STATISTICAL TEXT ANALYSIS ON THE HANDS OF BASIC SCHOOL STUDENTS AND TEACHERS: AN INTERDISCIPLINARY APPROACH

Humberto J. Bortolossi and José Felipe E. B. dos Santos
Fluminense Federal University, Brazil
hjbortol@vm.uff.br

In this paper, we present a free multiplatform software of our own development (available at http://www.uff.br/cdme/desktop/lpp/lpp-en.html) that provides an interactive environment in which Basic School students and teachers can experiment, explore and enjoy, through a very simple interface, the use of Statistics in a real-world application, namely, statistical text analysis. We also provide a set of activities to be used with the software in the classroom (these activities were successively refined from several workshops with Basic School students and teachers). Through this articulation, students can learn statistical concepts in the linguistic context and, at the same time, learn linguistic concepts practicing Statistics.

BACKGROUND

Researchers and curriculum documents have agreed to several aspects that should guide the teaching of Statistics in Basic School and Teacher Education. For instance, in USA, the GAISE document (Garfield et al., 2005) gives six recommendations to produce statistically educated students, namely, (1) emphasize statistical literacy and develop statistical thinking; (2) use real data; (3) stress conceptual understanding, rather than mere knowledge of procedures; (4) foster active learning in the classroom; (5) use technology for developing concepts and analyzing data; (6) use assessments to improve and evaluate student learning. In Brazil, the National Curriculum Parameters (NCP) present three skills that must be developed in High School Mathematics (Brasil, 2002, 2006): (1) representation and communication; (2) investigation and comprehension; (3) sociocultural science contextualization (including the understanding and the use of technological resources in human culture). For Statistics, the NCP emphasize its interdisciplinary aspect: "Statistics must be seen as a set of ideas and procedures for applying mathematics to real-world problems, more especially those from other areas".

However, how address all these recommendations? How to articulate skills, contents, technology and interdisciplinarity? The communities of Mathematics Education and Statistics Education have been conducting several researches in order to obtain answers to these questions. In this paper, we present our contribution to this line of action: a free software and a set of activities (all of our own) that promotes the GAISE and NCP recommendations using, for this, an interdisciplinary approach for the areas of Statistics and Linguistics.

