palgrave
communications
HUMANITIES | SOCIAL SCIENCES | BUSINESS

ARTICLE

https://doi.org/10.1057/s41599-019-0311-0 OPEN

Quantification of textual comprehension difficulty with an information theory-based algorithm

Louise Bogéa Ribeiro[1,2], Anderson Raiol Rodrigues[1,3], Kauê Machado Costa[4] & Manoel da Silva Filho[1]

ABSTRACT Textual comprehension is often not adequately acquired despite intense didactic efforts. Textual comprehension quality is mostly evaluated using subjective criteria. Starting from the assumption that word usage statistics may be used to infer the probability of successful semantic representations, we hypothesized that textual comprehension depended on words with high occurrence probability (high degree of familiarity), which is typically inversely proportional to their information entropy. We tested this hypothesis by quantifying word occurrences in a bank of words from Portuguese language academic theses and using information theory tools to infer degrees of textual familiarity. We found that the lower and upper bounds of the database were delimited by low-entropy words with the highest probabilities of causing incomprehension (i.e., nouns and adjectives) or facilitating semantic decoding (i.e., prepositions and conjunctions). We developed an openly available software suite called *CalcuLetra* for implementing these algorithms and tested it on publicly available denotative text samples (e.g., articles, essays, and abstracts). We propose that the quantitative model presented here may apply to other languages and could be a tool for supporting automated textual comprehension evaluations, and potentially assisting the development of teaching materials or the diagnosis of learning disorders.

[1] Institute of Biological Sciences, Federal University of Pará, Belém, Pará, Brazil. [2] Contemporary Art Museum, Federal University of Pará, Belém, Pará, Brazil. [3] Institute of Health Sciences, Federal University of Pará, Belém, Pará, Brazil. [4] National Institute on Drug Abuse Intramural Research Program, National Institutes of Health, Baltimore, USA. Correspondence and requests for materials should be addressed to M.da S. F. (email: manoel@ufpa.br)

Introduction

Textual comprehension is a cornerstone of learning that contributes to a range of educational and developmental phenomena, including language acquisition, building vocabulary and literacy (Nowak et al., 2000). Currently, a major limitation in the field is that skills and competencies associated with textual comprehension tend to be evaluated via the use of subjective criteria, typically related to outcomes, such as spelling test scores (Sigman et al., 2014).

Textual comprehension is often defined as mental representations made by semantic connections in a text wherein codes are translated and combined (Kendeou et al., 2014). This process is determined not by concepts of reading difficulty but by rules pertaining to statistical linguistics. Features such as entropy, fatigue and ultimately comprehension, can thus be gauged from the mathematical relations of word usage. Zipf (1935) established specific laws on word frequency over a body of textual information. He observed that only a few words are used often, and a higher number is used very rarely. These rules were mathematically formalized, and it was noted that the most common words in a text corpus occurred twice as often as the second most frequently used word and this pattern repeated itself for subsequent frequency rankings (the word in the position n appears 1/n times as often as the most frequent one).

Shannon (1948) further described how statistical linguistics impact on the user and communicator, specifically in relation to entropy and fatigue. He proposed a five-point law by which accuracy and scores can be estimated and configured owing to the conformity of the text (Shannon, 1948). First, the information source which produces a message or sequence of messages must be communicated to a receiving terminal. Second, a transmitter that operates on the message being transmitted must in some way produce a signal transference suitable for transmission over the given channel. Third, the channel itself must then act as a means to mediate the transmitted signal from the transmitter to the receiver. Fourth, the receiver ordinarily performs the inverse operation of the transmitter, in turn reconstructing the message sent from the original signal. The fifth and final element was the destination, which was comprised of the intended person, reader or observer. These functions are the overarching process framework in which Zipf's equation on textual comprehension operates.

Information theory-based studies have thus sought to measure the amount of information, informational entropy (Montemurro and Zanette, 2002, Montemurro and Zanette, 2011, Debowski, 2011, Kalimeri et al., 2015) and semantic information in a message (Bar-Hillel and Carnap, 1952, D'Alfonso, 2011, Bao et al., 2011, Marcelo and Damian, 2010), showing—from databases of texts—the universality in entropy across languages and a link between semantic information and statistical linguistics (Montemurro and Zanette, 2016, Montemurro, 2014). For example, linguistic analysis has found that the letter a is frequently used in English; however, it contains little information and has high entropy as the more similar are the letter occurrences, the higher is the entropy (Rosenfeld, 2000). However, these principles have yet to be systematically applied for the development of software suites capable of objectively extracting statistical patterns from text databases and using them to infer textual comprehension.

Another issue in the field is the scarcity of instruments developed for non-English languages. In this study, we focused on the Portuguese language. Some tools designed for English, such as the Flesch-Kincaid grade level and Coh-Metrix measurements, have been adapted to evaluate textual intelligibility in Portuguese (Gasperin et al., 2009). Although English has complex computational constructions, it has been observed that these are used infrequently (Thorne and Szymanik, 2015). However, despite these tools and the recent availability of Portuguese words repositories (Ferreira, 2014, Wiktionary, 2017, Mark, 2011, Github, 2017), there have been relatively few systematic analyses of the correlation between letter occurrences and syllabic structures in Portuguese (Hartmann et al., 2014, Soares et al., 2017, Rabêlo and Moraes, 2008).

Using an information theory-based approach, we developed a software tool (CalcuLetra) for quantifying average textual reading comprehension difficulty, measured under the premise that the words most frequently used in writing are the words most familiar to readers. We hypothesized that the level of reading comprehension was related to high probability (high degree of familiarity), low-entropy words, which was tested using information theory's objective quantitative parameters for given word occurrences in a bank of academic theses in Portuguese. We consider entropy as a degree of uncertainty about textual comprehension. Therefore, a highly comprehensible text would be defined by having many low-entropy, high probability words. We also collected samples of Portuguese denotative texts (e.g., articles, essays, and abstracts) as a test database for the developed software. We propose that the quantitative model presented in this paper can objectively evaluate the relative complexity of a given text, thus, constituting an objective evaluation tool for textual semantic decoding and comprehension level complexity.

