

USING VISUAL TEXT MINING TO UNCOVER RELEVANT ISSUES IN THE BRAZILIAN EDUCATION SYSTEM

Leila Abuabara

Universidade Federal de São Paulo (UNIFESP) & Instituto Tecnológico de Aeronáutica (ITA)
São José dos Campos-SP, Brasil
leila.abuabara@unifesp.br

Alberto Paucar-Cáceres

Manchester Metropolitan University (MMU)
All Saints Campus, Manchester, United Kingdom of Great Britain and Northern Ireland
A.Paucar@mmu.ac.uk

RESUMO

A comunicação de mídia gera continuamente conteúdo sobre questões de interesse social. Por vezes, são informações de alta qualidade e em vasto volume. Neste estudo, exploramos um conjunto de entrevistas organizadas por uma tv pública brasileira ocorridas no período pré-eleitoral. Estas entrevistas envolveram especialistas em Educação no Brasil discutindo uma variedade de problemas impulsionados pelo baixo desempenho na avaliação escolar PISA 2018. O objetivo é construir um *corpus* referente a estas entrevistas e explorá-lo a fim de identificar os temas mais relevantes aí tratados e que merecem atenção. Para isso, usamos ferramentas de *mineração visual de texto* (VTM) com auxílio do *software* Iramuteq para análise lexicográfica básica e multivariada. Dentre os resultados, destacamos as categorias de temas que emergiram a partir dos *contextos linguísticos*: (i) a importância da leitura, (ii) cuidados na primeira infância, (iii) práticas pedagógicas, (iv) dinâmica do ensino básico, e (v) políticas públicas para o ensino superior.

PALAVRAS-CHAVE. mineração visual de texto, Iramuteq®, dados de mídia, Educação, Brasil

Tópicos: AdP&ED (PO na Administração Pública e Educação), BDA (Big Data e Analytics)

ABSTRACT

Media communication is continuously generating content of issues of social interest. Sometimes it is high quality information and in vast volume. In this study, we explore a set of interviews organized by a Brazilian public television taking place in the pre-election period. These interviews involved experts in Education in Brazil discussing a variety of issues driven by low performance in the Program for International Student Assessment in 2018. The aim is to build a *corpus* referring to these interviews and explore it to identify the most relevant themes addressed therein and that deserve attention. For this, we use *visual text mining* with support of Iramuteq software for basic and multivariate lexicographical analysis. Among the findings, we highlight the categories of themes that emerged from *linguistic contexts*: (i) the importance of reading, (ii) early childhood care, (iii) pedagogical practices, (iv) dynamics of basic education, and (v) public policies for higher education.

KEYWORDS. Visual text mining, Iramuteq®, media data, Education, Brazil

Paper topics: AdP&ED (OR in Public Administration and Education), BDA (Big Data and Analytics)

1. Introduction

Visualization is a research method [Eppler e Burkhard 2008] that also has its relevance within Operational Research (OR) in several contexts since graphic representation can accommodate both *quantitative* and *qualitative* information. This latter includes arguments, ideas and opinions which can be easily found in the media. For use of analytical visualization tools, large amount of textual data is required, which has been abundantly and continuously being produced by the media communication through interviews, newspaper articles, documentaries, amongst others. We can generally find high quality content of issues of social interest such as sustainability, economy, policy, health & wellness, and so on. In this study, we explore a set of interviews organized by a Brazilian public television and which involved experts in the field of Education in Brazil discussing a variety of issues around this topic. This discussion was resulted from the Program for International Student Assessment (PISA) 2018 Report that revealed that 15-year-old Brazilian students had low proficiency in all disciplines compared to the other 78 countries that also participated in the assessment [INEP, 2020]. These interviews took place during the Brazilian presidential and State governors' pre-electoral campaigns in 2018 aiming to awake candidate's interest to this central issue and possibly for them to consider these topics in their campaign proposals. Thus, we present the research question addressed in this article:

RQ: *What are the most relevant themes discussed by education experts that could be part of future government projects?*

Methodologically, from the set of interviews, we carefully assembled a related *corpus*. To support the analytical investigation, we use visual text mining tools available at Iramuteq® software. Text mining process can extract instigating and non-trivial patterns or knowledge discovery from unstructured textual documents [Tan 1999]. The aim of this research is to explore this *corpus* and understand the most relevant themes addressed therein. Thus, we summarize this set of interviews into few issues to be addressed by governments in achievable future projects. In addition to the candidates for public office, the answer to the research question above shall be also of interest to education managers in the public and private spheres or any individual inserted in the field of education. And, why not say, the whole of society shall be interested in educational public policies.

