Error annotation in COPLE2

Iria del Río

Workshop on Interoperability of Second Language Resources and Tools
University of Gothenburg
6–8 December, 2017

Introduction

- COPLE2 is a corpus of Portuguese FL/L2 that encompasses written and spoken data produced by foreign learners of Portuguese.

- The learning data was collected in evaluation tests or accreditation exams between 2010-2014, at FLUL. We continue to collect and include new data.

- COPLE2 constitutes a good resource for teachers and researchers, since it provides empirical data to:

  (i) identify general errors in the learning of Portuguese L2;

  (ii) develop textbooks and other material targeting specific groups of students;

  (iii) implement teacher training materials by taking into account the analysis of the errors;

  (iv) illustrate the writing-speech interaction.

Corpus Design

Level	**Written**	**Spoken**	**Total**
A1	82	10	92
A2	413	0	413
B1	312	0	312
B2	201	1	202
C1	38	1	39
Total	**1,046**	**12**	**1,058**

Corpus Design
Written Subcorpus

The **written subcorpus** is composed by:

- **1,046 free essays** produced by **482 students.**

- **15 different L1s** (number of students in brackets): Chinese (129), English (65), Spanish (58), German (46), Russian (30), French (33), Japanese (24), Italian (32), Dutch (13), Tetum (9), Arabic (8), Polish (13), Korean (6), Romanian (9) and Swedish (7).

- **Different proficiency levels:** A1 (9%), A2 (39%), B1 (30%), B2 (19%), C1 (3%).

- **Different genres and topics**: 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%), literary critic (0,2%).

Corpus Design
Metadata

- Regarding the learner:

- Age (18-40 years old).
- Nationality (relevant for languages that are spoken in different countries).
- Knowledge of other foreign languages.
- Period of time studying Portuguese.

- Regarding the task:

- Task description (diagnostic, mid-term or final test; accreditation exam).
- Time limit for writing or untimed.
- Access to reference tools (dictionaries, grammars, notes, etc.).

Transcription

Written Subcorpus

- All the essays were scanned and manually transcribed.
- The transcriptions are encoded in XML, following the TEI guidelines, and anonymized.
- Each file is composed by a header (with the metadata), and the transcription, which includes:

  (i) all the changes made by the student (deletions, additions, etc.);

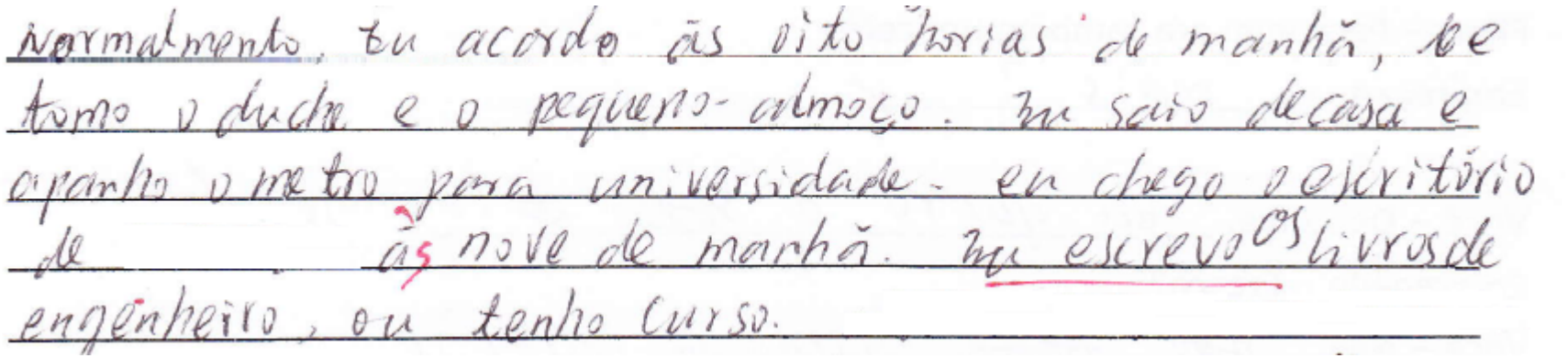
  (ii) the correction and comments made by the teacher.

Transcription
Written Subcorpus

Excerpt from the XML file

<p>Normalmento, Eu acordo às oito horias de manhã, <del hand="zh010">t</del> e tomo o duche e o pequeno-almoço. Eu saio de casa e apanho o metro para universidade, eu chego o escritório de XX <del hand="corrector">á</del><add hand="corrector">às</add> nove de manhã. <hi hand="corrector" rend="underlined">Eu escrevo <add hand="zh010">os</add></hi> livros de engenheiro, ou tenho curso.

