Acentua Fácil: A system to support Brazilian Portuguese Accentuation

Gabriella Selbach Staniecki[1], **Larissa de Freitas Astrogildo**[1], **Tiago Thompsen Primo**[1]

[1]Centro de Desenvolvimento Tecnológico – Universidade Federal de Pelotas (UFPEL)
96010-610 – Pelotas – RS – Brasil

{gsstaniecki,larissa}@inf.ufpel.edu.br,tiagoprimo@gmail.com

Abstract. *The main objective of Acentua Fácil is to assist students in the writing process and teachers in the learning process. This article presents a system to aid the accentuation process for Brazilian Portuguese. The input of the system is texts. These texts are analyzed through the use of grammatical rules and artificial intelligence(IA) techniques such as tokenization. For each word in the text, we show the graphic accentuation errors, together with explanations of how it should be a correct word.From the results we were able to use some validation metrics, such as histograms and confusion matrix, both of which are used when working with AI*

1. Introduction

Brazilian Portuguese is a complex language for students to learn and for teachers to support their students during their journey to learn the language. Their benefits aim to improve and develop the student's writing and speaking abilities, especially for native speakers.

Due to the plurality that we have in our language, the teacher must pass on to the student the formal writing concept. During the first years of literacy and learning/understanding the writing process, the school has a fundamental role in developing the individual.

Because of this, the National Curriculum Guidelines[1] was created as a goal to establish a collaboration between the spheres of government to guide the school curricula and ensure a common formation for all. This document addresses the students' needs at each educational stage, encompassing what they should learn, arguments to guide early childhood education, high school, EJA, and others.

Even the schools following the National Curriculum Parameters (NCP) still present difficulties in content such as graphic accentuation and scoring, either by personal or methodological factors. It is not uncommon for students to arrive in high school without knowing how to accentuate a word correctly.

According to data released by the INAF of 2018, a study carried out by Instituto Paulo Montenegro (IPM) and the NGO Ação Educativa, 13% of people who reach or finish high school is considered functionally illiterate, for example, people who have difficulty in making use of reading and writing. The National Institute of Educational

[1] http://basenacionalcomum.mec.gov.br/images/BNCC_EI_EF_110518_versaofinal_site.pdf

DOI: 10.5753/wave.2020.212081

Studies and Research Anísio Teixeira (INEP) holds the Basic Education Evaluation System (SAEB), where Portuguese and Mathematics subjects at 5th, 9th, and high school levels are assessed. Some alarming data, for example, even after 12 years of schooling, about 70% of students finish Basic Education without reading and understanding a simple text.

Today, there are some tools available whose goal is to help students in their writing process. Among them, we can mention traditional spelling checkers that display some errors, tools such as [Guide and Ferreira 2016] that make a more detailed analysis of spelling and grammar in texts, and academic works about Natural Language Processing and Computational Linguistics such as [Benevides 2017] and [Guide and Ferreira 2016]. However, none of the work cited is intended to assist students and teachers.

For this reason, this article presents a tool, called Acentua Fácil, which aims to help students and teachers. Students can use it as a support in their writing process, explaining what should be corrected from their own mistakes. Teachers can use it as support for their classes to determine which points their students have more difficulties.

This article is structured as follows: the second section presents the theoretical reference (Phonetics and Phonology); the third section presents the related works; the fourth section presents the tool developed (Acentua Fácil). Finally, the last section presents the conclusion.

2. Phonetics and Phonology

In the literature, different definitions are found addressing the issue of graphic accentuation to Brazilian Portuguese. The main distinction occurs in phonetic and phonological terms.

The term phonetic is related to physical aspects such: duration (the time it takes to pronounce a consonant or vowel), intensity (energy with which a consonant or vowel is expressed), and frequency (wave or vibrations that expresses a consonant or vowel) [Martins 1988]. These three paradigms have a great influence on the accent being placed on the word.