The research presented in this article is part of a wider project in which, in a sequential phase, we intend to evaluate the content of these interviews through Soft Operational Research approaches [Rosenhead and Mingers 2001] within a multimethodological framework [Mingers and Brocklesby 1997] still to be proposed. Therefore, the contribution of this research, albeit partial, is to explore media data as a prelude to a more focused application of a Soft OR methodology.

Beyond this introduction which contextualizes the research problem, the article is organized as follows. Section 2 highlights the literature review related to visualization tools within OR field illustrating with some published applications. We also explore text mining approach and a key concept of this research that is *language*. In section 3, we detail the material used in the research. In section 4, we present the methodology that is divided in two main stages: (a) corpus organization, and (b) visual text mining application. We present and discuss very interesting findings in section 5. Finally, we conclude our research in section 6 presenting some limitations and future avenues to this study.

2. Literature Review

Diagrammatic representation of information is a convenient way to summarize information and support in generating hypotheses, recognizing trends, and understanding patterns. The practice of visualization tools can occur at different stages of a research and contribute to its direction [Basole *et al.* 2022]. In OR, traditional and static data visualization methods have been widely present particularly with *quantitative* data in many ways such as bar charts, pie charts, line charts, histograms, scatter plots and so on. *Qualitative* data also have their place in mental models,

flowcharts, word clouds, social network, tree maps and on. More recently, data science brought us new interactive and dynamic techniques able to work with huge volumes of data [Ertug *et al.* 2018]. The process of extracting relevant information from databases is called *knowledge discovery in databases* (KDD) [Quoniam *et al.* 2001].

The method selected in this research is *visual text mining* (VTM) which has been recently becoming very usual in digital humanities specially providing a “distant reading” of the situation [Moretti 2013]. This is because textual data mining is a means of reduction input data dimensionality, beyond being a quick way to have an overview of the data. Although this benefit, the outcome provided from this approach generally needs human intervention specially for qualitative interpretation and validation.

In *text mining* (also referred as *automated text analysis* or *text data mining* or *text analytics*), a set of documents (*corpus*) is considered as unstructured database that is normalized and become structured data suitable for analysis. In this context, values are obtained from the frequency of words and terms, amongst other statistical outcomes. Furthermore, texts are considered unstructured data (a bag of words, or a vector of word) and from that the frequency of words shall be counted and relations of co-occurrence shall be identified to generate clusters and then providing an idea of the *linguistic context* based on a dictionary.

An important concept in this research is (human) *language* and its interpretation from the perspective of text mining, which differs from other approaches such as in Soft OR, for instance. Analytical work using language (more written or transcribed than spoken) is of interest in various contexts and has been receiving increasing attention [Jackson and Moulinier 2002]. The approach applied in this research is known as *unsupervised learning* in which the general question is “*what is going on?*” with a particular subject (in our case, Education). Thus, an unlabeled data set is subject to analysis and clustering and so hidden patterns can be recognized without the need of a human intervention (so “*unsupervised*”) [Delua 2021]. However, features of language such as synonymies (different words with the same meaning) and polysemy (a same word with different meanings) shall receive special attention since they are not easily identified. In addition, elements linked to human cognition and a specific context or even metaphor are not easily understood in text mining. Analysts need to contextualize the outcomes in order to identify and interpret the insights.

3. Research Material

The *research material* is composed of a set of 37 interviews (data raw) with experts in education in many fields dated September and October 2018, the Brazilian presidential and governors’ pre-electoral period, and were organized by public television broadcaster. Among the interviewed are senior professors, pedagogues, members of educational councils, the Ministry of Education and Culture (MEC), journalists and sociologists specialized in Education, and educational consultants. The complete playlist can be found at: https://tvcultura.com.br/playlists/221_de-olho-na-educacao-de-olho-na-educacao.html. The context included discussions on current problems in the Brazilian education system and educational projects for the future governments to be considered in their campaign proposal.

Each interview considered a central theme to be debated such as “*How is education in Brazil nowadays?*”; “*Is private education better than public education?*”; “*Can any course be done remotely?*”; “*Higher education incentive programs*”, “*What is the main problem of public higher education?*”, “*What is the solution to the lack of vacancies in the public daycare centers?*” and so on. All interviews were conducted by the same interviewer, lasted around 20 minutes, and followed a structure organized as follows:

Start: composed by the introduction to the program, to the central theme, and the interviewee.