Excerpt from the manuscript



TEITOK interface tool

- The XML files were imported to the TEITOK (Tokenized TEI Environment) platform for visualization, linguistic annotation and search functions.
- The corpus was firstly automatically tokenized.
- Automatic POS annotation and lemmatization were performed (Neotag).
- For the **written subcorpus**, TEITOK interprets the XML encoding to enable:

(i) the visualization of different versions of the texts: XML, transcription (faithful to the handwritten document), final version intended by the student, the correction of the teacher;

(ii) the linguistic annotation: lemmatization; PoS, orthographic normalization, error codification;

(iii) the image of the handwritten essay, on request;

(iv) corpus search (for word, lemma, POS, error code, metadata, etc.).

TEITOK: written subcorpus

Arabic/ar001CVMTD.xml

ar001CVMTD

Student Information

Native language	Arabic
Proficiency	intermediate

• edit teiHeader

View options

Text: Transcription | Student form | Teacher form | Orthographically corrected form | Syntactically corrected form | Lexically corrected form -

Show: Colors | <pb> | Images - Tags: POS tag (ort) | Lemma (ort) | POS tag (synt) | Lemma (synt) | Lemma (lex) | CINTIL pos

Edit the information about each word of this file by clicking on the word in the text below, or click here to edit the raw XML

Cascais de 05 de Julho de 2010

Caro Nuno,

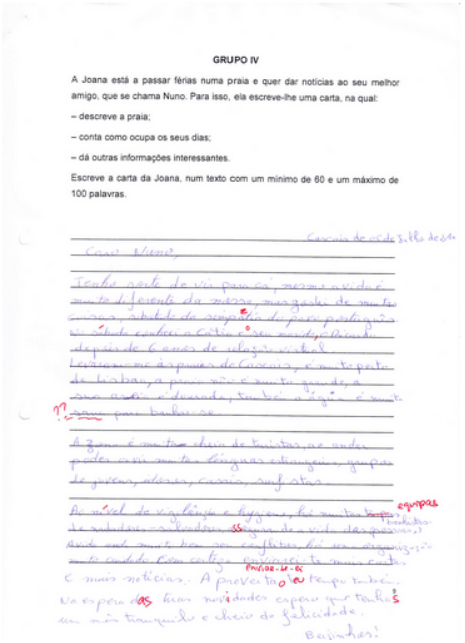
Tenho sorte de vir para cá, mesmo a vida é muito diferente da nossa, mas gostei de muitas coisas, sobretudo da simpatía do povo português. No sábado conheci a FF e o seu marido, o MM depois de 6 anos de relação virtual.

Levaram-me às praias de Cascais, é muito perto de Lisboa, a praia não é muito grande, a sua areia é dourada, também a água é muito sana para banhar-se.

A zona é muita cheia de turistas, ao andar podes ouvir muitas línguas estrangeiras, grupos de jovens, idosos, casaís, surfistas.

Ao nível de vigilância e hygiene, há muitas equipas de nadadores-salvadores, assegurando a vida dos banhistas. A vida anda muito bem sem conflitos, há uma organização muito cuidado. Com certeza enviar-te-ei mais cartas e mais notícias. Aproveita o teu tempo também. Na espera das tuas novidades espero que tenhas um mês tranquilo e cheio de felicidade.

Beijinhos!



GRUPO IV

A Joana está a passar férias numa praia e quer dar notícias ao seu melhor amigo, que se chama Nuno. Para isso, ela escreve-lhe uma carta, na qual:

– descreve a praia;

