BJMKT
BRAZILIAN JOURNAL OF MARKETING

Received: 13 Dec. 2022/**Approved:** 29 Dec. 2022
Responsible editors: Julio Araujo Carneiro da Cunha & André Torres Urdan
Evaluation Process: *Double Blind Review*
e-ISSN: 2177-5184
https://doi.org/10.5585/remark.v21i5.21290

TUTORIAL ARTICLE

Check for updates

USE OF IRAMUTEQ FOR CONTENT ANALYSIS BASED ON DESCENDING HIERARCHICAL CLASSIFICATION AND CORRESPONDENCE FACTOR ANALYSIS

Marcos Rogério Mazieri
Universidade Nove de Julho (Uninove)
São Paulo, SP - Brazil.
marcosmazzzieri@gmail.com

Luc Marie Quoniam
Université du Sud Toulon-Var
La Garde, Provence-Alpes-Côte d'Azur, FR - França
mail@quoniam.info

David Reymond
Université du Sud Toulon-Var
La Garde, Provence-Alpes-Côte d'Azur, FR - França
dreymond@univ-tln.fr

Katia Cinara Tregnago Cunha
Universidade Nove de Julho (Uninove)
São Paulo, SP - Brazil.
katiapatentes@gmail.com

Objective: To present content analysis based on descending hierarchical classification and factorial correspondence analysis as complementary and sequential techniques, with possible application in the area of innovation management.

Method: Content analysis based on Computer-Aided Text Analysis (CATA) techniques using the IRAMUTEQ software Originality/Relevance: The research paradigm usually influences the researcher's own knowledge bases and domain methodologies. Manual text analysis can eventually be attributed to the interpretivist paradigm and CATA techniques can eventually be attributed to the post-positivist paradigm, however, there seems to be no reason to make this distinction.

Results: Presentation of a framework that demonstrates the technological choices of the most innovative companies in the world (Google, Apple and Amazon), common and different choices, among them.

Theoretical/methodological contributions: Development of a method for analyzing the technological choices of the most innovative companies in the world, applying content analysis based on descending hierarchical classification and factorial analysis of correspondence in a sequential and complementary way.

Social / management contributions: Decision-making for innovation management can be revised according to the technological choices presented. Competitive advantage has distinctiveness by nature. We demonstrate that it is not a problem to master the same technologies, that is, companies can have similar technological domains and still have different marketing approaches.

Keywords: Content analysis, descending hierarchical classification, correspondence factorial analysis, patents, innovation management, digital transformation, Iramuteq

How to cite the article

American Psychological Association (APA)

Mazieri, M. R., Quoniam, L. M., Reymond, D., & Cunha, K. C. T. (2022, Oct./Dec.). Use of iramuteq for content analysis based on descending hierarchical classification and correspondence factor analysis. *Brazilian Journal of Marketing*. 21(5) 1978-2011. https://doi.org/10.5585/remark.v21i5.21290.

Introduction

The general objective of this article is to present content analysis based on descending hierarchical classification and factorial correspondence analysis as complementary and sequential techniques, with possible application in the area of innovation management. The research problem that is intended to be used to exemplify the use of methods is based on the challenges involved in understanding the technological choices of companies considered to be the most innovative in the world. It is not possible to intuitively know whether the most innovative companies in the world make identical choices, if at all, how much they are similar. If they make different choices, what are the core technologies and how are the different technological choices characterized? Is there technological overlap? The research objective that can be stated for these problems is: To analyze and discriminate the technological choices of the most innovative companies in the world. Content analysis is the technique used to discriminate homogeneous thematic groups in a text, which are then organized into classes. Each class absorbs the most similar themes, so that, at the end of the procedure, there will be as many classes as the number of themes discriminated within a certain text or set of texts. The formation procedures of thematic classes for textual data analysis can be performed manually (Bardin, 1977) or with the help of computers (Short et al., 2010; Miraballes et al., 2018; Miraballes & Gámbaro, 2018; Campion & Champion, 2020). In order to evaluate the content analysis based on Computer-Aided Text Analysis (CATA) techniques, we intend to explain how textual data processing can be mathematically carried out, with a view to the area of applied social sciences in the area of innovation management to facilitate the later stages of interpretation of information to be carried out by the researcher. There is some controversy between the interpretivist view and the post-positivist view on the use of CATA techniques which we will comment on, largely using the assumption that text analysis is a systematic process that is neither doctrinal nor normative. Content analysis is not about finding something covered by the text, but what the text expresses and this is explicit information (Bardin, 2009). We are arguing in this article that CATA techniques can contribute to both interpretivist and post-positivist research. For this, we will present the operation of some of the most common text analysis algorithms.

The techniques responsible for automated processing are numerous. However, it was in the French school of data analysis that text analysis techniques, as a type of data, became evident and therefore such techniques were chosen as part of the object of analysis of this article. The main techniques used in the French school of data analysis originate from applied

mathematics, especially descending hierarchical classification (CHD) and correspondence factor analysis (CFA). The CFA has been known since the 1930s, verified in the work of Hirschfield (Hirschfeld, 1935), when it was called "treatment of data without average". The CFA technique came to be used especially by French researchers for the analysis of textual data after the 1973 work of Jean-Paul Benzécri (Benzécri, 1973) was published, whose main objective was to obtain graphical representations of the lines and columns of a 2 x 2 contingency table. The use of the CFA technique in the area of social sciences was intensified in the 1980s by works that showed cases of use in sociology, anthropology and psychology (Cibois & Jambu, 1981). The Descending Hierarchical Classification is also a statistical technique known since 1950, however, it was from the work of Máx Reinert (Reinert, 1990a, 1990b, 1995, 2007) that the use of analysis technique using CHD was intensified in the areas of social sciences.