Methods

Model parameters. We assumed each word's frequency f and occurrence probability P determined a corresponding knowledge probability P_k and unknowledge probability P_u from the correlation of scales between $0 \leq P \leq 4.987$ and $0 \leq P_k \leq 100\%$, with the average occurrence probability being $P_a = 0.003$ ($P_{k=}$ 50%) and $P_u = 100 - P_k$. Given their irregular distribution, the words were divided into two groups G1 ($50\% \leq P_k \leq 100\%$) and G2 ($0 \leq P_k \leq 50\%$) to avoid distortions in the mean. Thus, using linear interpolation, for G1, $Pk = \frac{(\mathrm{P} - 0.003) \times 50}{4.987} + 50$ and for G2, $Pk = \frac{\mathrm{P} \times 50}{0.003}$. The information quantity $I = log(1/P_u)$ and information entropy $h = I \times P_u$ were also measured for each word (Rosenfeld, 2000). Equivalence formulas had to be obtained as different letters could have similar entropies but generate words that have different degrees of familiarity. This implies that the amount of information and entropy could be the same in texts with different comprehension difficulty classifications. For example, both isolated letters such as c and m have similar informational entropy, thus both their amount of information and entropy can be the same for an easily comprehensible text or not. However, in a given text we can have more unusual words with the letter c than the letter m with different entropies. Thus, variables P_k and P_u needed to be formulated to establish entropy in semantic-communication terms. It was necessary to correct any misrepresentation of the words that were more likely to appear in texts at lower reading levels. Such inferences were confirmed using probabilistic averages for text quantification in a relatively broad database rather than focusing on isolated words.

CalcuLetra. All equations, words and their respective quantitative parameters were stored in the software suite *CalcuLetra*, which we have made openly available (https://github.com/LouiseBogea/CalcuLetra). From a text, it shows the number of letters, words, sentences and paragraphs, as well as letter or word frequency f and occurrence probability P. From each word it also shows the knowledge probability P_k and unknowledge probability P_u, information quantity I and entropy h. From each variable, it shows the minimum, maximum and average values. Thus, the average P_k of a text was calculated from the sum of each word's

P_k. The stored P_k values did not change when new text samples were entered. Words in the entered texts not stored in *CalcuLetra* are automatically excluded to avoid distortions in quantification analysis. We can import a.xls or.csv file to expand the *CalcuLetra* database or to reset it. Thus, we collected denotative texts samples in Portuguese ($n = 3330$) as test data and inserted to *CalcuLetra* to validate the quantitative model. Further information about the text sample sources is given in Supplementary Table S3.

Sample selection. Academic theses published during the last 10 years (2007–2017) were manually collected ($n = 1032$) from the Brazilian Digital Library of Theses and Dissertations and in the Digital Library of Theses and Dissertations at the University of São Paulo (USP). Their word usage statistics were fully quantified ($n = 33,101$) and used as training data. Further information about the repositories is shown in Supplementary Tables S1 and S2. Foreign and misspelled words or words not included in Portuguese dictionaries were manually removed as were figures, proper nouns, numbers and other symbols (e.g., () [] { }? + // = *<>:; ‘ % @! # - & |) to avoid any discrepancies in the statistical analyses. Words were categorized as either denotative or connotative based on their explicit definitions in dictionaries. We chose academic theses (denotative language) to avoid analysis errors caused by the figurative meaning of words.

Model application. Training words were stored in the information-processing software *CalcuLetra*, which was developed for this study using C#, Microsoft Visual Studio 2012, with SQL Server Management Studio 2012 being used to manage the database. SQL was used to filter out words that had 17 or more letters, those that occurred less than 4 times and those that had an occurrence probability of $P \leq 0.0001$ due to misspelling or mixed words. Duplicate words, prefixes and suffixes were also removed by their very low occurrence to avoid distortions in the mean (e.g., pós-, sub-, vice-, intra-, inter-, pro-, pre-, anti-, ante-, -mente, -issimo, -oso, -osa).

Validation and data analysis. BioEstat 5.3 was used to evaluate comparisons within the training database and between quantitative parameter analyses, with the emphasis placed on P_k and h variations (Ayres et al., 2007). A decision tree was used to establish the P_k and P_u ranges in relation to entropy (Breiman, 2017). This is a commonly used predictive/exploratory alternative for regression and classification problems, as decision trees can handle both categorical and numerical variables using both classification and regression. The decision trees algorithm assisted in predicting and classifying the data based on the values of the response variables. It also executed successive binary partitions to obtain subsets that are increasingly homogeneous concerning the response variable and placed the results in a hierarchical tree of decision rules for prediction or classification, with each data point partition being a node. There are internal nodes and terminal nodes, which are called leaves. The main steps in the algorithm were: (1) generating the root node (from the entire database); (2) determining the best way to divide the data into two groups (find the nodes to be divided); (3) choosing an attribute that best classified the data by considering the single paths between the root nodes and each subsequent node; and (4) creating and designing the nodes and the associated branches. The algorithm then returned to step 2.

If all cases in each terminal node show identical values, the impurity of the node will be minimal; the homogeneity will be maximal and the prediction will be perfect. One way to control division is to allow the division to continue until all terminal nodes are pure or contain no more than a specified minimum number of cases or objects. In most cases, the interpretation of summarized results in a relatively simple tree. Thus it is useful not only for rapid classification of new observations but can also often produce a much simpler model to explain why these observations are classified or predicted in a particular way.

Because decision trees are nonparametric, they do not assume an underlying distribution and can provide a relatively unbiased initial inspection of the training data, providing an image of the structure and thereby improving the interpretation of the results. Here, the quadratic relationships between entropy and P_k were modeled using a Gaussian regression model (Montgomery et al., 2006). Deviation measures were found, including the mean absolute deviation (MAD), the simple mean of the absolute differences between real and adjusted values; mean square deviation (MSD), the simple mean square of the differences between actual and adjusted values; and mean absolute percent error (MAPE), the percentage error between the actual and adjusted values.