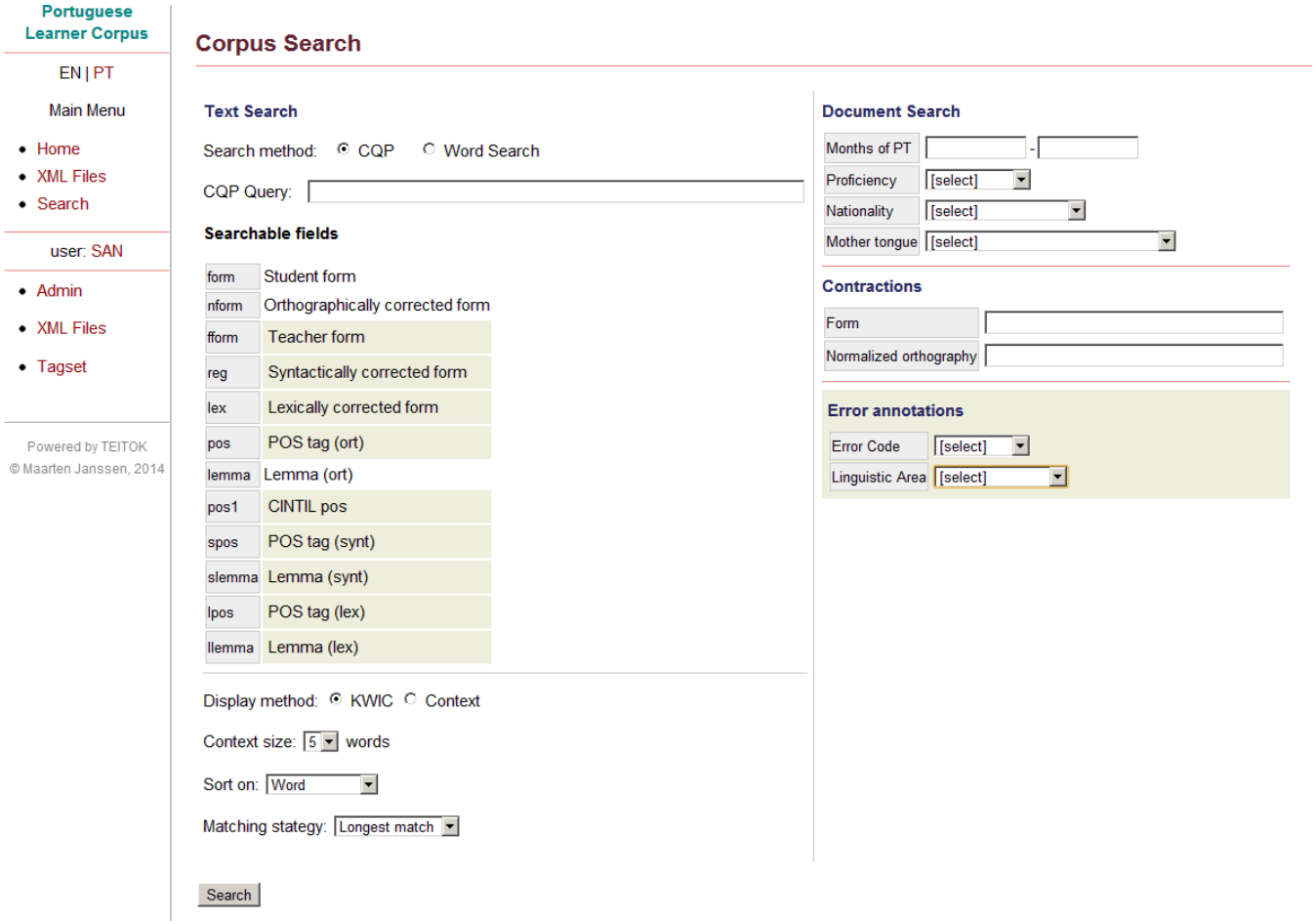
– conta como ocupa os seus dias;

– dá outras informações interessantes.

Escreve a carta da Joana, num texto com um mínimo de 60 e um máximo de 100 palavras.

TEITOK: query system

- Corpus search uses CQP, which allows to combine different types of information, making it possible to perform complex and powerful search queries.



Error annotation

- Error tagging is an important step in learner corpora annotation since it helps to identify problematic areas in the learning process and provides useful data for many areas of study (Díaz-Negrillo and Thompson, 2013).

- However, it is **not** present in all learner corpora because:

  - It is a high time-consuming task performed manually.

  - There is no standard for it, schemas are created for particular projects.

Error annotation

Development of error tagging systems through the years

- **Technically**: from in-line, flat systems like in the *Cambridge Learner Corpus*:

[…] *lawyers, doctors, etc, <#UA>they</#UA> hardly earn #50,000 a year.* (CLC; Nichols, 2003: 576).

to standoff, multi-layer architectures like in the FALKO corpus.

word	dass	nur	er	…	konnte	durch	dieses	Tor	eingelassen	werden	
target					durch dieses Tor eingelassen werden konnte						
word order identification					x						
word order description					MF_RSK						
word order explanation					transfer						

Error annotation

Development of error tagging systems through the years

- **Conceptually:** less variation. Most of the systems focus on the same linguistic areas:

	CLC	FreeText	Louvain	*NICT JLE*
Phonology	–	–	–	–
Punctuation	√	√	√	–
Spelling	√	√	√	–
Grammar	√	√	√	√
Lexis	√	√	√	√
Pragmatics	√ (register)	√ (register)	√ (register)	√ (non-verbal cues)
Discourse	√ (pronoun reference)	√ (cohesion)	√ (unnatural discourse)	√ (self-corrections)

From Díaz-Negrillo & Fernández Domínguez, 2006

Error annotation

Basic questions for error tagging:

1 **What is an error and what is not? - GUIDELINES!**

- Error tagging implies a high level of interpretation.
- It is really important to define clearly what we consider an error.
- Important to differentiate between error tagging and generating a "corrected" version of a text.

Error annotation

Basic questions for error tagging:

1 **What is an error and what is not? - GUIDELINES!**

- Pilot annotation experiment showed a high degree of disagreement on the identification of errors.

Total errors	Agreement	Disagreement
484	303	181

  - Disagreements concerning punctuation constituted the 44% of the total.
  - Cause: annotator B tended to penalize more the use of commas than annotator A.

Error annotation

Excerpt from our annotation guidelines:

“Always try to minimize changes in the student’s text. We try to find a compromise between what the student wrote and what is correct in Portuguese.

For example, in: *O filme **é** tres horas. (The movie **goes** three hours)*

We prefer this correction: *O filme é de tres horas.* (*The movie goes on for three hours).*

Rather than: *O filme **dura** tres horas. (The movie lasts three hours).*

Because this way we show how to use correctly the student’s choice instead of showing how to use correctly a different verb (and structure).”

Error annotation

Basic questions for error tagging:

2 **Do we classify considering the possible cause of the error or the linguistic area affected by the error (or both)? – GUIDELINES!**

Some types of errors are easy to classify (agreement errors), but others are not. The same linguistic reality can be interpreted in different ways.

mesos instead of *meses* (months)

We have an orthographical error, because the word does not exist in Portuguese. However, we have also a problem in the inflectional suffix of a noun: the student used a productive masculine plural suffix in Portuguese ("os" like in "livro" 'book' > "livros" 'books'). So it is the possibility that the student simply wrote wrongly a word or that he used the wrongly suffix.

Error annotation

Basic questions for error tagging:

3 **Which should be the scope of an error?**

Take agreement errors: should we annotate only the word that exhibits the error or the whole structure in agreement?

bom apresentação 'good$_{\textbf{masc}\text{-sing}}$ presentation$_{\textbf{fem}\text{-sing}}$'

Probably the problem here is that the student interprets "apresentação" as masculine (it ends in "o") and therefore uses the masculine adjective. But the error is only visualized in the adjective and no in the noun.

Error annotation in COPLE2

- COPLE2 is, to our knowledge, the first learner corpora for Portuguese that offers error annotation.

- Our error annotation system takes advantage of the corpus architecture and the possibilities that the TEITOK environment offers.

Error annotation in COPLE2

Two-step system:

1. Manual normalization of errors at three linguistic levels: Spelling, Grammar, Lexis.

    - In-line, flat and token-based annotations inside the XML file.
    - Quick annotation system (reduced number of decisions for the annotator).
    - In progress: 68% of the corpus has been normalized.

2. Automatic annotation using fine-grained tags generated from all the information encoded in the corpus.

Manual normalization

- Three linguistic levels of annotation: orthographical, grammatical and lexical. The three levels can be filled for a given token at the same time.
- The annotation consists on the addition of the correct word form with its lemma and POS.
- The different levels provide different visualizations of the text, where the introduced corrections replace the student forms.
- It is a multi-layer system with different levels of annotation that work bottom-up allowing for different representations of the learner form.
- It assumes up to three target hypothesis where the reference linguistic system is the target native language.

Manual normalization:

orthographical level

Token value (w-174): noveidades

XML	Raw XML value	nov<del hand="corrector">e</del><
form	Student form	novedades
fform	Teacher form	novidades
nform	Orthographically corrected form	novidades
reg	Syntactically corrected form	
lex	Lexically corrected form	
pos	POS tag (ort)	NFP
lemma	Lemma (ort)	novidade
spos	POS tag (synt)	
slemma	Lemma (synt)	
lpos	POS tag (lex)	
llemma	Lemma (lex)	
error	Error code(s)	