Intermediary information: composed by two main non-sequential moments. One where the interviewer makes questions, and the expert answers them. Another where a reportage including statistical evidence, some projects related to the theme

and small testimonials from other specialists, teachers, students, and parents are presented.

Closure: farewell including a quote for reflection from a famous or relevant person in society related to education.

Secondary analysis: data from the media

Interviews, debates, magazine articles and other publicly available media materials are a rich and inexhaustible source of data and information on various issues of concern to societies nowadays. They are happening all the time and generally deal with political, historical, economic, educational issues, amongst many others. This sometimes-high-quality material counts on the contribution from members of the industry, consultants, experts, and governments involved in those issues.

This sort of data is generally produced by professionals, in particular journalists specialized in these areas. The non-direct interaction with the interviewees may avoid any bias due to censorship, sentiment, contamination by interference, influence, or conflict of values [Goldenberg 2011]. More importantly, by approaching face to face experts or major stakeholders in an interview, a degree of knowledge in the field is required to have a fruitful and credible process, which was not the case of the authors of this research nor is it so common. This makes the use of secondary data a reason of advantage [Georgiou and Stevaux 2008].

Despite the advantages described above, in addition to the financial and time saving resources and the difficulties of conducting this sort of research alone, even so, there may be criticisms of this type of research material which was produced within a rigid structure and format with a controlled rhythm (as media events involve staging, camera management, previously recorded videos and so on) instead of a purely natural dialogue. Even though, we may assume the complex nature of the issues involved in these interviews, debates and reports, and the possible benefits to explore this issue with specialized tools.

Terms definition

We establish some terms that are used throughout this research, as follows:

- **Data corpus** refers to all data collected to this research. For this study, the *corpus* was built by the researchers herein authors and comprises the collection of 37 transcribed interviews.
- **Data set** refers to all data from the *corpus* that are being used for a particular analysis.
- **Data item** refers to each individual piece of data collected, which together make up a data set or the data *corpus*. Here, *item* refers to each of the 37 individually transcribed interviews. We remark that the software brings the term "text" to the data item. So, we can interchange terms throughout the document without loss of generality.
- **Data extract** refers to an individual coded chunk of data, which has been identified within, and extracted from, a data *item*.

4. Methodology

The methodology has two sequential and delimited stages. Each stage is composed by a set of activities. First stage (A) refers to the pre-processing and preparation of the textual data corpus. Second stage (B) refers to the analytical software application. Figure 1 outlines the proposed methodology which is detailed below.

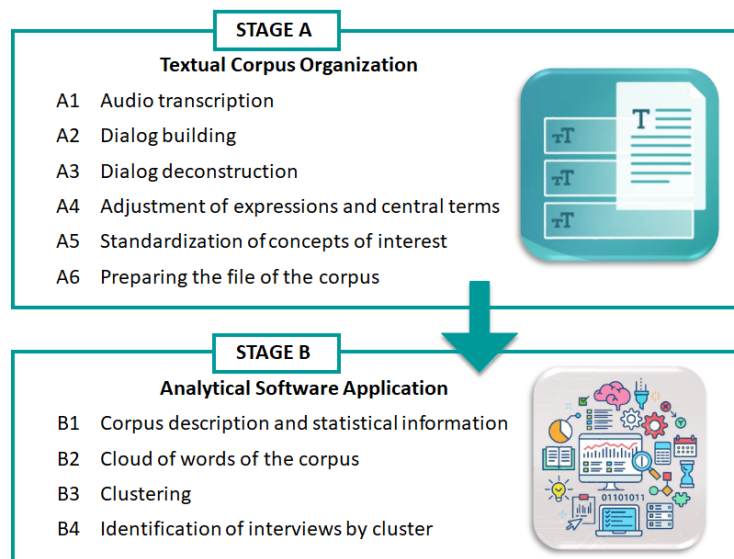


Figure 1. Methodology in two stages (elaborated by the authors).

Stage A: Textual Corpus Organization

The process of organizing the *corpus* contributed to the familiarity with data. In addition, being careful with data brings rigor to the research. Four sequential activities were performed as detailed below:

A1. Audio transcription

We feed (<https://downsub.com/>) the URL (Uniform Resource Locator) address which is the location of the webpage for each individual interview. Subtitles are automatically downloaded in TXT format. The accuracy of the transcription was checked by watching the interviews at reduced speed. When necessary, adjustments were made manually such as correction of a word or removal of any irrelevant data (e.g. sticks, songs).