As the test data was inserted in *CalcuLetra*, each text's readability was determined by each word's objective parameters from the stored database and their average P_k compared to quantify the comprehension difficulty. To identify the approximate points at which there was a change in the slope between the number of denotative texts inserted in *CalcuLetra* and the average knowledge probabilities, a binary segmentation method was used that identified the changes in the mean and variance of the knowledge probabilities as the number of texts grew (Scott and Knott, 2006). For each change point interval, the relationship between the knowledge probabilities and texts was described using linear regressions. For all analysis, including tables and figures, we used R (version 3.2.4) (R Core Team, 2013).

Results

Representative word groups. We considered the upper and lower bounds of the training database as the most relevant (Fig. 1), as they showed the words with a $P_k \geq 0.7$ (G1) that contributed to comprehension and those with a $P_u \geq 0.3$ (G2) which were likely to cause a communication failure.

In the dispersion chart that connects the data pairs, we found a positive and significant correlation ($r = 0.99$, p-value < 0.001) between the entropy and P_k, i.e., the greater the entropy, the greater the P_k and vice versa. Conversely, we found a significant negative correlation ($r = -0.99$, p-value < 0.001) between the entropy and P_u, i.e., the higher the entropy, the lower the P_u and vice versa.

In Fig. 2, we evaluated the cut-offs for P_k (a,c) and P_u (b,d), for which the entropy distribution was homogeneous by employing a machine learning regression tree technique, with the stipulation that the minimum number for the group division be 5000 to avoid an overly large tree, facilitating data interpretation. Classification and regression tree algorithms aim to achieve the best possible predictive accuracy. Thus, the division at each node was found and generated improvement in forecast accuracy by finding sub-groups in which the groups are homogeneous (intra-group) and heterogeneous (inter-group). Operationally, the most accurate forecast was defined as the one with the lowest rate of classification errors.

As Fig. 2 shows, from the cut-off points identified by the regression tree (panels 2a, b), entropy was described in the bands for knowledge probability (Table 1) and for the probability of being unknown (Table 2). Also, descriptive of each variable is represented in boxplots (panels 2c, d). Further information about boundaries in boxplots is shown in Supplementary Table S4.

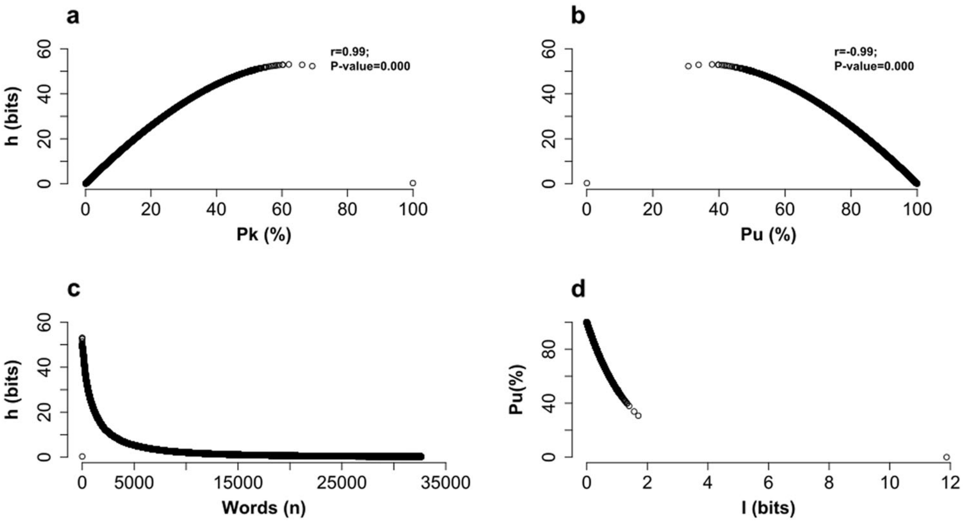

Fig. 1 Distributions of the probabilities, information, and entropy for each word stored in CalcuLetra. **a**, **b** Harmonic distribution of entropy, except for deviations in G1 ($P_k \geq 0.7$), which predominantly included conjunctions, prepositions, and articles. This is due to a lack of data in this group; inverse results for **b**. The distributions had two atypical points: there was an observation that presented a high value for P_k (99.97) and a low value for h (0.31), and there was an observation with a high value for P_u (100.00) and a null h (0.00). **c** Around 17.33% of the texts had entropy measures greater than 10, and these were also the texts with the lower word counts. For instance, entropy equaled 10 when the number of words was 3259, and the entropy equaled 5 when the number of words was 5724. A four-word count was associated with maximum entropy (53.05). **d** Monotonic decreasing and non-linear relationships between the variables. When P_u was maximized, I was minimized