Manual normalization:
grammatical level

Token value (w-17): uma

XML	Raw XML value	um<add hand="corrector">a</add>
form	Student form	um
fform	Teacher form	uma
nform	Orthographically corrected form	
reg	Syntactically corrected form	uma
lex	Lexically corrected form	
pos	POS tag (ort)	BUMS
lemma	Lemma (ort)	um
spos	POS tag (synt)	BUFS
slemma	Lemma (synt)	
lpos	POS tag (lex)	
llemma	Lemma (lex)	
error	Error code(s)	

Manual normalization:
lexical level

Token value (w-130): ~~tropas~~ **equipas**

XML	Raw XML value	<del hand="corrector">tropas</del>
form	Student form	tropas
fform	Teacher form	equipas
nform	Orthographically corrected form	
reg	Syntactically corrected form	
lex	Lexically corrected form	equipas
pos	POS tag (ort)	NFP
lemma	Lemma (ort)	tropa
spos	POS tag (synt)	
slemma	Lemma (synt)	
lpos	POS tag (lex)	
llemma	Lemma (lex)	equipa
error	Error code(s)	

Distribution of errors: preliminary results

- **723 texts and 118,743 tokens.**

- **14,176 errors** (11.9% of total tokens), with the following distribution:

  - 6,465 orthographical (45.6%);
  - 6,036 grammatical (42.6.%);
  - 1,675 lexical (11.8%).

Manual normalization

Limitations of the token-based normalization system:

- There are errors that go beyond the token.
- There are errors that operate at higher linguistic levels (semantics, pragmatics).
- The classification is too general: it is difficult to make investigation on particular linguistic issues with only three categories.

COPLE2 tagset for error annotation

Two levels of information:

- **Linguistic area**: Spelling, Grammar and Lexis.

  - Areas present in most of the tagsets and strongly linked to second language learning.
  - Relatively easy to identify from the annotation point of view.

- **Error category** (and subcategory in some cases): agreement, verbal tense, word order, etc.

  - Common categories in error tagsets.
  - General categories not restricted to specific theoretical frameworks that can be sub-specified/merged/deleted in later stages of annotation.

COPLE2 tagset for error annotation

The tagset follows the principles stated in Granger (2003b):

- **Consistent**: IAA experiments with a general value of κ = 0.84.
- **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.
- **Flexible**: hierarchical categories, structured in two levels. Easily adjustable.
- **Reusable**: it accounts for general categories that describe common errors in three linguistic areas. It can be adaptable to close languages like Spanish.

COPLE2 tagset for error annotation

- Position-based tags:
  - First position= linguistic level: Spelling-S; Grammar-G; Lexis-L.
  - Subsequent positions= type and subtype of error: agreement, lexical choice, etc.

Spelling + Stress Mark = SS; Grammar + Agreement + Gender = GAG;

Lexical + Lexical Choice = LC

- 38 tags (so far).
  - Spelling = 11 tags; Grammar = 25 tags; Lexis = 2 tags.
  - It is possible to increase the number of tags if required (the current process of error normalization indicates that new linguistics levels may be necessary).
- Standoff xml annotation + annotation of multiple tokens + annotation of overlapping spans.

COPLE2 tagset for error annotation

LINGUISTIC CATEGORIES	POSITION BASED	DESCRIPTION OF THE TAG	ERROR EXAMPLES
S_StressMark	S + S	It is used when there is an extra or a missing accent mark.	*diferentes* ***paises*** *e povos*
S_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***
S_Grapheme_Deletion	S+G+D	This tag includes all errors concerning the choice of the correct grapheme, and it is used if: a) 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***
S_Grapheme__Substitution	S+G+S	c) a grapheme is wrongly used instead of another grapheme (again, this does not apply to inflectional/derivational suffixes).	***spportei***
S_Grapheme_Transposition	S+G+T	Two graphemes have exchanged positions.	***apíses***
S_Capitalization	S + C	The word is written in lower case and should be capitalized or the opposite.	*a liberdade de que fala* ***pessoa***
S_WordBoundarySplit	S+B+S	One word incorrectly split.	***última mente*** *não falamos*
S_WordBoundaryMerged	S + B + M	Two or more words wrongly merged.	***fimdesemana***
S_PunctConfused	S + P + C	A wrong punctuation mark is used.	*a quem lhe ouvir* ***.*** *por exemplo*
S_PunctRedundant	S + P + R	Punctuation mark is not necessary.	*Assim* ***,*** *foi a minha modesta leitura*
S_PunctMissed	S + P + M	A punctuation mark is missed.	*é tal grande* ***que*** *às vezes*