The term phonological is related to the theoretical aspects of the language: lexicon [Jr Câmara 1970], metric [Bisol 2012], and morphological [Lee 1995]. In this work, the phonological pattern will be used for the accent attribution since its greater application happens using rules, as explained by [Benevides 2017].

[Bechara 2009] deals with the issues of word tonicity and syllabic division. Considering this subject, some nomenclatures are important; they are monosyllable, oxytone, paroxytone, and proparoxytone. Also, in its grammar, the author explains how the tonicity of the word and its syllabic division influence the accent's position. The author explains how the tonicity of the word and its syllabic division influence the accent's position.

2.1. Brazilian versus Portuguese

Even having similarities the Portuguese of Brazil and Portugal have some very significant differences, especially in the speech where the phonetic issue is different, we use the pronoun before the verb, but in Portugal, the structure is the opposite. Specifically in matters of accentuation, the Portuguese do not use the circumflex accent, which can generate some strangeness. But what can be considered most different are some words that receive accents that for us Brazilians are totally meaningless, such as: sémen, xénon, vómer, fénix, ténis, bonús. Other words are accented like dêmos and íman and the Lusitanians are not [Batista 2018].

3. Related Works

The research of [Nunes 2017] surveyed with high school students from the state of Pernambuco to raise questions about the orthography of the Portuguese language. The most frequent mistakes in the corpus produced were regular mistakes. Regular mistakes are those that can be easily solved by learning orthographic rules, for example, graphic accentuation, absence of the letter ‘r’ in the infinitive (‘comer’ [eat], ‘vender’ [sell], ‘amar’ [love]), among others.

The research of [Fiorio et al. 2019] developed software called Linguistic. This software aims to perform the analysis and extraction of characteristics of the Portuguese language’s texts, using structural levels of the language as a function of words, grammatical levels, and morphological classes. After performing some tests, the authors observed that the tool satisfactorily implements the extraction of characteristics, presenting some analysis errors that should be solved in future works.

The research of [Guide and Ferreira 2016] investigated the accentuation in Brazilian Portuguese using and developing computational mechanisms to validate their hypotheses. The analysis was based on solutions arising from theoretical phonology. In this work, different models were constructed, divided into two groups of a probabilistic nature: (1) uses the idea and N-gram based on the probability of size chains “n”, being a simple pattern model; (2) uses a Naive Bayes classifier that takes into account the vector of traits that includes morphological, prosodic, and segmented characteristics of words. Finally, the author compares the classifier’s performance in the accentuation task with other authors’ values, showing that the proposed tool obtained a higher value than the others.

Still, we can find some systems available online such as Imaginei[2], and Descomplications[3], where students can submit their texts to be analyzed. The processing/analysis performed by each of these platforms can take about a day or more. Finally, none of the platforms is the result displayed to the student immediately.

Also, there are automatic brokers such as Grammarly[4], where the student can check whether their text is written correctly or not. The first checker performs a deeper analysis of the text explaining the error it found, showing an evolution of the user in writing, but is available only for the English language. The second proofreader is available for Portuguese, but it only does the spelling correction; it does not give a detailed explanation of how to correct the error.

In Acentua Fácil, as in related works, correcting students’ mistakes is a point of extreme relevance. The tool automatically displays the student’s errors in the text, which can be observed and explored at the moment, which can significantly assist the writing process. Acentua Fácil will also make available materials related to grammar rules to