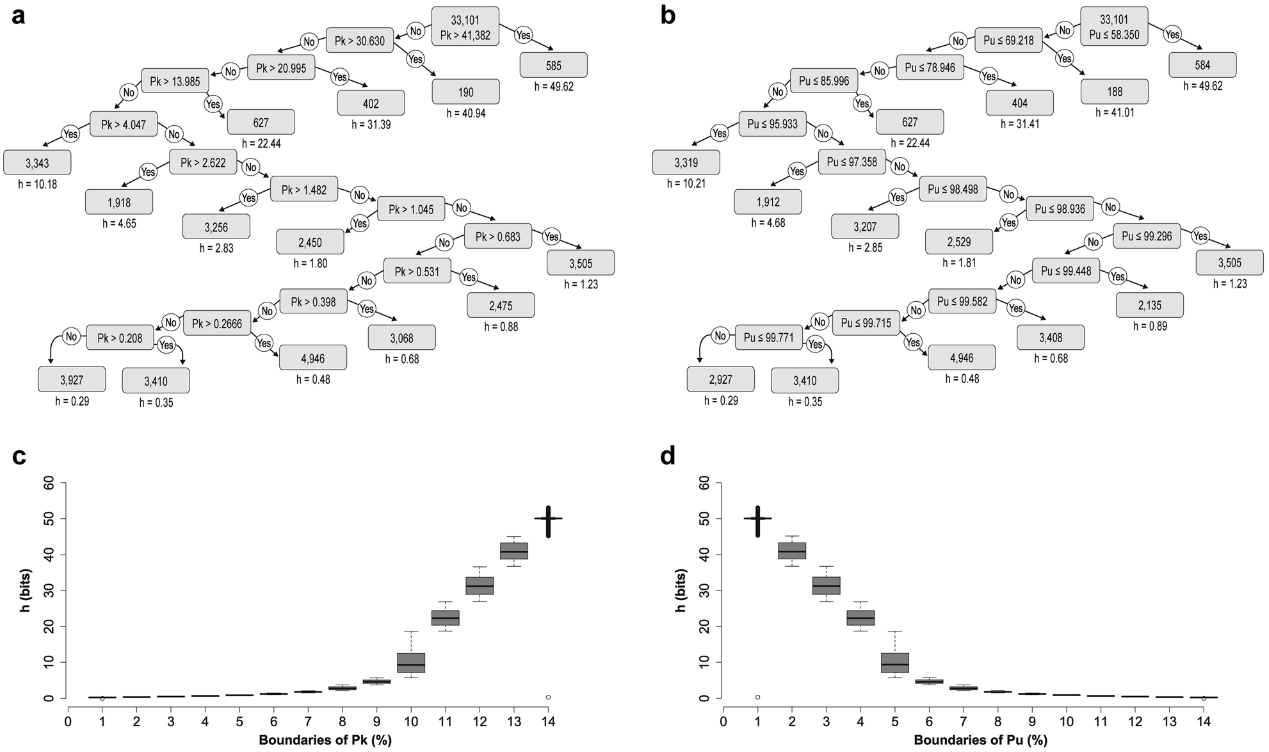

Fig. 2 Regression tree (**a**, **b**) for entropy as a response variable and knowledge probability as an explanatory variable, also probability of being unknown as an explanatory variable, and descriptive of variables in boxplots (**c**, **d**). In each sheet, entropy is represented as follows: h = leaf mean. Such tree method results can be summarized in a series of logical if-then conditions, unlike the usual regression models that define a prediction equation for the relationship between variables. In addition, there is no implicit assumption that the relationships between the predictor variable and the dependent variable are linear

Quadratic regression. To describe the relationship between the entropy and the knowledge probability, we used a quadratic regression model to determine the knowledge probability value that maximized entropy, with the knowledge probability variable being centered on the average. The adjusted model was found to have a variance inflation factor of 6.24, indicating the absence of multicollinearity, and thus was considered to be well adjusted (Fox, 2008). Therefore, the model equations—Eq. (1) for entropy

Table 1 Description of entropy by knowledge probability

Pk range	Entropy (*h*)								
	N	Avg.	SD	CI-95%	Min.	1° Q	2° Q	3° Q	Max.
(41.3823; 99.9730]	585	49.615	2.419	[49.377; 49.780]	0.315	50.008	50.057	50.179	53.049
(30.6300; 41.3823]	189	40.941	2.469	[40.599; 41.270]	36.739	38.780	40.784	43.269	45.030
(20.9950; 30.6300]	402	31.388	2.797	[31.124; 31.655]	26.925	28.924	31.221	33.705	36.601
(13.9980; 20.9950]	627	22.436	2.335	[22.254; 22.615]	18.717	20.366	22.318	24.387	26.860
(4.0466; 13.9980]	3343	10.183	3.541	[10.070; 10.302]	5.746	7.143	9.319	12.514	18.694
(2.2616;4.0466]	1918	4.652	0.565	[4.627; 4.676]	3.760	4.160	4.609	5.113	5.718
(1.4820; 2.6116]	3256	2.829	0.465	[2.813; 2.845]	2.150	2.420	2.769	3.226	3.732
(1.0450. 1.4820]	2450	1.799	0.178	[1.793; 1.807]	1.526	1.635	1.797	1.961	2.122
(0.6833; 1.0450]	3505	1.228	0.147	[1.224; 1.233]	1.011	1.092	1.228	1.364	1.500
(0.5316; 0.6833]	2475	0.882	0.062	[0.879; 0.884]	0.794	0.820	0.875	0.930	0.982
(0.3983; 0.5316]	3068	0.677	0.055	[0.675; 0.679]	0.602	0.629	0.684	0.739	0.765
(0.2666; 0.3983]	4946	0.484	0.054	[0.482; 0.485]	0.411	0.437	0.492	0.521	0.574
(0.2083; 0.2666]	3410	0.354	0.023	[0.353; 0.355]	0.329	0.329	0.355	0.384	0.384
[0.000; 0.20833]	2927	0.286	0.014	[0.286; 0.287]	0.000	0.274	0.274	0.300	0.300

We found that in the lower P_k range ([0.000; 0.20833]), the average *h* was 0.286 [0.286; 0.287], and the minimum and maximum values were 0.000 and 0.200; in the highest P_k range ([41.3823; 99.9730]), the average *h* was 49.615 [49.377; 49.780], and the minimum and maximum values were 0.315 and 53.049. The decision tree was unable to differentiate the atypical points due to a lack of data with high knowledge probability. Confidence Interval is the Bootstrap for the average

h and knowledge probability P_k and Eq. (2) for unknown probability P_u—were defined as follows:

$$\begin{aligned}\overline{Pk} &= Pk - 3.210436\\ \widehat{h} &= 4.6457866 + 1.4350649^{*}\overline{Pu} - 0.0101171^{*}\overline{Pu}^{2}\end{aligned} \quad (1)$$

$$\begin{aligned}\overline{Pu} &= Pu - 96.78956\\ \widehat{h} &= 4.6457866 - 1.4350649^{*}\overline{Pu} - 0.0101171^{*}\overline{Pu}^{2}\end{aligned} \quad (2)$$

In this sense, when P_k increased by 10 units, *h* had an average increase of approximately 13.92 units. This function allowed for the maximum points to be identified and the entropy to be estimated from the knowledge probability. It was found that the position that maximized entropy was a knowledge probability of 70.92% ($P_u = 30.74\%$); as after this value, as P_k or P_u increased, *h* tended to decrease Table 3.