A2. Dialog building

Once again, we watched the interviews aiming of building a dialogue. Thus, paragraphs and punctuation were inserted, beyond identifying who the speech belongs to (interviewer or interviewee).

A3. Dialog deconstruction

For each *data item* of the corpus, we performed a series of manipulations to prepare the data *corpus* to be inputted in the analytic software. Thus, the dialog corresponding to each interview was deconstructed and became a long and single paragraph in a block of sentences. The manipulation activities included:

- Each interview (data item) was rewritten in a single paragraph.
- Removal of signs such as question marks (?), apostrophes ('), semicolons (;), exclamation points (!), hyphens (-), dollar signs (\$), percentages (%), parentheses (), square brackets [], slashes (/) and other math symbols, leaving only periods (.) and commas (,).
- The questions were retrieved in the answers, as the example below:

Original Dialogue:

Question: *What is students' favorite discipline?*

Answer: *Science.*

After manipulations:

The favorite discipline of students is science.

- Removal of interjections and filling words, particularly when the interviewee repeats the same word or expression very frequently as a reflective purpose or just a habit with no meaning. E.g. *so, like, I think that, ah, well, perhaps, isn't it*, etc.

A4. Adjustment of expressions and central terms

Some terms are formed by two or more words. And, it would not make sense to analyze them separately. Thus, we connected them through underscoring. In other cases, we replace them by their acronym. All these cases are specified in Table 1.

Table 1. Manipulation of expressions and central terms.

Terms referring to	Process adopted	Examples
Places	Underscoring	São_Paulo, Distrito_Federal, Rio_de_Janeiro, Minas_Gerais, Rio_Grande_do_Norte, Mato_Grosso, Estados_Unidos, Coreia_do_Sul, América_Latina
Persons	Underscoring	Paulo_Freire, Itamar_Franco, Luiza_Erundina
Institutions and Organizations	Underscoring	Congresso_Nacional, Conselho_Nacional_de_Educação, Todos_pela_Educação, Instituto_Ayrton_Senna, Caixa_Econômica, Banco_Mundial, Universidade_de_Stanford, Centro_Paula_Souza, Universidade_Mackenzie
Projects and Disciplines	Underscoring	Plano_Nacional_de_Educação, Sistema_Nacional_de_Avaliação
Compound words or usual terms	Underscoring (kept the singular form of the term)	pré_escola, sala_de_aula, ser_humano, ponto_de_vista, bem_estar, dia_a_dia, mão_de_obra, salário_mínimo, rede_social, pensamento_computacional, projeto_de_vida, itinerário_formativo
School year	Underscoring	primeiro_ano, segundo_ano, quinto_ano, sexto_ano, nono_ano
Known terms by their initials	Use of the acronym	<ul style="list-style-type: none"> ◦ <i>Base Nacional Comum Curricular</i> replaced by “BNCC”. ◦ <i>Ministério da Educação</i> replaced by “MEC”. ◦ <i>Lei de Diretrizes e Bases</i> replaced by “LDB”. ◦ <i>Instituto Brasileiro de Geografia e Estatística</i> replaced by “IBGE”. ◦ <i>Produto Interno Bruto</i> replaced by “PIB”. ◦ <i>Ensino para Jovens e Adultos</i> replaced by “EJA”.

A5. Standardization of concepts of interest:

Some terms are mentioned in different ways by different interviewees. However, some of them are of particular interest in this research. Due to this, we standardized them throughout the corpus. They are listed in Table 2.

Table 2. List of adopted standardized terms.

adopted term	equivalent terms in the original text
educação_infantil	educação infantil, ensino infantil, anos iniciais, escola infantil
ensino_fundamental	ensino fundamental
ensino_médio	ensino médio, curso médio, escola média
ensino_técnico	curso técnico, médio técnico
ensino_superior	curso universitário, educação superior, curso superior
educação_básica	educação básica, ensino básico
fundamental_I	fundamental I, ensino fundamental anos iniciais
fundamental_II	fundamental II
ensino_a_distância	curso a distância, curso online, curso remoto

A6. Preparing the file of the corpus:

Each data *item* corresponds to the transcription of an interview. The items of the corpus were arranged in chronological order in a single TXT file. According to software requirement, a command line containing four asterisks and a notation was used to separate the items. Thus, the command line of the first item is ****** *doc_01**; the second item ****** *doc_02**, and so on up to the last item ****** *doc_37**. Each data *item* ready to be entered into the software is designated as an 'initial context unit' (originally 'unité de contexte initial') [Reinert 1990].