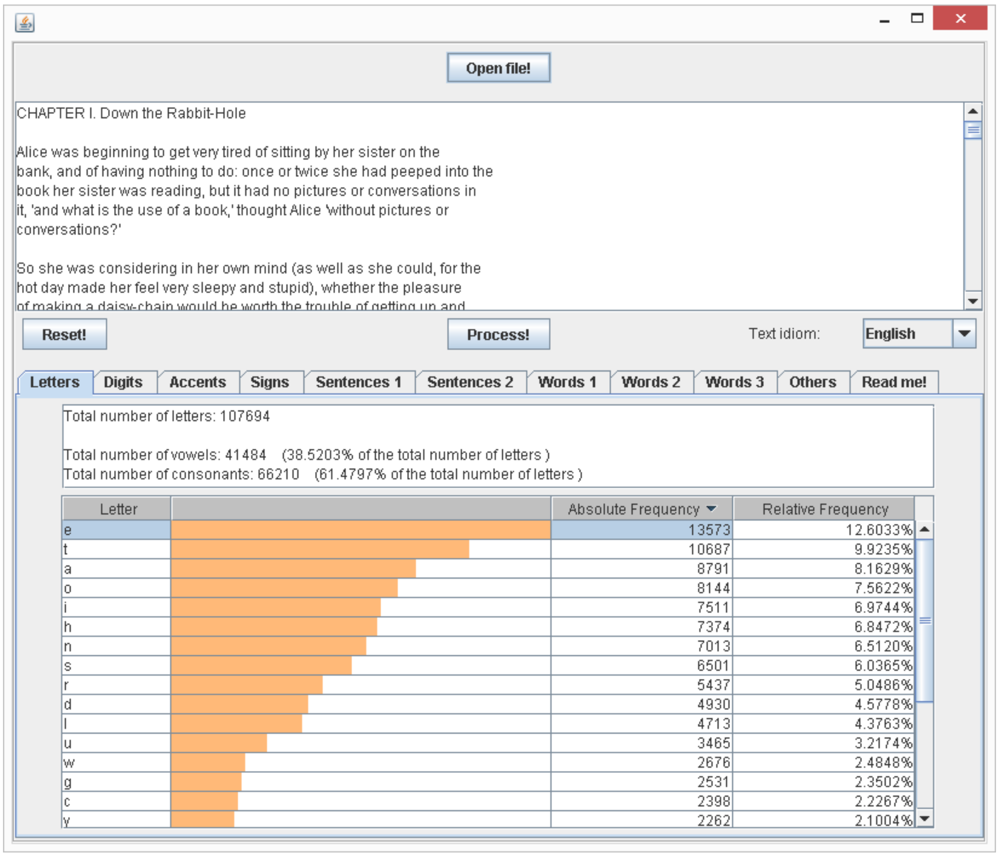
THE SOFTWARE LPP

Our software, written in the Java language, is divided into four modules that share a very simple graphical interface: basically, there is an input area where the user can enter text typing it directly or open a text file from his/her computer or, yet, use the technique of "copy and paste " (ctrl + c / ctrl + v).

Any kind of text can be entered: free entire books, poems, speeches, song lyrics, movies subtitles, thousands of digits of real numbers, etc. By pressing the "Process!" button, the given text will be then processed, and the results of a statistical analysis will be displayed numerically and graphically in several tabs.

Figure 1 shows the graphical user interface of the Module 2 (the main module of the software). This module counts the number of letters, digits, accents, punctuation marks, words and periods of a text. It also calculates the number of letters per word and words per period (showing mean, mode, median, variance and standard deviation of these quantitative variables), the longest periods, the shortest periods, the longest words and the shortest words. Data reports given in tables can be sorted by clicking repeatedly on the corresponding column header.

In: M.A. Sorto (Ed.), *Advances in statistics education: developments, experiences and assessments. Proceedings of the Satellite conference of the International Association for Statistical Education (IASE),* July 2015, Rio de Janeiro, Brazil. ©2015 ISI/IASE iase-web.org/Conference_Proceedings.php



CHAPTER I. Down the Rabbit-Hole

Alice was beginning to get very tired of sitting by her sister on the
bank, and of having nothing to do: once or twice she had peeped into the
book her sister was reading, but it had no pictures or conversations in
it, 'and what is the use of a book,' thought Alice 'without pictures or
conversations?'

So she was considering in her own mind (as well as she could, for the
hot day made her feel very sleepy and stupid), whether the pleasure
of making a daisy-chain would be worth the trouble of getting up and

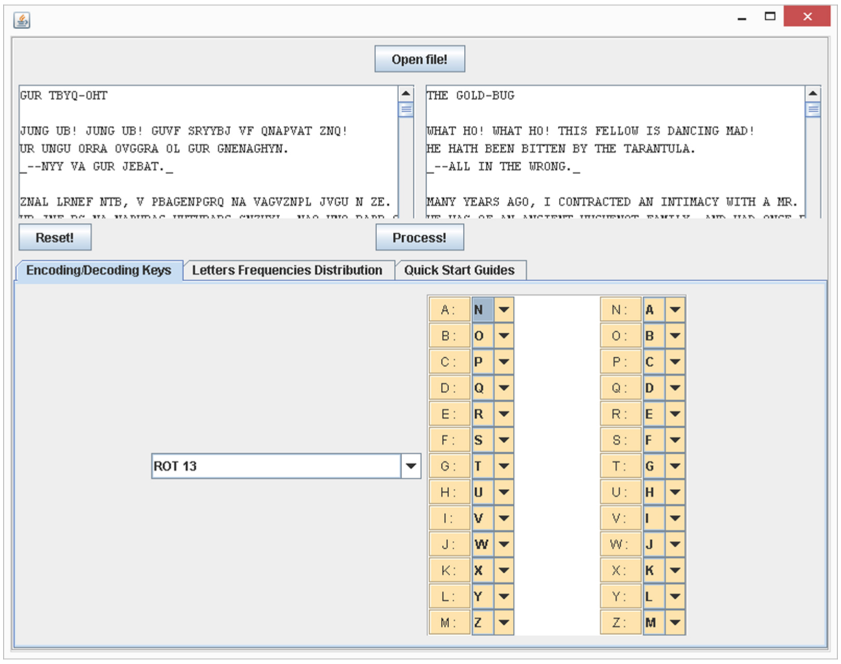
Total number of letters: 107694

Total number of vowels: 41484 (38.5203% of the total number of letters)
Total number of consonants: 66210 (61.4797% of the total number of letters)