Figure 3 shows the comparison of actual and adjusted values for P_k (a) and P_u (b). Note that after the inflection point, the adjusted curve did not exactly follow the actual curve because of the small number of points examined in this range. This was the main limitation of the training database; however, despite the missing values in G1 ($P_k \geq 0.7$), it was possible to infer that *h* tended to decrease for familiar words.

We found MAD and MSD estimates to be low, suggesting proximity between the adjusted values and the actual data. Also, the MAPE estimate was that the forecast was on average 3.009% incorrect, which is satisfactory for terms of fit quality.

Quantification of the texts in CalcuLetra. Test data (e.g., articles, essays, abstracts, and news) were entered in *CalcuLetra* ($n = 3330$) in order to extract the quantitative word values based on the stored objective parameters in the training database (Fig. 4). This process inferred the difficulty of semantic comprehension by the average P_k values. Texts were already organized syntactically in standard Portuguese. We could set $P_k \leq 0.3$ for texts with a higher chance of causing incomprehension.

We used a binary segmentation method to identify the changes in the mean and variance of P_k as the number of texts grew (Scott and Knott, 2006). Similar to decision trees, this method searches for the values that make the groups more homogeneous in relation to the mean and variance internally and are heterogeneous between the groups. In this way, we identified points 24, 551, 2724, 3152, 3295, and 3330. Note that within each change point range, the relationship between P_k and the texts was strictly linear. Therefore, the six linear regressions were adjusted according to the time intervals identified in the change points. When the number of texts increased by 10 units in the interval [1; 24], P_k increased by 3.00 units on average; in the interval (25; 551), P_k increased by 0.03 units on average; in the interval (552; 2724), P_k increased by 0.10 units on average; in the interval (2725; 3152), P_k increased by 0.03 units on average; in the interval (3153; 3295), P_k increased by 0.30 units on average; in the interval (3296; 3330), P_k increased by 3.00 units on average. In all models, the R^2 value was perfect as one variable thoroughly explains the variability in the other.

Discussion

Textual complexity analyses often include several domains (Crossley et al., 2017). In this paper, comprehension difficulty was quantified in order to infer an objective measure of the complexity of denotative written material in Portuguese using texts related to educational language standards. Text comprehension was estimated based on the correlations between the semantic representations of the words and their respective probabilities to ensure that the textual complexity was related to the semantic entropy of the words and their familiarity degree. To the best of our knowledge, this was the first investigation on Portuguese textual comprehension difficulties using quantitative methods and objective parameters. The use of a quantitative model that incorporates comprehension in terms of word complexity measures may help overcome many of the problems associated with the reliance on qualitative standards and subjective evaluations (Gathercole and Alloway, 2006). Importantly, entropy analysis in our model revealed that the relationship between word comprehensibility and probability is not linear (Smith, 2012). This means that word features, such as a word's relationship to the sentence, is non-linearly related to textual features, such as occurrence. The model presented here can elucidate that relationship through the use of regression analysis (Gastón and García-Viñas, 2011).

We chose academic texts due to their potentially more comprehensive vocabulary and lower use of figurative language. We acknowledge that using a corpus of academic words may have biased our sampling, as it is possible that university students intentionally use unusual words to demonstrate a more elaborate vocabulary. There is a possibility that we did not accurately

Table 2 Description of entropy by the probability of being unknown

Pu range	Entropy (*h*)								
	N	Avg	SD	CI-95%	Min	1° Q	2° Q	3° Q	Max
[0.0265; 58.3500]	584	49.622	2.414	[49.407; 49.783]	0.315	50.008	50.057	50.180	53.049
(58.3500; 69.2183]	188	41.009	2.456	[40.638; 41.354]	36.756	38.813	40.830	43.305	45.171
(69.21833; 78.9466]	404	31.414	2.815	[31.151; 31.690]	26.925	28.924	31.230	33.742	36.739
(78.9466; 85.9966]	627	22.436	2.335	[22.262; 22.617]	18.717	20.366	22.318	24.387	26.860
(85.9966; 95.9333]	3319	10.215	3.533	[10.102; 10.331]	5.771	7.183	9.371	12.564	18.694
(95.9333; 97.3583]	1912	4.679	0.568	[4.655; 4.704]	3.786	4.186	4.637	5.166	5.746
(97.3583; 98.4983]	3207	2.854	0.464	[2.838; 2.870]	2.176	2.446	2.795	3.252	3.760
(98.4983; 98.9366]	2529	1.810	0.186	[1.803; 1.817]	1.526	1.661	1.797	1.961	2.150
(98.9366; 99.2966]	3505	1.228	0.147	[1.223; 1.233]	1.011	1.092	1.228	1.364	1.500
(99.2966; 99.4483]	2135	0.896	0.055	[0.893; 0.898]	0.820	0.846	0.901	0.930	0.982
(99.4483; 99.5816]	3408	0.689	0.062	[0.687; 0.691]	0.602	0.629	0.684	0.739	0.794
(99.5816; 99.7150]	4946	0.484	0.054	[0.482; 0.485]	0.411	0.437	0.492	0.521	0.574
(99.7150; 99.7716]	3410	0.354	0.023	[0.353; 0.355]	0.329	0.329	0.355	0.384	0.384
(99.7716; 100.00]	2927	0.286	0.014	[0.286; 0.287]	0.000	0.274	0.274	0.300	0.300