Stage B - Analytical software application

In this second stage, we explored IRaMuTeQ (*Interface de R pour les Analyses Multidimensionnelles de Texts Et de Questionnaires*), an open-source and open-code software for lexicographic textual analysis (www.iramuteq.org) which allows statistical analysis of textual corpus that is anchored in R software (www.r-project.org). In this research, we explored basic and multivariate lexicographical analysis. For all analysis, the following parameters were adopted:

Table 3. Iramuteq selected parameters.

Parameter	Selected
lematization	Yes
active forms (code 1)	(i) adjectives, (ii) unrecognized forms, and (iii) verbs
supplementary forms (code 2)	All remaining forms (e.g. adverbs, articles, pronouns, conjunctions, prepositions etc.)
standard dictionary	Portuguese
idiom	Portuguese
defined characters	utf8_all languages

B1. Corpus description and statistical information

We evaluated the statistical information and parameters provided by the software. This activity is also a way to verify if the software is reading the corpus correctly and if the corpus is considered coherent from a linguistic perspective. Parameters include number of documents, total number of words, number of words per document, frequency, and rank according to Zipf Law [Powers 1998]. The software relies on a dictionary to classify words (into adjectives, articles, preposition, verb, pronouns, auxiliary verbs and so on) as well as for the *lemmatization process*, when conjugated verbs, plural/singular words and certain suffixes are reduced into simple modes.

B2. Cloud of words of the corpus

Cloud of words shows up visually the most frequent terms of a textual corpus enabling a quick understanding of data at hand. It can be useful in preliminary analysis. However, it is recommended to associate it to other methods [McNaught and Lam 2010]. We evaluate three bands of words frequency of active forms, according to Cassettari *et al.* [2015]:

- **Trivial or basic terms:** define the central themes of the corpus.
- **Interesting terms:** define the emergent themes of the corpus.
- **Noise:** non-relevant terms.

B3: Clustering

In this activity, the corpus is divided into *text segments* (data extract) of a few lines (1 to 3), respecting punctuation breaks whenever possible. Each *text segment* is designated as an 'elementary context unit' (originally 'unité de contexte élémentaire') [Reinert 1990]. Through the hierarchical classification method, an iterative process that mixes several techniques (segmentation, hierarchical classification, dichotomy according to factor analysis and dynamic clouds) text segments are grouped into *lexical classes*. Each class constitutes a 'linguistic context' (not to be confused with 'situation context'), which gives us a sense of language use and its collective representation [Reinert 1990]. This clustering process is known as *Descending*

Hierarchical Classification (DHC) which is hierarchically organized according to the relevance of the term in each group (also called ‘class’). We attribute a label to each class. The interpretation of the classes can be done by understanding the most representative *text segments* of each class.

B4: Identification of a set of items per cluster

As a last activity, we identified the set of interviews (*data set*) that are part of each class (or cluster). Each set may deserve future deeper analysis according to the research interest.

5. Findings and Discussions

The data corpus is in Portuguese, as are the results provided by the software. Whenever possible, we present the results translated into English. Stage A resulted in the data corpus to be entered in the software (available with the corresponding author upon request). It is especially a handmade work which counted with the experience and sensibility of the researchers. Therefore, we present the outputs of activities of Stage B, as follows. We start presenting the corpus information and textual statistics in Table 4 (activity B1).

Table 4. Corpus information, textual statistics

Iramuteq information		Interpretation
Number of texts	37	Number of interviews (data item) in the corpus.
Number of text segments (TS)	2.243	The text segments have from one to three lines, automatically sized by the software.
Occurrences	79.333	Total number of words in the corpus
Number of forms	4.193	Number of distinct words.
Number of hapax	1.668 (2,10% of occurrences and 39,78% of forms)	<i>Hapax</i> are the words that occur just once in the corpus.
Average occurrences per text	2.144,14	Average of words per interview (data item).

We follow exhibiting the cloud of words (activity B2) which was built limited to the terms classified as *active forms* (adjectives, verbs, and unrecognized forms).

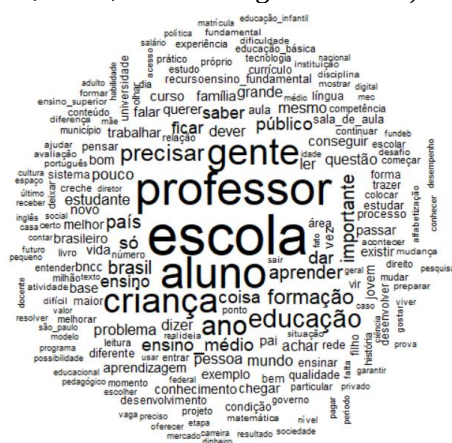


Figure 2. Word cloud of the data corpus.

