7th. Eicefala

A labeled dataset for analysing deviant orthographic forms in texts written by children

Adelaide H.P. Silva (UFPR/DELLIN, LIAMF/CNPq);
Fabiano Silva; Wanderlan Carvalho (UFPR/DInf/LIAMF);
Lourenço Chacon (UNESP/Marília)

7th. Eicefala

- Why orthography?

  - spelling is taken as a parameter to evaluate whether an individual is literate or not;

  - children are submitted to an examination that evaluates their performance in spelling twice in Elementary School (3rd and 5th grade)

7th. Eicefala

- In spite of that...

  - 34% of Brazilian children do not achieve the minimum score to be considered literate;

  - there is no planning about the spelling competence expected of students in each grade (Morais, 2000).

7th. Eicefala

- Considering the scenario

  - we started the elaboration of a system that learns the patterns of "errors" produced by children of Elementary School

7th. Eicefala

- Objetives

  - to use computational tools as an aid to the spelling teaching process;

  - to elaborate a system that

    - identifies the most recurrent types (patterns) of "errors" in written productions of children;

    - simulates orthographic “errors” (deviant forms), and relate them to the canonical written forms.

7th. Eicefala

- “Error” typology

  - 1st step towards the elaboration of the system;
  - based on the typology provided by Chacon, Pezarini (2018) for written productions of children in Elementary School

7th. Eicefala

- Chacon, Pezarini (2018):

  - different types of “errors”;

  - processes of omissions ("pene" for "pente"); transpositions ("porfessor" for "professor"); substitutions ("sebola" for "cebola")

7th. Eicefala

- Chacon, Pezarini (2018):

  - gradiency in “errors” can be seen, e.g., in the use of a grapheme that represents a sound in an unexpected position, as well as in the use of a grapheme in the expected position, but representing a different sound than expected.

7th. Eicefala

- Dataset

  - consists of the written language database assembled by the "Language Studies Research Group" (Chacon, 2018);

  - contains texts produced especially for it by children from the first to the fifth yeal in a public school in the city of Marília (SP).

7th. Eicefala

- Dataset
  - texts result from a task of rewriting a story read by the teacher for children in each grade;
  - all departing texts are the same for all children, in all grades;
  - 04 different productions were collected during one year.

7th. Eicefala

- Dataset

  - 128 texts (65 texts from 3rd grade and 63 texts from 5th grade);

  - 45561 words (total);

  - 356 word/text (mean);

  - focus on texts from 3rd and 5th grades.

7th. Eicefala

- Data analysis

  - orthographic words as units of analysis ("erros" apply within words and encompass syllables and segments);

  - manual classification of "errors", following Chacon, Pezarini (2018).

7th. Eicefala

- Data analysis

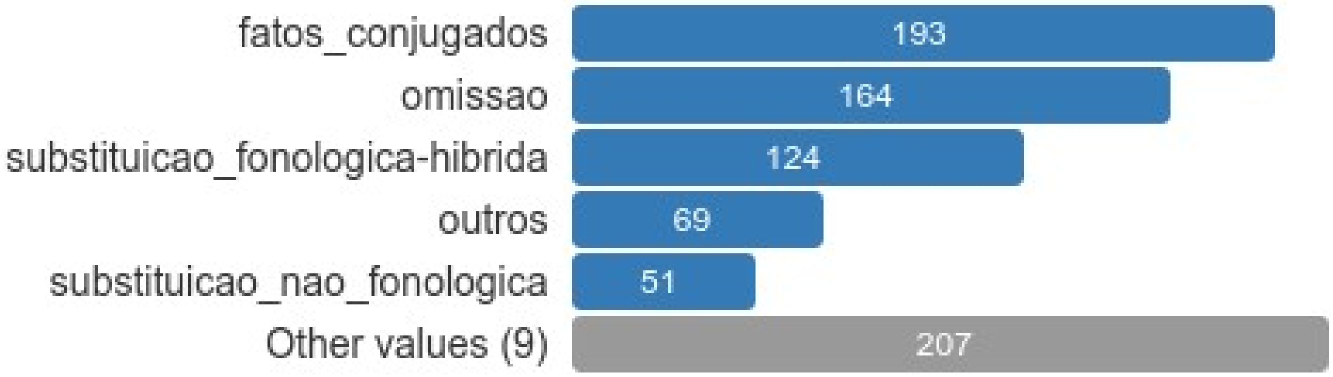
  - profiling using Pandas (open source library specific for data analysis running on Python);

  - profiling allows different grouping of “errors” types according to the school grade.