The lower range for P_u ([0.0265; 58.3500]) had an average *h* of 49.622 [49.407; 49.783], of which the minimum and maximum values were 0.315 and 53.049. The decision tree was unable to differentiate the atypical points due to a lack of data with low knowledge probability. In the highest range for P_u (99.7716; 100.00), the average *h* was 0.286 [0.286; 0.287], and the minimum and maximum values were 0.000 and 0.300. Confidence Interval Bootstrap for the average

Table 3 Bounds of the training database

ID	Word	F	P (%)	Pk (%)	Pu (%)	I (bits)	h (bits)
G1							
107504	Com	240327	2.739839	97.843672	2.156328	5.535279	0.119358
110952	Para	232542	2.651089	94.674283	5.325717	4.230880	0.225324
118876	Que	227632	2.595119	92.675482	7.324518	3.771122	0.276216
G2							
112540	Percolação	40	0.000456	0.760000	99.240000	0.011005	0.010921
227	Alvarenga	64	0.000730	1.216667	98.783333	0.017660	0.017445
159	Albumina	289	0.003295	5.491667	94.508333	0.081487	0.077012

$P_k \geq 0.7$: Group of words predominated by connective elements for textual cohesion (e.g., conjunctions and prepositions); $P_u \geq 0.3$: Group of words predominated by elements that were semantically meaningful for textual coherence (e.g., nouns and adjectives)

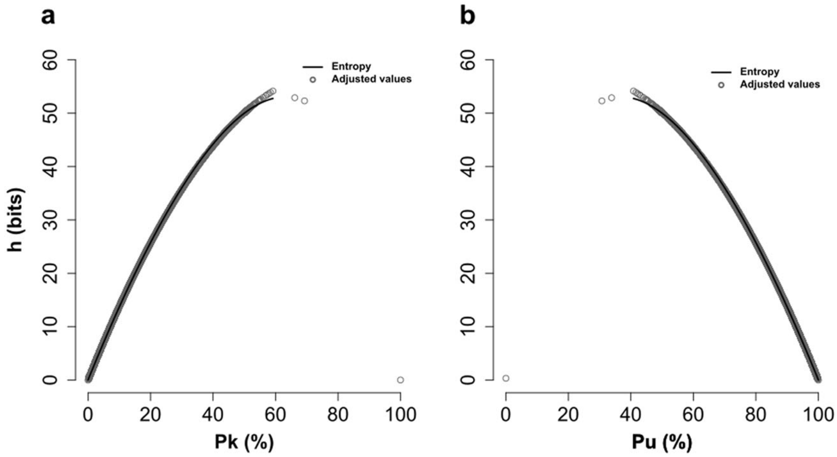

Fig. 3 Adjusted regression model: knowledge probability **a** and probability of being unknown **b**. The adjusted values by the regression analysis show that the model can accurately reproduce the actual entropy values. The adjusted model shaped the data characteristics up to a 60% knowledge probability. The lack of adjustment for the higher values of P_k reflected the lack of samples with probabilities higher than 60% and the presence of atypical observations ($P_k = 99.97$ and $h = 0.31$)

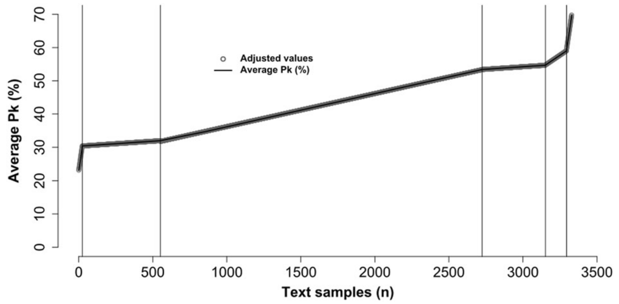

Fig. 4 Random register of denotative texts in CalcuLetra and the average P_k correlations. The average P_k for the test data was $0.3 \leq P_k \leq 0.5$, and texts with an average $P_k \leq 0.3$ were unlikely to be understood by most neurotypical readers. The statistical significance level was set at $P_k \geq 0.23$. The total range for the average P_k values was from 0.23 to 0.69, and the overall average P_k was 0.42. In the comparison of actual and adjusted P_k values, the mean percentage error was very low, indicating almost perfect prediction. MAD (0.024) and MSD (0.002) estimates were low, suggesting that the adjusted values were in close proximity to the real data. Also, the MAPE estimate indicated that the average forecast was incorrect by 0.064%

sample more common words that would exhibit higher degrees of familiarity. However, we highlight that one can use any corpus (including from other languages) as training data for *CalcuLetra* to quantify textual comprehension difficulty. Corpus selection by itself does not directly impact the validity of our quantitative model, and future studies may systematically address how the training dataset can affect comprehension estimates.

Our results found that textual labor was quantifiable by a low degree of familiarity, a low amount of information, and low entropy, which together predicted a high incomprehension probability (summarized in Fig. 5). In a study of word frequency in books, the amount of information in 300 to 3000 unpredictable words was reported to be 0.2 bits per word, which was congruent with the findings in our dataset (Marcelo and Damian, 2010). Our results are also consistent with several available Portuguese word-frequency lists (Ferreira, 2014, Wiktionary, 2017, Mark, 2011, Github, 2017).