Letter	Absolute Frequency	Relative Frequency
e	13573	12.6033%
t	10687	9.9235%
a	8791	8.1629%
o	8144	7.5622%
i	7511	6.9744%
h	7374	6.8472%
n	7013	6.5120%
s	6501	6.0365%
r	5437	5.0486%
d	4930	4.5778%
l	4713	4.3763%
u	3465	3.2174%
w	2676	2.4848%
g	2531	2.3502%
c	2398	2.2267%
y	2262	2.1004%

Figure 1. Statistics of "Alice's Adventures in Wonderland" by Lewis Carroll

The first module is a special adaptation of the main one to study the Caesar cipher, a simple cryptology substitution technique. More specifically, this module allows the user to decode/encode a text by an arbitrary permutation of the letters (Figure 2). As it is known (Cozens & Miller, 2013), this kind of cryptology technique may be easily broken using the frequency analysis of the letters (for instance, since the letter "e" is the most frequent in lengthy and regular texts in English, the most frequent letter in the ciphertext is probably the letter "e").



```
GUR TBYQ-OHT

JUNG UB! JUNG UB! GUVF SRYYBJ VF QNAPVAT ZNQ!
UR UNGU ORRA OVGGRA OL GUR GNENAGHYN.
_--NYY VA GUR JEBAT._

ZNAL LRNEF NTB, V PBAGENPGRQ NA VAGVZNPL JVGU N ZE.
```

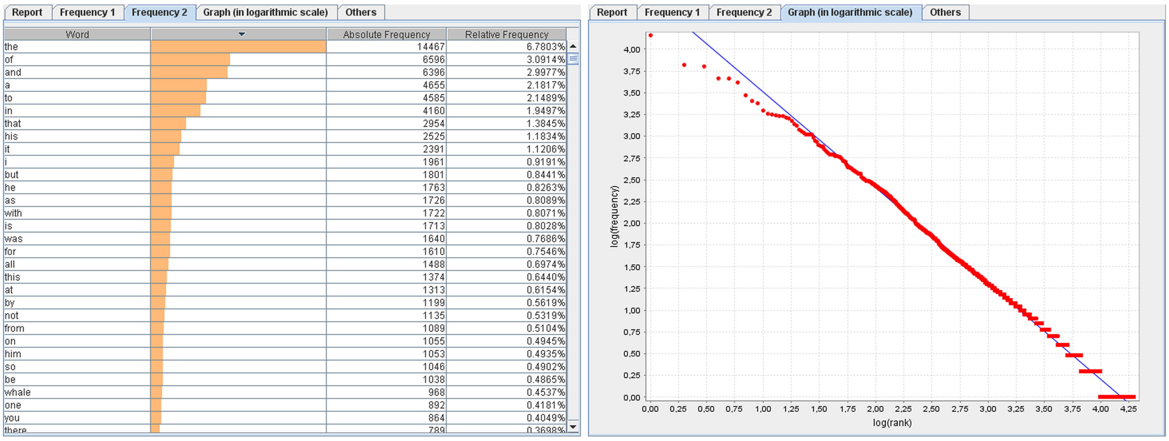
```
THE GOLD-BUG

WHAT HO! WHAT HO! THIS FELLOW IS DANCING MAD!
HE HATH BEEN BITTEN BY THE TARANTULA.
_--ALL IN THE WRONG._

MANY YEARS AGO, I CONTRACTED AN INTIMACY WITH A MR.
```

ROT 13

A :	N	N :	A
B :	O	O :	B
C :	P	P :	C
D :	Q	Q :	D
E :	R	R :	E
F :	S	S :	F
G :	T	T :	G
H :	U	U :	H
I :	V	V :	I
J :	W	W :	J
K :	X	X :	K
L :	Y	Y :	L
M :	Z	Z :	M

Figure 2. "The Gold-Bug" by Edgar Allan Poe decrypted using the cipher ROT 13

The third module was designed to study the surprising Zipf's Law (Grzybek, 2006). This empirical power law, proposed by the linguist George Kingsley Zipf (1902-1950) of Harvard University, suggests that in a text with a large number of words, the frequency f of occurrence of a word as a function of its position r in a list sorted by frequency of occurrence has the following form: $f = C/r^a$, where C and a are constants, with the value of a close to 1 (Clauset, Shalizi & Newman, 2009). Note that, in the variables $y = \log(f)$ and $x = \log(r)$, the Zipf's Law is expressed as an affine function: $y = b + a\,x$, with $b = \log(C)$. Thus, the coefficients a and b can be estimated, for example, using the method of Least Squares. This whole process is automated in the software. Figure 3 illustrates the Zipf's Law for the novel "Moby Dick" of Herman Melville (C = 40536.4574 and a = 1.1025).