COPLE2 tagset for error annotation

LINGUISTIC CATEGORIES	P. BASED	DESCRIPTION OF THE TAG	ERROR EXAMPLES
G_UnnecessaryWord	G + U	The written word is unnecessary.	*eu vou **a** organizar uma festa*
G_OmittedWord	G + E	Omission of a necessary word.	*[a] fala do dia a dia do cidadão*
G_WrongWord	G + W	Cases where the lemma (not the POS) selected by the learner is not correct, according to the grammatical surrounding context.	*numa altura em que ninguém sem tempo **por** nada*
G_WrongCategory	G + C	Wrong POS.	*não vive nas **selvagens** com tantos riscos*
G_WrongStructure	G + S	Error affecting more than one word (not a MWE) that can't be classified using the other subcategories.	*A industria de seviços (**hotel, restaurante, transporte**, etc)*
G_Agreement_Gender	G + A + G	Agreement error afecting gender.	*os ideáis humanitarias*
G_Agreement_Number	G + A + N	Agreement error afecting number.	*tem paisagens **lindíssima***
G_Agreement_Gender&Number	G + A + B	Agreement affecting gender and number.	*pode ser palavras **bo***
G_Agreement_Person	G + A + P	Agreement error affecting person.	*os portugueses não **podia***
G_WordOrder	G + O	The error affects the order of constituents.	*não lêem **livros muitos***
G_Verb_Tense	G + F + T	Incorrect tense.	*Sempre **havia** e sempre havrá*
G_Verb_Mode	G + F + M	Incorrect mode.	*uma battaria nova **deva** durar*
G_Verb_Tense&Mode	G + F + Z	Incorrect tense and mode.	*senhor prometeu-me que **irão** funcionar*
G_Verb_FiniteNoFinite	G + F + F	Confusion between finite and no finite.	*cada vez mais **melhorar***
G_Verb_Aspect	G + F + A	Incorrect aspect (use of imperfect instead of perfect simple, other cases?).	*o quarto **estivo** muito frio*
G_VerbalConstruction_Voice	G + F + V	For confusions between the active and the passive voice, this tag is used.	*?*
G_VerbalConstruction_Periphrasis	G + V + H	Error in periphrasis.	*E espero que não **va acontecer***
G_VerbalConstruction_ComplexForm	G + V + X	Error building a complex verb form: wrong auxiliary, use of gerund instead of participle, etc.	*?*
G_VerbalConstruction_Clitization	G + V + K	Error in clitized forms.	***descobrimos**-as*
G_PronounClitic_Case	G + F + C	Error in case (pronoun).	*visitou-**lhe** ontem*
G_PronounClitic_Person	G + F + P	Erron in person (pronoun).	*Liguei ás meninas e disse-**lhe** que vieram*
G_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*
G_SuffixDerivation	G+ D	Wrong use of derivational suffixes.	***a legalizacion** desse assunto*
G_SuffixInflection	G + I	Wrong use of inflectional suffixes.	*e uma familia **aberte** às outras pessoas*

COPLE2 tagset for error annotation

LINGUISTIC CATEGORIES	POSITION BASED	DESCRIPTION OF THE TAG	ERROR EXAMPLES
L_LexicalChoice	L + C	Used word exists in the language and the POS is correct, but the lemma is not right since it is r	*Se não* ***tiver*** *medidas de proteção*
L_UnexistentWord	L + U	A word that doesn't exist in Portuguese but is not a spelling error. This error can be a word from the L1, a word that is created mixing the L1 and Portuguese... etc, but it is not a typo where it is possible to identify easily the intended word.	*e* ***estabilitamos*** *a melhor relação*

Automatic generation of tags

- We take advantage of the corpus architecture and the possibilities of the TEITOK environment. Tags will be automatically generated (at least partially) using the information encoded in the corpus: original student form; POS; normalization of errors.

- Inference process for the automatic generation of a tag:

- Student form = *um* + Grammatically normalized form = *uma* > problem at the grammatical level > **First letter of the tag = G.**

- POS student form = *BUMS* + POS grammatically normalized form = *BUFS* > problem in the gender of the determiner > agreement problem affecting gender > **second and third letter of the tag = AG.**

Final tag = **GAG**

Error annotation in COPLE2: next steps

- Decide the scope of some errors (for example, agreement errors).
- Generate the fine-grained tags semi-automatically.
- Include syntactic and discourse annotation.
- Used the annotated data to develop automatic strategies to identify and (ideally) classify some types of errors.
- Use the annotated data in NLP tasks and tools.

Muito obrigadɵa!

ŧThanks a lot!

Tack saå mycket!