Previous studies have shown that complex aspects of information organized in symbolic sequences can be quantified (Montemurro and Zanette, 2002, Montemurro and Zanette, 2011, Debowski, 2011, Kalimeri et al., 2015). In this work, we ranked semantic decoding difficulties using established equivalence formulas. Differently from other research that sought to quantify the entropy of word order in sentences, the results of which showed that there was constant relative entropy despite vocabulary diversity, this study set the average ratio for the Portuguese denotative text comprehension (Montemurro, 2014). In using an average ratio for text comprehension, deviations due to vocabulary diversity can be overcome. While this may suggest that the parameters may only function and possibly depend on the restricted description of a language, this can be altered to suit any specified vernacular. In this sense, the findings of the research are

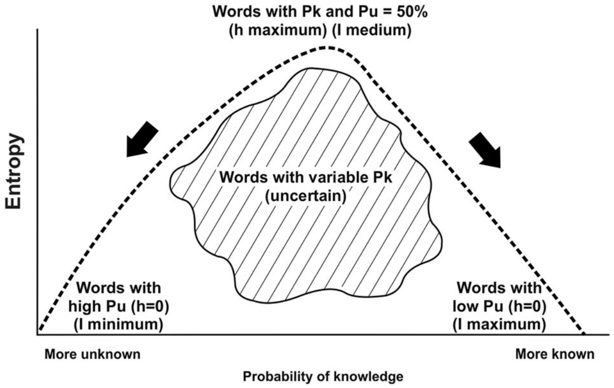

Fig. 5 Overview of the database parameters. Lower and upper bounds of the training database were delimited by low-entropy words with the highest probabilities of causing incomprehension (i.e., nouns and adjectives) or facilitating semantic decoding (i.e., prepositions and conjunctions). Despite the missing values in G1($P_k \geq 0.7$), the score resulted from eliminated words at the bounds. This avoided distortions in the mean, while the linear interpolation made in this range could be deduced based on the harmonic distribution that was established in other bands

not restricted to Portuguese comprehension, and parameters can be applied to any linguistic modality that incorporates a defined semantic code (language) (Zwaan, 2016). This generalizability is because, despite the known differences in structure and vocabulary of any given language system, the impact of word order within the structure of the language is thought to be a statistical linguistic universal (Fox, 2008). In deviating from reliance on word ordering and concentrating instead upon a quantified flow of informatics sequencing, word usage and cognitive forms such as repetition and rehearsal are acknowledged in the function of comprehension (Zwaan, 2016). As described, we can speculate on several applications for our software suite. For example, an objective quantification of textual comprehensibility could be used to the development of diagnostic and therapeutic tools for learning disorders such as dyslexia (Harley and O'Mara, 2016). Thus, the use of a quantitative tool that identifies words in relation to their embedded discourse might provide a valuable objective guideline for the quantitative evaluation of learning disorder severity, considering current therapeutic practices are based mainly on qualitative measures (Kirkby et al., 2011). In addition, by using the regression analysis technique employed in this study, it may be possible to estimate the optimal rate of acquiring a new word from textual representations concerning its occurrence throughout different texts, which could be used as benchmarks for clinical studies and diagnostics (Oliveira and Gomes, 2010).

Our software might also be used in standard education. In the latest edition of the Programme for International Student Assessment (2015), Brazil ranked 59th in reading out of 70 countries, revealing that many students have insufficient comprehension skill. The National Indicator of Functional Literacy (INAF) reports that 68% of the 30.6 million Brazilians between 15 and 64 years old who have studied up to 4 years, and 75% of the 31.1 million who have studied up to 8 years, remain at basic literacy levels. The use of inadequate teaching material and improper screening for learning disabilities are likely contributors to these failures in public education outcomes. Specifically, the annual National Textbook Program (PNLD) has been heavily criticized, especially regarding the creation and selection of textbooks (Di Giorgi et al. 2014). Using *CalcuLetra*, teachers might perhaps select more appropriate course material based on quantitative entropy parameters (Cidrim and Madeiro, 2017). Of course, which entropy levels might be best suited for specific grades or educational goals would have to be empirically determined. *CalcuLetra* could also be applied for the creation of textbooks, essay materials, and the evaluation of automated textual comprehension and student production. It could also be used to quantify the functional acquisition of specific vocabularies using different training datasets. While our model was not yet been fitted nor compared to the comprehension of actual human readers, it is plausible that it could be used to monitor the development of a given student's textual comprehension skill by analyzing the production of the student with progressively complex training datasets inserted in *CalcuLetra*. Understanding trends and differences in textual representation can also add to ongoing theoretical debates about the role of automated textual analysis in education. Indeed, automated textual analyses may contribute to exploring intersectional perspectives on social groups studying in Portuguese, such as specific ethnic groups and individuals with learning disabilities.

Conclusions

By applying information theory principles, we developed a new textual analysis software suite (*CalcuLetra*) and demonstrated that the comprehension of Portuguese denotative texts could be quantified using objective parameters. Our results revealed groups of words with low entropy that either cause incomprehension or facilitate semantic decoding. We propose that the methodology and software developed may eventually be used as an auxiliary evaluation tool for teaching materials and textual comprehension assessments, as well as for the study and therapy of learning disorders. This model may also be adapted to other languages to evaluate the difficulty of semantic decoding and semantic complexity.

Data availability

The datasets generated and/or analyzed during the current study, and the relevant code, are available on https://github.com/LouiseBogea/CalcuLetra.

Received: 3 February 2019 Accepted: 12 August 2019
Published online: 03 September 2019

References

Ayres M, Ayres Jr M, Ayres DL, Santos AAS (2007) Bioestat 5.0 aplicações estatísticas nas áreas das ciências biológicas e médicas. IDSM, Belém

Bao J, Basu P, Dean M, Partridge C, Swami A, Leland W, Hendler JA (2011) Towards a theory of semantic communication. 2011 IEEE Netw Sci Workshop 1:110–117

Bar-Hillel Y, Carnap R (1952) An outline of a theory of semantic information. Res Lab Electron Tech Rep 247:221–274

Breiman L (2017) Classification and regression trees. Routledge, Abingdon

Cidrim L, Madeiro F (2017) Information and Communication Technology (ICT) applied to dyslexia: literature review. Rev CEFAC 19(1):99–108

Crossley SA, Skalicky S, Dascalu M, McNamara D, Kyle K (2017) Predicting text comprehension, processing, and familiarity in adult readers: New approaches to readability formulas. Discourse Process 54:340–359

D'Alfonso S (2011) On quantifying semantic information. Information 2:61–101

Debowski T (2011) On the vocabulary of grammar-based codes and the logical consistency of texts. IEEE Trans Inf Theory 57:4589–4599