[2] https://www.imaginie.com.br/

[3] https://descomplica.com.br/

[4] http://www.grammarly.com/

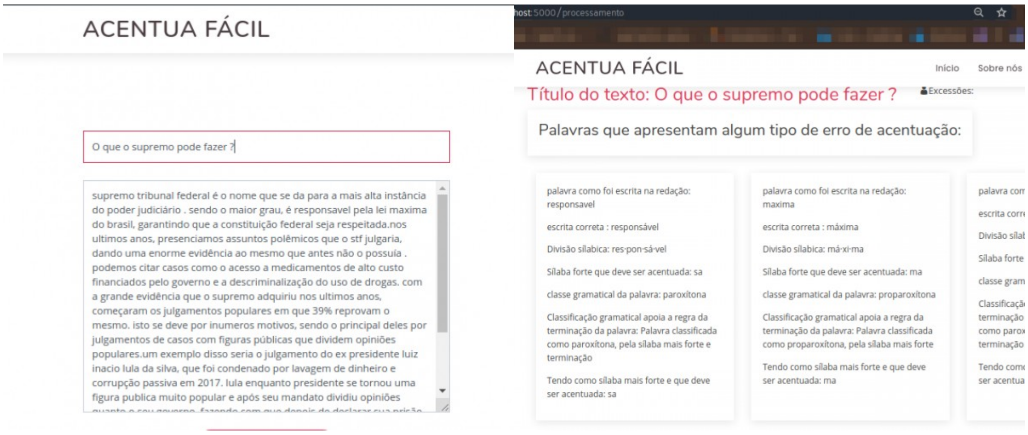
know where he has some writing problems and learns from his mistakes.

4. Acentua Fácil

According to the challenges described in the introduction and related works, the computational system, called Acentua Fácil, was developed. The system can assist students and teachers in graphic accentuation, focusing on words classified as monosyllables, oxytones, paroxytones, and proparoxytones. The development of the Acentua Fácil can be separated into two steps: I - Processing Texts, II- Interface.

At the Processing Texts step, techniques such as tokenization, speech, and marking parts were used. Still, web crawlers were built to serve as a basis for further analysis and validation of what was developed. The first accent correction is made using the Portuguese Language Dictionary (VOP)[5], regular expressions were implemented to identify which words would fit which rules. The linguistic rules are defined in the grammar of [Bechara 2009].

On the other hand, the Interface step is a website, where a text can be submitted for analysis, and important information can be checked to build a good text. For that matter, this system gives tips on a graphic accent.



O que o supremo pode fazer ?

supremo tribunal federal é o nome que se da para a mais alta instância do poder judiciário . sendo o maior grau, é responsavel pela lei maxima do brasil, garantindo que a constituição federal seja respeitada.nos ultimos anos, presenciamos assuntos polêmicos que o stf julgaria, dando uma enorme evidência ao mesmo que antes não o possuía . podemos citar casos como o acesso a medicamentos de alto custo financiados pelo governo e a descriminalização do uso de drogas. com a grande evidência que o supremo adquiriu nos ultimos anos, começaram os julgamentos populares em que 39% reprovam o mesmo. isto se deve por inumeros motivos, sendo o principal deles por julgamentos de casos com figuras públicas que dividem opiniões populares.um exemplo disso seria o julgamento do ex presidente luiz inacio lula da silva, que foi condenado por lavagem de dinheiro e corrupção passiva em 2017. lula enquanto presidente se tornou uma figura publica muito popular e após seu mandato dividiu opiniões

host:5000/processamento

[redacted]

Título do texto: O que o supremo pode fazer ?

Palavras que apresentam algum tipo de erro de acentuação:

palavra como foi escrita na redação: responsavel

escrita correta : responsável

Divisão silabica: res·pon·sá·vel

Sílaba forte que deve ser acentuada: sa

classe gramatical da palavra: paroxítona

Classificação gramatical apoia a regra da terminação da palavra: Palavra classificada como paroxítona, pela sílaba mais forte e terminação

Tendo como sílaba mais forte e que deve ser acentuada: sa

palavra como foi escrita na redação: maxima

escrita correta : máxima

Divisão silabica: má·xi·ma

Sílaba forte que deve ser acentuada: ma

classe gramatical da palavra: proparoxítona

Classificação gramatical apoia a regra da terminação da palavra: Palavra classificada como proparoxítona, pela sílaba mais forte

Tendo como sílaba mais forte e que deve ser acentuada: ma

palavra com

escrita corr

Divisão silab

Sílaba forte

classe gram

Classificaçã terminação como parox terminação

Tendo com ser acentua

Figure 1. The screen of the Acentua Fácil.

When we finished the first version of the system, we decided to conduct surveys to see the performance of Acentua Fácil to find accentuation errors.