The analysis of the word of clouds is made through Table 5 where we included the frequency range of terms and their importance in the analysis.

Table 5. Cloud of words analysis.

Group	Relevance	Frequency (f)	Total of terms	Examples
1 st (central sphere in the cloud)	Trivial (or basic)	$f \geq 26$	275	school (680), teacher (557), schoolchild (516), child (407), teaching (157), student (144) etc.
2 nd	Interesting	$26 > f \geq 5$	814	grade (25), progression (25), financing (24), fifth_year (22), preschool (20), language (20), stimulate (19), inequality (16) etc.
3 rd (outermost part of the cloud)	Noise (or non-relevant)	$f < 5$	2.598	appreciation (4), trajectory (4), television (4), minimum_wage (4), social_network (4) etc.

Third outcome is the clustering of terms (activity B3). From a total of 2.243 text segments (TS), 1.929 were utilized by the software, meaning a utilization rate of 86% for the construction of classes. This figure (above 70%) means that the corpus is suitable to this analysis. The content was categorized into 5 classes as shown in Figure 3.

Initially, the corpus branches into 2 sub-corpora (A and B). Sub-corpus A splits into two sub-corpora: A1 which contains **class 5**, labelled “*The importance of reading*”; and sub-corpus A2, which contains **class 4**, labelled “*Early childhood care*” and **class 1**, labelled “*Pedagogical practices*”. Sub-corpus B is divided into two (B1 and B2). The first includes **class 3**, named “*The dynamics of education*”, and the second contains **class 2**, named “*Public policies for higher education*”. The boxes include the number of text segments (TS) that made up each class, as well as the equivalent percentage. The labels assigned to each class are based on the sets of words that make up the class, as well as the titles of the interviews.

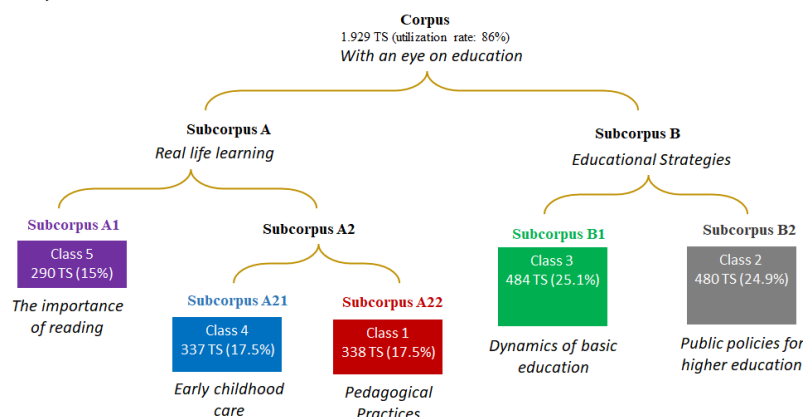


Figure 3. Clustering

By expanding the information in Figure 3, we have Figure 4 with the twenty most relevant terms in each class. Eventually, we skipped certain terms of minor meaning (e.g. take, well), replacing it by the subsequent one. The percentage shown in the right column means how much that sub-corpus (or class) concentrates that term in relation to the corpus. Based on these lists of terms and words, we gain insights into key issues in the research context. The words in **class 5** highlight the foundation of basic education that is our native language, Portuguese, and the various forms of communication and entertainment such as reading a book, writing a text, telling a story. The words and terms in **class 4** refer to very young children who are being introduced into the school environment (nursery), but the family still has a very important role in their education. The terms and words in **class 1** refer to the pedagogical practices that shall be present in the school routine. The terms and words in **class 3** refers to the educational management, including huge figures of Brazilian statistics. Finally, the terms and words in **class 2** refer to higher education. Although it is not part of the basic education, there are Government funding programs that cover this level of education, which is strategic for the future of a country.

