Error annotation in a Learner Corpus of Portuguese

Iria del Río & Amália Mendes
University of Lisbon, Center of Linguistics - CLUL
igayo@letras.ulisboa.pt, amaliamendes@letras.ulisboa.pt

Abstract

We present the error tagging system of the COPLE2 corpus and the first results of its implementation.. The system takes advantage of the corpus architecture and the possibilities of the TEITOK environment to reduce manual effort and produce a final standoff, multi-level annotation with position-based tags that account for the main error types observed in the corpus. The first step of the tagging process involves the manual annotation of errors at the token level. We have already annotated 47% of the corpus using this approach. In a further step, the token-based annotations will be automatically transformed (fully or partially) in position-based error tags. COPLE2 is the first Portuguese learner corpus with error annotation. We expect that this work will 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.

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, 2004) and provides useful data for many areas of study (Díaz-Negrillo and 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 and 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 process is still a challenge.

We present the error annotation system for the COPLE2 corpus 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. 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; section 3 presents the COPLE2 corpus and its error annotation system; in section 4 we show our first annotation results; finally, section 5 presents the conclusions and future challenges.

2. Related work

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 and 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[1], 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) points 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, 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, and that in

[1] For a detailed review see Díaz-Negrillo and Fernández-Domíguez (2006).

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 and 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.

3. Error annotation in the COPLE2 corpus

3.1 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 (Janssen, 2012). All the information is stored together with the original texts in XML files that can be searched through the CQP query language (Christ et al., 1999).

3.2 Error annotation in the COPLE2 corpus

For error annotation in COPLE2 (del Río et al., 2016) we take advantage of the corpus architecture and the information already annotated, as well as of 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 fine-grained, standoff, multi-level system. The token-based system makes possible a quick and simple annotation, supports the visualization of the corrected text and complex queries using CQP. But, what is more important: it allows for the automatic generation of the fine-grained annotation system 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.

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). For example, given the input: *um cidade* ('a_{MASC} $city_{FEM}$') instead of *uma cidade* ('a_{FEM} $city_{FEM}$), where what we have is an agreement problem between a determiner and a noun, the annotator introduces the correct word form for the determiner (*uma* instead of *um*) and the correct POS, but the lemma remains the same (*um*). 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). All the error 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). However, taking into account what we discussed in section 1, it is clear that this system presents some problems for error annotation: it only works at the token level[2]; it offers a limited categorization and description of errors types; and it is limited to three linguistic areas, while some errors go beyond those areas.

Due to these limitations, the token-level annotation is complemented with a fine-grained, standoff, multi-level 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. The tagset designed for this system is similar to the taxonomies described in Tono (2003),

[2]Although, as we will see in next section, we have found that only a small percentage of the total errors identified so far in COPLE2 did not fit at all in a token based interpretation.

Nicholls (2003) or Dagneaux et al. (2005). It contains 38 tags and it is structured in two levels of information: (i) general linguistic area affected; (ii) error category (and subcategories in some cases). Level 1 includes (for the moment) the same three linguistic areas that 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: the first letter corresponds to level 1 and the subsequent letters to level 2. For example, for agreement errors affecting gender, the tag is "GAG" which stands for "Grammar + Agreement + Gender".

Most of the tags and their corresponding corrections can be automatically generated (at least partially) comparing the original form of the student with the corrections (plus lemma and POS) introduced at the token level. The first letter of the tag can be always generated just checking the linguistic level where the corrections were added. The subsequent letters of the tag can be inferred using the linguistic information annotated in the corpus. For example, we have an error tag for accentuation marks (SS). For this error type, we can compare the student form and the *orthographically corrected form* to check if the difference affects only accentuation marks and, in that case, assign the corresponding letters to the error tag (SS). With this strategy, we take advantage of the TEITOK and COPLE2 possibilities to automatically produce a detailed error annotation with low manual effort.

4. Results of error annotation at the token level

We have started the error annotation at the token level.[3] 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 go 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:

(1) *As praias são muito **lindos**,* […] > *lindas*[4] ('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 level, affecting a grammatical structure (the sentence, in this example). 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:

(2) *Foi uma expêriencia que eu nunca **tenho esquecido*** > *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 tagset from the token-based annotations. We will do it through conversion scripts that take as input all the 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:

(3) *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. One example is the tag LN (Lexical+Nonexistent_Word), for the cases where the student created a new word (that does not exist in Portuguese) using recognizable morphological processes, as in:

(4) *e **estabilitamos** a melhor relação > estabelecemos* ('and we **establish** the best relation').

In this example the student created a new verb *estabilitar*, probably from the adjective *estável* (stable), instead of using *estabelecer* (to establish).

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-level 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

[3] For the moment, only one annotator is performing the task. In the future, we would like to count with at least two different annotators.

[4] All the examples are from COPLE2 and have the following format: the error is marked in bold, the correction is shown after the ">" symbol and a translation in English (with the corresponding correction) is provided.

structures that correspond to a certain tag, automatizing the generation. A second line of work is related to the addition of new linguistic levels 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.

6. 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).

7. Bibliographical 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).

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

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

del Río, I., Antunes, S., Mendes, A. and 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. and 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* (pp. 249–264). Amsterdam & Philadelphia: John Benjamins.

Díaz-Negrillo, A. and Fernández-Domíguez, J. (2006). Error Tagging Systems for Learner Corpora. *RESLA*, 19:83--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* (pp. 123-145). Amsterdam & Atlanta: Rodopi.

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.

Kutuzov, A. and 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. and Adolphs, P. (2005). Multi-level annotation error annotation in a learner corpora. In *Proceedings of Corpus Linguistics* 2005 1, Birmingham (England), July 2005, 14-17. Rosen, A., Hana, J., Štindlová, B. and Feldman, A. (2013). Evaluating and automating the annotation of a learner corpus. *Language Resources and Evaluation*, pp. 1--28.

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.

8. Language Resource References

Boyd, A., Hana, J., Nicolas, L., Meurers, D., Wisniewski, K., Abel, A., Schöne, K., Štindlová, B. and 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.

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.

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

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

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.