This article shows analysis made between the corrections of ENEM style essays performed by UOL and the same texts submitted in the developed platform. For the analysis, we use histograms and confusion matrix.

4.1. Histogram

The histogram is a bar graph used in a data set with different values; we can visualize only the frequency distribution of data [Scott 2008]. In this work, the histogram was used to display how many errors were found in the ENEM style essays. Being considered only the question of whether or not errors exist, disregarding any classification.

[5] http://www.portaldalinguaportuguesa.org/

4.2. Confusion Matrix

When we build a model and evaluate how to correct the data presented, we can implement a confusion matrix. We start from a test set with known and true data. We compare this information with ours, so we can verify how many samples were classified correctly or not. [Artasanchez and Joshi 2020]

The confusion matrix is considered a table, working from two sets, one with true data (Test) and the other with the model (Predicted) and the classification of the information. From this classification, we can extract fundamental metrics for validation of what we are developing. Here in this work, we will use a binary classification of 0 or 1, so we can have the following classes:

- True positive (TP): we predicted that the output would be 1, and the true data is also 1.
- True negative(TN): we predicted that the output would be 0, and the true data is also 0.
- False positive(FP): we predicted that the output would be 1, but the true data presented 0 as output.
- False negative(FN): we predicted that the output would be 0, but the true data presented 1 as output.

To evaluate a classifier is used as main metrics accuracy, precision, recall, e f1-score. The accuracy is the most general value that we have; it tells us how much of a hit in total our model presented. The precision evaluates how accurate our model's positive data was, whereas the recovery refers to the amount of data predicted as positive and the actual positive. The recall tells us what proportion of positive values our model found correctly. The f1-score will be the harmonic mean between precision and recall.

4.3. Preliminary Tests

As preliminary tests were captured, the first 100 essays of UOL. These essays were processed using Acentua Fácil and analyzed. We built two datasets. The first dataset contains an essay, a dictionary with the corrections made, the total number of errors found in the essay, a flag whether or not the word has been corrected, and the type of correction. And the second dataset contains an essay, a flag whether or not the word has been corrected.

From each dataset, we could already extract the respective values (real and predicted) and ensure that the data were aligned. It was verified word by word of each essay. And only after this step that the vectors for the generation of the confusion matrix were obtained.

Our confusion matrix follows a binary structure being the data 0 for cases where there was no correction and 1 for cases where a correction occurred. The real set is extracted from the UOL. For example, the following words [órgão, próprio, outro, felicidade] could return the set [1,0,0,0] in it only the word organ was corrected. The predicted values represent the corrections made by Acentua Fácil, where it may or may not be equal to the real set, being, for example, [1,0,1,1]. Here we can notice that the last values were corrected but should not have been.

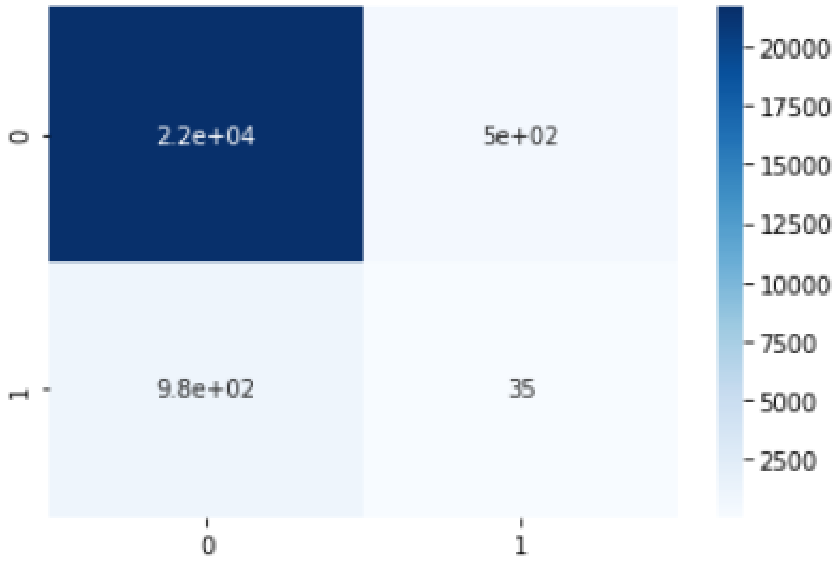
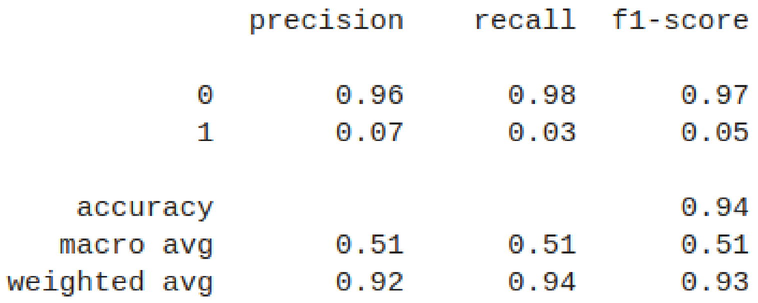