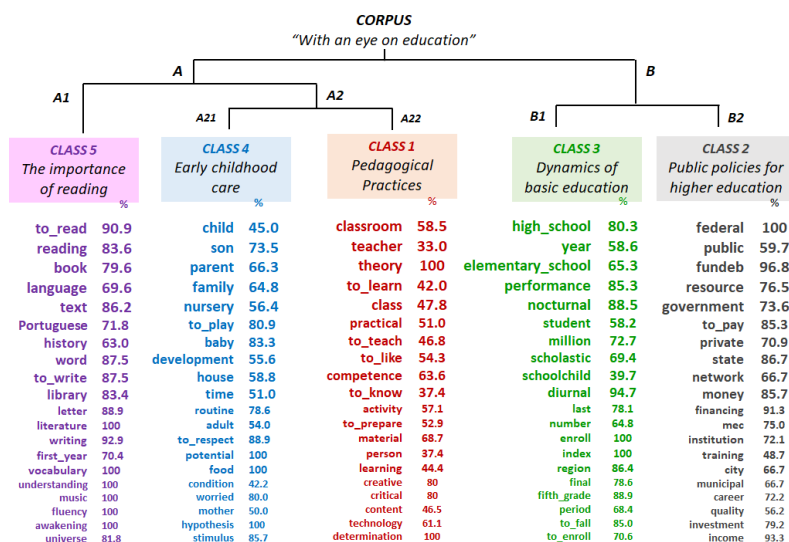


Figure 4. Dendrogram including the main terms of each class.

The relationship among terms and classes can also be seen on a two-dimensional plane (known as 'biplot'). The axes represent the *inertia* which is the percentage of variance. The proximity of terms in the graph indicates their association.

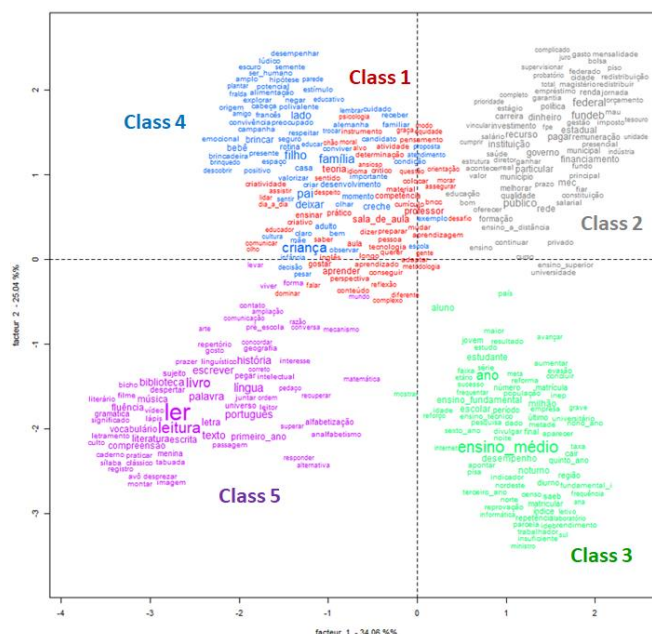


Figure 5. Factorial Correspondence Analysis

Last finding to be present is this report is the identification of the set of interviews (data set) that are related to each class (activity B4) in Table 6. This information is important for a more detailed analysis, that means, a "close reading" that will be part of a future and following phase of this research. The documents are presented in descending order of relevance. Relevance is measured by the amount of text segments that document contributes to each class.

Table 6. Data set that composes each class.

Class	Texts - interviews or data set (in descending order of relevance)
Class 5 (purple ink)	doc_27, doc_29, doc_33, doc_16, doc_13, doc_17, doc_02
Class 4 (blue ink)	doc_36, doc_32, doc_18, doc_10, doc_05, doc_02, doc_13
Class 1 (red ink)	doc_07, doc_20, doc_34, doc_22, doc_19, doc_28, doc_24, doc_03, doc_12, doc_06, doc_30
Class 3 (green ink)	doc_31, doc_11, doc_25, doc_23, doc_21, doc_26, doc_24, doc_14, doc_09
Class 2 (gray ink)	doc_35, doc_15, doc_37, doc_09, doc_30, doc_04, doc_08, doc_01

The analysis so far is initial and possibly considered superficial (insights). It was based on words, terms, and some previous knowledge on the problematic situation in the Brazilian education. However, it was possible to identify five distinct themes that deserve a dedicated further analysis. In addition, our corpus was broken in data items that facilitates the analysis by class (or cluster).

6. Conclusions and Future Avenues

Exploring data through diagrammatic representations can be convenient and interesting. In Operational Research it is commonly used with *quantitative* data such as statistical charts and graphics but also with *qualitative* data such as cloud of words and mental models. It can count on software support or not. In this research, we explored a set of televised interviews in Education that we transcribed and pre-processed to be inputted in the support software Iramuteq. We explored *basic* lexicographical analysis which uses terms (words / vocables) as unit of analysis, resulting in the cloud of words. Also, we investigated *multivariate* lexicographical analysis, when terms (words / vocables) were analyzed within the text segments, resulting in the *descending hierarchical classification* that took us the linguistic context.