7th. Eicefala

- Data analysis

  - Histogram with the distribution of the types of "errors" in texts of the 3rd grade

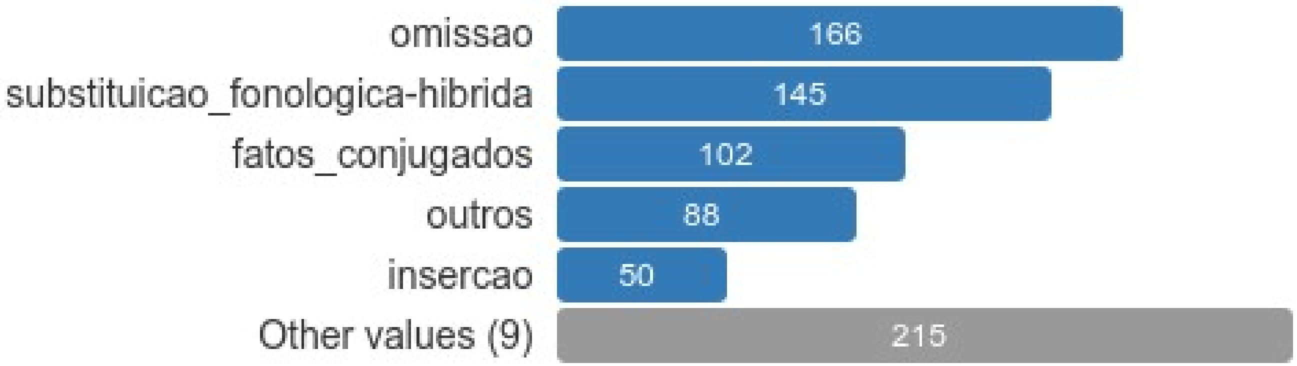


(Elaborated by the authors)

7th. Eicefala

- Data analysis

  - Histogram with the distribution of the types of "errors" in texts of the 5th grade



(Elaborated by the authors)

7th. Eicefala

- Discussion

  - decrease in the frequency of "combined facts" in the 5th grade (13%), compared to the 3rd. grade (24%)

  - “combined facts” is one class of “error” that comprises different types of “errors” in the same word.

7th. Eicefala

- Discussion

  - "omission" remains very close in both grades

  - BUT: an examination of data reveals that omissions in 5th grade occur mainly in the reduction of diphthongs of verbal inflected forms, <r> deletion or the 1st syllable of inflected forms of verb “to be” (e.g. “tava”). In the 3rd grade omissions seem to be less localized.

7th. Eicefala

- Discussion

  - "others" type: related to prosody (phonological word), it increases from the 3rd to the 5th grade.

  - Apparently, children start to build hypotheses about the segmentation of the speech chain more recurrently as they advance in formal education.

7th. Eicefala

- Discussion

  - profiling confirms the hypothesis that it is possible to find patterns of "errors" for the different grades of Elementary School;

  - patterns allow us to build automatic tools for accessing children's performance in the literacy process.

7th. Eicefala

- Expected results

  - help teachers understand the hypotheses underlying orthographic “errors” and identify the aspects to be worked on with the children;

  - help build guidelines for planning about the spelling competence expected from children in different grades of Elementary School.

7th. Eicefala

- Next steps

  - build automata for learning “errors” and associating them to the corresponding standard orthographic forms;

  - augment the corpus;

  - test the system with new data.

7th. Eicefala

- References

CHACON, L. Banco de Dados de Escrita no Ensino Fundamental I. Database. 2018.

CHACON L, PEZARINI I.O. Gradiência na correspondência fonema/grafema: uma proposta de caracterização do desempenho ortográfico infantil. In: César ABP, Seno MP, Capellini SA. Tópicos em Transtornos de Aprendizagem: Parte VI. Ribeirão Preto: Chacon, L.; Pezarini, I. O. Gradiência na correspondência fonema/grafema: uma proposta de caracterização do desempenho ortográfico infantil. In: César ABP, Seno MP, Capellini SA. Tópicos em Transtornos de Aprendizagem: Parte VI. Ribeirão Preto: Booktoy Livraria e Editora, 2018.

MORAIS, A. G. *Ortografia: ensinar e aprender*. Ed. Ática, São Paulo, 3. ed. 2000.