Word		Absolute Frequency	Relative Frequency
the		14467	6.7803%
of		6596	3.0914%
and		6396	2.9977%
a		4655	2.1817%
to		4585	2.1489%
in		4160	1.9497%
that		2954	1.3845%
his		2525	1.1834%
it		2391	1.1206%
i		1961	0.9191%
but		1801	0.8441%
he		1763	0.8263%
as		1726	0.8089%
with		1722	0.8071%
is		1713	0.8028%
was		1640	0.7686%
for		1610	0.7546%
all		1488	0.6974%
this		1374	0.6440%
at		1313	0.6154%
by		1199	0.5619%
not		1135	0.5319%
from		1089	0.5104%
on		1055	0.4945%
him		1053	0.4935%
so		1046	0.4902%
be		1038	0.4865%
whale		968	0.4537%
one		892	0.4181%
you		864	0.4049%
there		789	0.3698%

Figure 3. Zipf's Law for the novel "Moby Dick" by Herman Melville

The last module allows the analysis of the vocabulary richness of a text. More precisely, it plots a graph showing the number of different words $V(N)$ depending on the number N of words read from beginning to end of text, as well three fitting functions proposed to model $V(N)$ (Baayen, 2001). Figure 4 shows such graphs for the short story "The Celebrated Jumping Frog of Calaveras County" by Mark Twain. This text has $N = 2634$ words and $V(N) = 746$ different words, so its percentage of different words, also known as *type token ratio*, is equal to $TTR(N) = V(N)/N =$ 28.3219%.

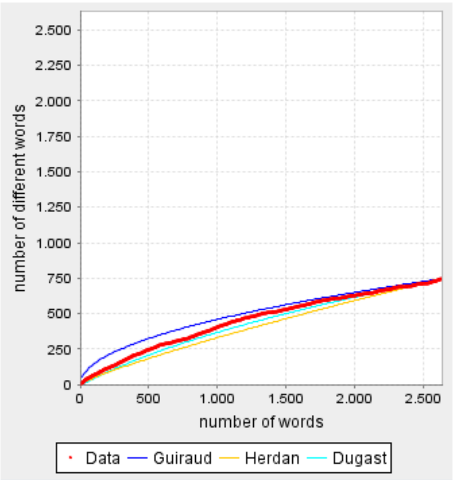


Figure 4. Vocabulary richness of the short story "The Celebrated Jumping Frog of Calaveras County" by Mark Twain

SAMPLE ACTIVITIES

We have elaborated a set of exercises to be worked out with the software. These exercises are available at http://www.uff.br/cdme/desktop/lpp/lpp-en.html as an RTF (Rich Text Format) document (so the teacher can make adjustments in the exercises according to the profile of his or her class). Next, we highlight some of them.

Sample Activity 1: Frequency Distribution of Letters, Cryptology and Combinatorics

As a first exercise, we encourage the students to investigate the frequency distribution of letters in texts with different sizes, languages (English, French, German, Spanish, Portuguese, etc.) and narratives (poems, short stories, speeches, song lyrics, movies subtitles, full books, newspaper articles). The objective is to make him or her realize that this distribution is a kind of "statistical signature" of the language of the text and, in this case, the sample size is very important (how does the distribution change with the sample size?).

After this investigation, the students may be presented to the Caesar cipher and about how to use the frequency distribution of letters of the encrypted text to decrypt it. They may play with these concepts exchanging encrypted messages with their classmates by email. This context also allows explorations in combinatorics: How many different keys can be built using permutations of the 26 letters of the alphabet? List some keys with the following characteristic: they decrypt messages that were encrypted by themselves (that is, keys associated with permutations that are involutions).

Still in this context, another interesting question is about the distribution of consonants and vowels in words. For instance, is there an English word with four or more letters where the number of vowels is greater than or equal to 80% of the total number of letters in the word? What about 88%? Acronyms, Roman numerals, onomatopoeias and words incorporated from other languages are not allowed.

Sample Activity 2: Lipograms

A *lipogram* is a literary composition (a kind of constrained writing) characterized by the deliberate omission of certain letters of the alphabet in the text (that is, these letters have zero frequency in the text). The challenge here is to create compositions where the most frequent letters of a particular language are not used. For instance, it is easy to write a long text in English without the letters "x", "k" and "z", but it is difficult to write a long one without the letter "e" (in Portuguese, it is difficult to write a long text without the letter "a").