Di Giorgi C, Militão SCN, Militão NA, Perboni F, Ramos RC, Lima VMM (2014) Uma proposta de aperfeiçoamento do PNLD como política pública: o livro didático como capital cultural do aluno/família. Ens Aval Pol Públ Educ 22 (85):1027–1056

Ferreira ABH (2014) Dicionário Aurélio. Editora Positivo, Curitiba

Fox J (2008) Applied Regression Analysis and Generalized Linear Models. Sage, Thousand Oaks, California

Gasperin C, Specia L, Pereira T, Aluísio S (2009) Learning when to simplify sentences for natural text simplification. Proc ENIA 1:809–818

Gastón A, García-Viñas JI (2011) Modelling species distributions with penalised logistic regressions: a comparison with maximum entropy models. Ecol Model 222(13):2037–2041

Gathercole SE, Alloway TP (2006) Practitioner review: Short-term and working memory impairments in neurodevelopmental disorders: diagnosis and remedial support. J Child Psychol Psychiatry 47:4–15

Github (2017) Frequency Words Hermit D. https://github.com/hermitdave/FrequencyWords/blob/master/content/2016/pt_br/pt_br_50k.txt. Accessed 20 Mar 2017

Harley TA, O'Mara DA (2016) Hyphenation can improve reading in acquired phonological dyslexia. Aphasiology 20(8):744–761

Hartmann N, Avanço L, Balage P, Magali D, Nunes MGV, Pardo T, Aluísio S (2014) A large corpus of product reviews in Portuguese: tackling out-of-vocabulary words. In: Ninth International Conference on Language Resources and Evaluation. European Language Resources Association (ELRA), Reykjavik, Iceland, pp 3865–3871

Kalimeri M, Constantoudis V, Papadimitriou C, Karamanos K, Diakonos FK, Papageorgiou H (2015) Word-length entropies and correlations of natural language written texts. J Quant Linguist 22:101–118

Kendeou P, Van Den Broek P, Helder A, Karlsson JA (2014) Cognitive view of reading comprehension: Implications for reading difficulties. Learn Disabil Res Pr 29:10–16

Kirkby JA, Blythe HI, Drieghe D, Liversedge SP (2011) Reading text increases binocular disparity in dyslexic children. PLoS ONE 6(11):e27105

Marcelo AM, Damian HZ (2010) Towards the quantification of the semantic information encoded in written language. Adv Compl Sys 13:135–153

Mark D (2011) A frequency dictionary of Portuguese. Routledge, London

Montemurro MA (2014) Quantifying the information in the long-range order of words: semantic structures and universal linguistic constraints. Cortex 55:5–16

Montemurro MA, Zanette DH (2002) Entropic analysis of the role words in literary texts. Adv Compl Sys 5:7–17

Montemurro MA, Zanette DH (2011) Universal entropy of word ordering across linguistic families. Plos ONE 6:e19875

Montemurro MA, Zanette DH (2016) Complexity and universality in the long-range order of words. Creat Univers Lang ArXiv abs 1503(1129):27–41

Montgomery D, Peck A, Viving G (2006) Introduction to linear regression analysis. John Wiley, New York

Nowak MA, Plotkin JB, Jansen VA (2000) The evolution of syntactic communication. Nature 404:495–498

Oliveira HG, Gomes P (2010) PT: automatic construction of a lexical ontology for Portuguese. In: Proceedings of 5th European Starting AI Researcher Symposium. Lisbon, Portugal, pp 199–211

R Core Team (2013) R: a language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria

Rabêlo LGN, Moraes RM (2008) Entropy and generation of approximation series using a JAVA tool. In: XXVI Brazilian Symposium on Telecommunications (SBrT). Brazilian Telecommunications Society, Rio de Janeiro, p 1–6

Rosenfeld R (2000) Two decades of statistical language modeling: Where do we go from here? Proc IEEE 88(8):1270–1278

Scott AJ, Knott MA (2006) Cluster analyses method for grouping means in the analysis variance. Biometrics 30:507–512

Shannon CE (1948) A mathematical theory of communication. Bell Syst Tech J 27:379–423

Sigman M, Peña M, Goldin AP, Ribeiro S (2014) Neuroscience and education: prime time to build the bridge. Nat Neurosci 17:497–502

Smith RD (2012) Distinct word length frequencies: distributions and symbol entropies. Glottometrics 23:7–22

Soares AP, Costa AS, Machado J, Comesana M, Oliveira HM (2017) The Minho Word Pool: norms for imageability, concreteness, and subjective frequency for 3,800 Portuguese words. Behav Res Meth 49:1065–1081

Thorne C, Szymanik J (2015) Semantic complexity of quantifiers and their distribution in corpora. In: Proceeding of the International Conference on Computational Semantics. International Wood Culture Society, London 64–69

Wiktionary (2017) Wordlist. En.wiktionary. https://en.wiktionary.org/wiki/Wiktionary:Frequency_lists/BrazilianPortuguese_wordlist. Accessed 20 Mar 2017

Zipf GK (1935) The psychobiology of language. Houghton-Mifflin, Oxford, England

Zwaan RA (2016) Situation models, mental simulations, and abstract concepts in discourse comprehension. Psychon Bul l Rev 23(4):1028–1034

Acknowledgements

This study has been funded by the Federal University of Pará (UFPA) through the Pro-rectory of Research (PROPESP). MSF is a research fellow of the National Council for Research and Development (CNPQ). The funding providers had no role in the study design, data collection, and analysis, decision to publish, or preparation of the paper. Funding provider's websites: http://www.ufpa.br/, http://www.cnpq.br/.

Author contributions

LBR, ARR, KMC, and MSF designed the research. LBR developed the new algorithm, wrote all the code, conducted the research, analyzed the data, interpreted the results, and wrote the first draft of the paper. ARR, KMC, and MSF supervised the study and contributed to the final version of the paper.

Additional information

The online version of this article (https://doi.org/10.1057/s41599-019-0311-0) contains supplementary material, which is available to authorized users.

Competing interests: The authors declare no competing interests.

Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.