Figure 2. The screen of the Confusion Matrix between manual corrections and corrections of Acentua Fácil .

We used Sklearn[6] library to plot the confusion matrix (Figure 2). With the data presented in Figure 2, it is possible to calculate the metrics mentioned in the session 4.2. These metrics are shown in Figure 3.



```
               precision    recall  f1-score

           0        0.96      0.98      0.97
           1        0.07      0.03      0.05

    accuracy                            0.94
   macro avg        0.51      0.51      0.51
weighted avg        0.92      0.94      0.93
```

Figure 3. Return of the classification_report function of the metrics package.

We can notice that for the positive values, our model registered low precision and recall. We have as the hypothesis that such fact occurred for the following reasons:

- The VOP dictionary that we are using in processing is not 100% complete and presents some faults both in the lack of some words and wrong correction of other words.
- More general cases have been treated in the processing. Brazilian Portuguese has several exceptions to its grammatical rules that still need to be studied more deeply and according to the possibility of being added to the tool's processing code.

Figure 4 shows the number of errors displayed in the manual correction by UOL; on the y-axis, there is the number of essays, and on the x-axis, there is the number of errors

[6] https://scikit-learn.org/

present in them, being the minimum error value zero and the maximum seven. Figure 5 shows the number of errors obtained by the correction on the Acentua Fácil; the y-axis represents the number of essays and the x-axis the number of errors corrected, being the minimum value zero, and the maximum twenty.

