hana, L. Nicolas, D. Meurers, K. Wisniewski, A. Abel, K. Schöne, B. Štindlová and C. Vettori. 2014. The MERLIN corpus: Learner Language and the CEFR. In *Proceedings of LREC*, Reykjavik, Iceland. pp.1281--1288.

Burnard, L. and S. Bauman. Eds. 2013. *TEI P5: Guidelines for Electronic Text Encoding and Interchange*. Text Encoding Initiative Consortium: Charlottesville, Virginia.

Council of Europe. 2001. *Common European framework of reference for languages: Learning, teaching, assessment*. Cambridge, U.K: Press Syndicate of the University of Cambridge.

Cresti, E. and M. Moneglia. Eds. 2005. *C-ORAL-ROM. Integrated Reference Corpora for Spoken Romance Languages*. Amsterdam/Philadelphia: John Benjamins Publishing Company.

Christ, O., B. Schulze, A. Hofmann and E. Koenig. 1999. *The IMS Corpus Workbench: Corpus Query Processor (CQP): User's Manual*. Institute for Natural Language Processing. University of Stuttgart. (CQP V2.2).

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

Delais-Roussarie E. and H. Yoo. 2010. The COREIL corpus: a learner corpus designed for studying phrasal phonology and intonation. In K. Dziubalska-Kołaczyk, M. Wrembel and M. Kul (Eds). *Proceedings of New Sound* 2010. Poznan, Pologne, pp. 100--105.

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

Granger, S. 1996. From CA to CIA and back: An integrated approach to computerized bilingual and learner corpora. In K. Aijmer, B. Altenberg and M. Johansson (Eds.). *Languages in Contrast. Text-based cross-linguistic studies.* Lund Studies in English 88. Lund: Lund University Press, pp. 37--51.

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. 2004. Computer learner corpus research: current status and future prospects. In U. Connor & T. Upton (Eds.), Applied Corpus Linguistics: A Multidimensional Perspective (pp. 123-145). Amsterdam & Atlanta: Rodopi.

Granger, S. 2015. Contrastive Interlanguage Analysis: a reappraisal. *International Journal of Learner Corpus Research*. Vol. 1:1. John Benjamins Publishing Company, pp. 7--24.

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

Hinrichs, L. 2006. *Codeswitching on theWeb. English and Jamaican Creole in e-mail communication*. Amsterdam/Philadelphia: John Benjamins Publishing Company.

Janssen, M. 2012. NeoTag: a POS Tagger for Grammatical Neologism Detection. In *Proceedings of LREC 2012*, Istanbul, Turkey.

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

Leiria, I. 2001. *Léxico – aquisição e ensino do Português Europeu língua não materna*. PhD Dissertation. Faculdade de Letras da Universidade de Lisboa.

Lozano, C. 2009. CEDEL2: Corpus Escrito del Español L2. In C. M. Bretones Callejas et al. (Eds). *Applied Linguistics Now: Understanding Language and Mind / La Lingüística Aplicada Hoy: Comprendiendo el Lenguaje y la Mente*. Almería: Universidad de Almería, pp. 197--212.

MacWhinney, B. 2000. *The CHILDES Project: Tools for Analyzing Talk. 3rd Edition*. Mahwah, NJ: Lawrence Erlbaum Associates.

Mendes, A., M. Généreux, I. Hendricks. 2014. *Manual for the CRPC on the CQPweb interface*. Manual 1.3. http://alfclul.clul.ul.pt/CQPweb/doc/CRPCmanual.v1_2_en.pdf.

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

Meurers, D. 2015. Learner Corpora and Natural Language Processing. In S. Granger, G. Gilquin and 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 and 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.

Schmidt, T. 2012. EXMARaLDA and the FOLK tools – two toolsets for transcribing and annotating spoken

language. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC)*. Istanbul, Turkey, pp. 236--40.

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