Among our findings, we highlight the corpus categorization in five main classes that are suggested to be addressed in future national policy projects. They are: (i) the importance of reading, (ii) early childhood care, (iii) pedagogical practices, (iv) dynamics of basic education, and (v) public policies for higher education. Although the findings present in this research answer the proposed research question, it was limited to provide perceptions. Thus, we recommend further analysis for assertive actions. We mean that the research shall move from a 'distant reading' to a 'close reading'.

There are possible disadvantages of limiting our research to visual representations of information. It can be considered oversimple, confusing or difficult to be interpreted. This may require visual literacy and existing guidelines knowledge. For such reason, we also do not disregard the association to other methods and tools in future research.

Acknowledgement: This research was financed in part by the *Coordenação de Aperfeiçoamento de Pessoal de Nível Superior - Brasil (CAPES)* - Finance Code 001.

References

- Basole, R., Bendoly, E., Chandrasekaran, A. and Linderman, K. (2022). Visualization in operations management research. *INFORMS Journal on Data Science*, 1(2): 172-187.
- Cassettari, R.R.B., Pinto, A.L., Rodrigues, R.S. e Santos, L.S. (2015). Comparação da Lei de Zipf em conteúdos textuais e discursos orais. *El profesional de la información*, 24(2): 157-167.
- Delua, J. (March 2021). Supervised vs. Unsupervised Learning: What's the Difference? Disponível em: <<https://www.ibm.com/cloud/blog/supervised-vs-unsupervised-learning>>.

Eppler, M.J. e Burkhard, R. (2008). Knowledge visualization. In: Jennex, E. (Ed.), Knowledge Management: Concepts, Methodologies, Tools, and Applications. IGI Global, pp. 987-999.

Ertug, G., Gruber, M., Nyberg, A. and Steensma, H.K. (2018). From the editors - a brief primer on data visualization opportunities in management research. *Academy of Management Journal*, 61(5): 1613-1625.

Georgiou, I. and Stevaux, P. (February 2008). Strategic Options Development and Analysis: The Case of the Brazilian Railways. *Final Report to GV PESQUISA*. FGV-EAESP, São Paulo.

Goldenberg, M. (2011). A arte de pesquisar: como fazer pesquisa qualitativa em Ciências Sociais. 12ª ed., Editora Record, Rio de Janeiro.

INEP. Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira. (2020). Relatório Brasil no Pisa 2018. Brasília.

Jackson, P. e Moulinier, I. (2002). Natural Language Processing for Online Applications: Text Retrieval, Extraction and Categorization. Amsterdam / Philadelphia: John Benjamins Publishing Company. ISBN 902724989X

Lopes, A. A., Pinho, R., Paulovich, F.V. e Minghim, R. (2007). Visual text mining using association rules. *Computers & Graphics*, 31: 316-326.

McNaught, C., e Lam, P. (2010). Using Wordle as a supplementary research tool. *Qualitative Report*, 15: 630-643.

Mingers, J. e Brocklesby, J. (1997). Multimethodology: Towards a framework for mixing methodologies. *Omega*, 25 (5): 489-509.

Moretti, F. (2013). Distant Reading, Editora Verso, Londres.

Powers, D.M.W. (1998). Applications and Explanations of Zipf's Law. In: Powers, D.M.W. (ed.) NeMLaP3/CoNLL98: New Methods in Language Processing and Computational Natural Language Learning, ACL: 151-160.

Quoniam, L., Tarapanoff, K., Araújo-Júnior, R.H e Álvares, L. (2001). Inteligência obtida pela aplicação de data mining em base de teses francesas sobre o Brasil. *Ciência da Informação*, 30 (2): 20-28.

Reinert, M. (1990). Alceste une méthodologie d'analyse des données textuelles et une application: Aurelia De Gerard De Nerval. *Bulletin of Sociological Methodology/Bulletin de Méthodologie Sociologique*, 26 (1): 24-54. doi:10.1177/075910639002600103

Rosenhead, J. e Mingers, J. (Eds) (2001). Rational analysis for a problematic revisited (2nd ed.). John Wiley & Sons, Ltd.

Tan, A.-H. (1999). Text mining: the state of the art and the challenges. In: *Proceedings of the PAKDD Workshop on Knowledge Discovery from Advanced Databases*, Beijing: 65-70.