There are already, in English, entire books written in the lipogram form: "Gadsby: A Lipogram Novel" (1939) by Ernest Vincent Wright and "A Void" (1994) translated by Gilbert Adair from the French original "La Disparition" (1969) by Georges Perec (both the original and the translation are written lipogrammatically). Therefore, as an excellent exercise in writing, students may be asked to compose their own lipograms using the software to check if the missing letter is really missing.

Sample Activity 3: Composing Texts with a Predetermined Mean for the Number of Letters per Word

The periods "Man, yes, you are mad, sad and bad." and "That is a good idea!" have both the same mean for the number of letters per word, namely, 3. The standard deviation of this quantitative variable is 0 for the first period (so all words have exactly the same number of letters) and about 1.26491 for the second one. As another interesting writing exercise, students may be asked to try to compose texts where the number of letters per word has a predetermined mean exactly (3, for example). We think that this exercise in the linguist context may help students to understand the concept of mean and standard deviation. Remark: Mike Keith wrote in 2010 the book "Not A Wake: A Dream Embodying π's Digits Fully for 10000 Decimals" in which the numbers of letters in successive words follow the digits of the number π (3.14159265358979323846...).

Sample Activity 4:

In this activity, we proposed that students be organized into teams and each team analyses statistically different books of a same author (these statistics include position and dispersion measures of the number of letters per word and words per period, vocabulary richness, etc.). The objective is to investigate whether different authors have or have not different statistics. Santos

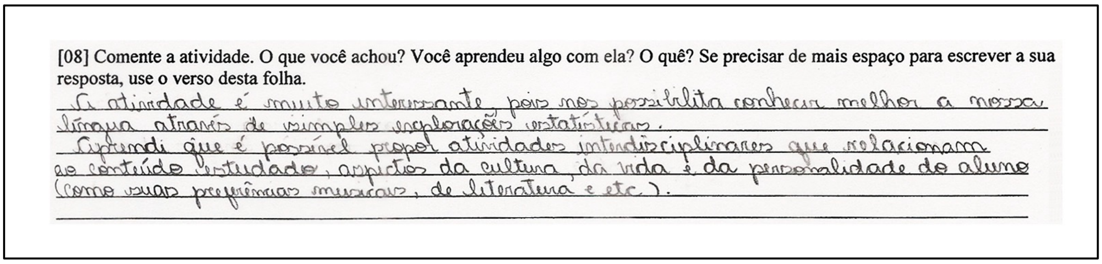
(2015), for instance, working with his High School students, discovered that the Brazilian writer Euclides da Cunha (1866-1909) has a peculiar characteristic: the median of the number of letters per word of his books is 5 while the same median for books of other contemporary writers (José de Alencar, Machado de Assis and Lima Barreto) is 4.

SOME CONCLUDING QUALITATIVE REMARKS

In 2013 and 2014, the authors tested (for the first time) the software and the proposed activities with two High School classes (Figure 5) with 59 students in total, one class of a technical course in Informatics with 8 students in total, and one class for in-service teachers of Mathematics with 12 teachers in total. A qualitative analysis of this experiment is available in Santos (2015). Here we provide some key remarks from this report.

Figure 5. High School students using the software LPP in the computer lab

- The use of Statistics and Mathematics in the study of languages (Literature and Linguistics) seems to be unknown to teachers and students in Basic School. In a survey applied in the beginning of his work, Santos (2015) reports that none of your 67 students knew any application of Mathematics and Statistics in Portuguese (they stated, however, the importance of Portuguese in the learning of Mathematics and Statistics). In addition, only one of the twelve in-service teachers in the survey knew an application of Mathematics and Statistics in Portuguese, namely, the Caesar cipher. Therefore, we believe that the activities we propose may serve as an excellent introduction to teachers and students of an important application of Statistics in our time: Natural Language Processing and Text Mining. The following testimony of one of the participating in-service teachers sums up this view and its didactical implication: "The activity is very interesting because it enables us to understand better our language through simple statistics explorations. I've learned it is possible to propose interdisciplinary activities relating aspects of culture, life and the student's personality (as his musical preferences, literature, etc.) to the study content." (Figure 6).