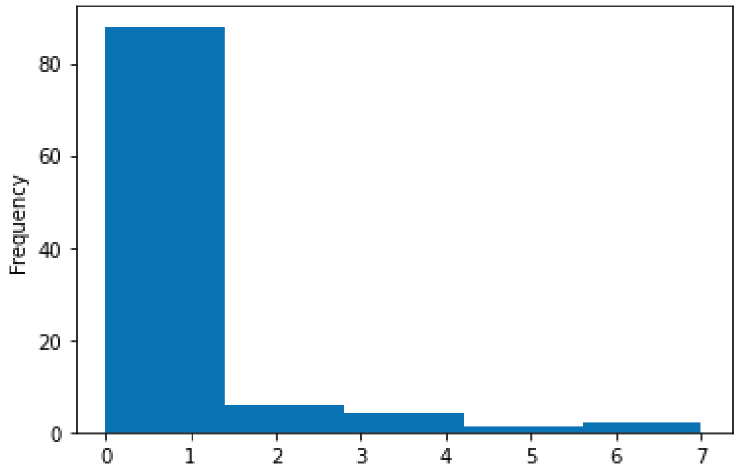


Figure 4. Histogram of manual corrections performed by UOL

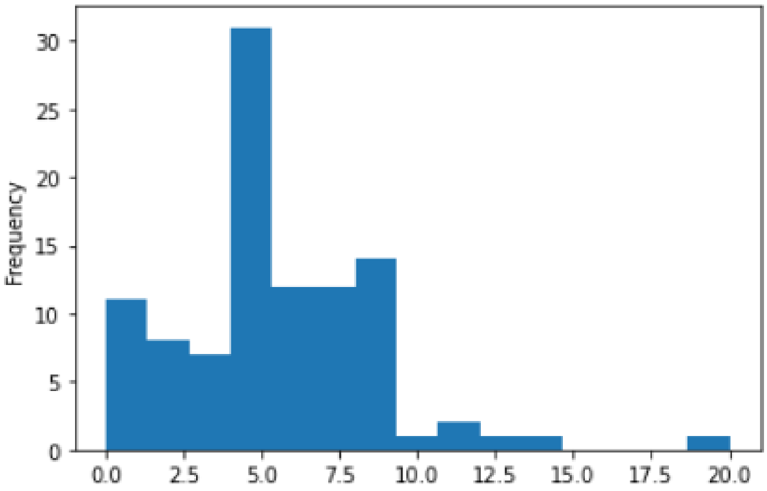


Figure 5. Histogram of automatic corrections made by Accentua Fácil.

Both histograms were obtained using bins metrics with a value of 15. This value was chosen because, with it, the graphs were better visually.

5. Conclusion

With the plurality of the Portuguese language, its teaching-learning cannot be considered a trivial task. Computational Linguistics is still taking small steps. Research and technological tools prove to be of paramount importance, contributing to the learning process, but there are still gaps where researchers can explore.

The work presented here enters as a contribution to diminishing these gaps. The aim was to investigate the issues related to graphic accentuation, carrying out the development of a system from grammatical rules, can show points of error and how to correct them.

With the analysis of the confusion matrix and the histogram, one can see the points that still need to be improved and those that have not been treated yet. It is intended to analyze the other ENEM essays collected from the UOL site, performing statistics with the data collected to observe the main errors found, which are the predominant classes of errors, and which aspects still need to be improved in the rules developed.

The system was thought to work initially with graphic accentuation, one of the points that are evaluated in the ENEM style newsrooms; as a future implementation, we thought to approach other aspects evaluated in the writing of this exam, such as punctuation, exchange of phonemes(s) (r, rr), among others.

References

Artasanchez, A. and Joshi, P. (2020). *Artificial Intelligence with Python: Your complete guide to building intelligent apps using Python 3.x, 2nd Edition*. Packt Publishing.

Batista, P. (2018). Principais diferenças entre o português de portugal e o português do brasil.

Bechara, E. (2009). *Moderna Gramática Portuguesa*, volume 1. Editora Nova Fronteira Participações S.A.

Benevides, A. d. L. (2017). O acento primário em pseudopalavras: uma abordagem experimental. Master's thesis, Universidade de São Paulo.

Bisol, L. (2012). O acento e o pé métrico binário. *Cadernos de Estudos Linguísticos*, 22.

Fiorio, R., Varela, P., Albonico, M., and Semler, J. (2019). Linguisticun: Uma ferramenta de auxílio ao ensino da língua portuguesa e à linguística computacional. *Brazilian Symposium on Computers in Education (Simpósio Brasileiro de Informática na Educação - SBIE)*, 30(1).

Guide, B. F. and Ferreira, M. B. (2016). Abordagem computacional para a questão do acento no português brasileiro. Master's thesis, Universidade de São Paulo.

Jr Câmara, J. M. (1970). *Estrutura da Língua Portuguesa*, volume 34. Editora Vozes.

Lee, S. H. (1995). *Morfologia e fonologia lexical do português do Brasil*. PhD thesis, Universidade Federal de Campinas.

Martins, M. R. D. (1988). *Ouvir Falar: Introdução à Fonética do Português*. Coleção Universitária Série Linguística. Editora Caminho.

Nunes, V. (2017). Ortografia da língua portuguesa: Uma análise em textos de estudantes do ensino médio. *Revista Signos*, 38(2).

Scott, D. (2008). *Histograms: Theory and Practice*, pages 47–94.