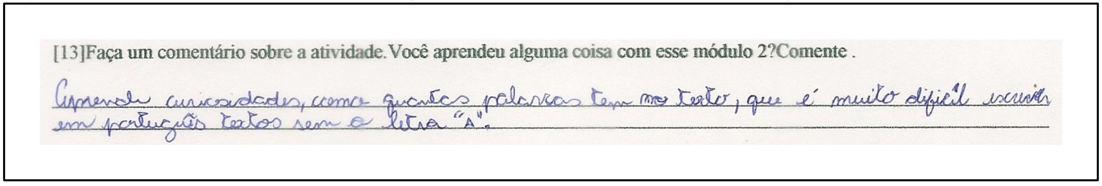


[08] Comente a atividade. O que você achou? Você aprendeu algo com ela? O quê? Se precisar de mais espaço para escrever a sua resposta, use o verso desta folha.

A atividade é muito interessante, pois nos possibilita conhecer melhor a nossa língua através de simples explorações estatísticas.
Aprendi que é possível propor atividades interdisciplinares que relacionam ao conteúdo estudado, aspectos da cultura, da vida e da personalidade do aluno (como suas preferências musicais, de literatura e etc.).

Figure 6. A testimony (in Portuguese) of one of the participating in-service teachers

- One practical advantage of our proposal refers to data collection in classroom: it is quite easy nowadays to collect free data on the Internet with different characteristics for statistical text analysis (indeed, in our experiment, the participating students and in-service teachers were able to analyze statistically 78 books of classical literature in Portuguese). Therefore, students may very easily experience how variability in texts is captured by Statistics. Even outliers may have a linguistic personification: a Portuguese lipogram, for instance, where the letter "a" does not appear. Here is a testimony of one of the participating High School students that realized this fact: “I learned interesting facts, such as how many words a text may have and that it is very difficult to write texts in Portuguese without the letter ‘a’.” (Figure 7).



[13]Faça um comentário sobre a atividade. Você aprendeu alguma coisa com esse módulo 2?Comente .

Aprendi curiosidades, como quantas palavras tem no texto, que é muito difícil escrever em português textos sem a letra “A”.

Figure 7. A testimony (in Portuguese) of one of the participating High School students

- Through the activities, students may appreciate the role of the computer in the experiments: try, for instance, to count by hand how many times the letter “e” appears in a book with 1 million letters. Without the help of computers, this would be a very boring and prone error task (it is said that Zipf employed his graduate students to count words for him). Indeed, the use of the computer in the Linguistic context allows students to play with big data, a not so common experience in the classroom.

While we have presented some examples of activities that explore Statistics in the linguistic context but, of course, the subject is so rich that others possibilities of projects unfolds. Here are some examples that we plan to try with students in a next time: for pairs of characters in a literary text, who is cited more? (Romeo or Juliet? Don Quixote or Sancho Panza? Frodo Baggins or Sam Gamgee? Etc.) What appears more, “this” or “that”? Is the word “not” always among the 30 more frequent words in a long English text?

We hope this article can disseminate among teachers and researches in Statistical Education the so unknown interdisciplinary connections between Statistics and Linguistics.

As future work, we intend to make a more in-depth quantitative study to try to measure the effect of the software and the proposed activities on student learning.

ACKNOWLEDGMENTS

The second author greatly acknowledges CAPES/MEC (Brazil) for its financial support for the PROFMAT/UFF/SBM Master’s Degree Program.

REFERENCES

Baayen, R. H. (2001). *Word Frequency Distributions*. Dordrecht: Kluwer Academic Publishers.

Brasil. (2002). *PCN+ Ensino Médio: Orientações Educacionais Complementares Aos Parâmetros Curriculares Nacionais. Ciências da Natureza, Matemática e Suas Tecnologias*. Brasília: Ministério da Educação.

Brasil. (2006). *Orientações Curriculares para O Ensino Médio: Ciências da Natureza, Matemática e suas Tecnologias.* Brasília: Ministério da Educação, Brasília.

Clauset, A., Shalizi, C. R., & Newman, M. E. J. (2009). Power-Law Distributions in Empirical Data. *SIAM Review*, *51* (4), 232-236.

Cozzens, M., & Miller, S. J. (2013). *The Mathematics of Encryption: An Elementary Introduction*. Washington, DC: American Mathematical Society.

Garfield, J. et al. (2005). *Guidelines for Assessment and Instruction in Statistics Education (GAISE) College Report*. Alexandria, VA: American Statistical Association.

Grzybek, P. (2006). *Contributions to the Science of Text and Language: Word Length Studies and Related Issues*. The Netherlands: Springer.

Santos, J. F. E. B. (2015). *Matemática, Estatística e Linguística: Um Relato de Experiência Interdisciplinar no Contexto da Escola Básica*. Master’s thesis, Instituto de Matemática e Estatística, Niterói: Universidade Federal Fluminense.