In the area of applied social sciences, especially in the area of innovation management, there are research projects that require the analysis of textual data to analyze the content of a given set of data. Obviously as in all areas of knowledge, the increase in processing capacity, reduction of storage costs and the availability of software applications facilitated the use of these techniques and removed technical barriers, in a way encouraged researchers and practitioners to use analysis of automated content, however, what on one hand brought access and facility to apply the analysis techniques on the other hand made some analysis results hermetic, of partial understanding and in some cases of misunderstanding. The point is that researchers in the areas of social sciences are naturally not lexical, linguistic specialists, or professionals in the area of mathematics, nor are they computer engineers, although they need to discuss these areas of knowledge mentioned in order to be able to design and perform an aided content analysis by computer, which term is known as CATA. There are still the different aspects of studies that can be originated in textual analysis; content analysis, semantic analysis and discourse analysis. Discourse analysis consists of examining the ideological constructions of the text producers, where each text reflects the worldview of its producers and, therefore, this article is not about discourse analysis. Semantic analysis consists of examining the lexical characteristics of texts in order to seek a systematic understanding of concepts and information based on the constitution of the semantics and lexicon of the language and, therefore, this article deals with semantic analysis. Content analysis consists of examining the thematic classes that can be systematically grouped with interpretation induced by theory, experience or both, and this article will deal with content analysis in the field of innovation management, specifically in digital transformation.

We should point out that textual analysis is not synonymous with content analysis directly. Textual analysis is the organization of data contained in the text. The assumption of textual analysis is that texts are formed by data, symbolized by all spelling, grammar and other linguistic structures at the lowest level, which we call the lexical level. The combination of text data produces information contained in the text that does not have meaning, located at the middle level, which we call textual level. The chain of information present in the text follows a logic that produces sense or meaning, according to the researcher's experience, at a high level that we call content level. Texts are lexical representations of transcribed verbal natural language, as well as the lexical representation of human cognition in text form. The lexical analysis of a text can be obtained by breaking down the forms of the text (words), with the aim of allowing the classification of such forms that compose it in order to generate information derived from the text data, but which are not visible in the natural form of the text human reading. In this article, word and form were considered synonymous, since the algorithm used in the empirical stage of this research uses this nomenclature. Content analysis is the interpretation of information found in the text, guided by theory or empirical knowledge, therefore, under the inductive and data-driven approach or guided by *ex ante* propositions.

Due to the multidisciplinary involved in a computer-aided textual analysis, better understanding the specific mathematical procedures that are the basis of textual data processing algorithms can make the level of content analysis more familiar to researchers, both interpretivists and post-positivists. In the area of applied social sciences, content analysis has been carried out with the aid of established and competent manual content analysis protocols, described by several researchers, and in particular by Bardin (Bardin, 1977). Manual content analysis follows the assumption that the researcher has the ability to find content in texts, manually classifying textual forms, looking for conjunctions and disjunctions of parts of the text. The technique is traditional and used for many years and is dependent on the experience and energy of the researcher to classify all textual information. For larger amounts of text, such as one hundred or two hundred texts on an A4 format page, it can be even more challenging for the researcher, both due to the time and energy to be spent on the analysis, as well as the need to maintain the stability of the classification judgment criteria, established by the researcher in the content analysis protocol.

The research paradigm normally influences the researcher's own knowledge bases and domain methodologies. Manual text analysis can eventually be attributed to the interpretivist paradigm and CATA techniques can eventually be attributed to the post-positivist paradigm,

however, there seems to be no reason to make this distinction. This thought may be one of the misunderstandings that drive away researchers who are interpretivists of CATA techniques, while at the same time leading researchers who are post-positivists to partial, superficial or mistaken analyzes when using CATA techniques. In textual analysis aided by algorithms, all the processing of the classification of forms, involving obtaining conjunctive and disjunctive tables, is carried out through quantitative techniques, using Chi2 tests. Evidently, this analysis reaches only the lexical and textual levels, but not the content level. The content level is perceptual and in this case, it is more useful for the production of knowledge when carried out by the researcher. We argue that CATA techniques can be used both in the interpretivist research paradigm and in the post-positivist paradigm, since the automation involved in the analysis by CATA techniques interferes in the layers of the lexical and textual level, without prejudice to the content level.

Theoretical Backgroud Lexical, textual and content level

Lexicon is the name of the set of words in a given language. In the case of a textual analysis, it refers to the set of forms present in the text or in the set of texts under analysis. The linguistic challenges of shifting the meaning of forms over time must be considered, since it is an inherent property of the linguistic heritage transmitted between generations. Therefore, analysis of texts produced in a time interval greater than 15 or 20 years, between the oldest text and the newest text of the set of texts, need to be carried out in order to control the periods, dividing the texts by subperiods, for example. Lexicon and vocabulary are not synonymous, as the lexicon forms the vocabularies of a language. The lexicon is composed of semantic forms grouped into classes of forms, according to the grammar of each language. Nouns, adjectives and verbs are classes of forms that exist in many languages. Analysis at the lexical level consists of separating the classes of forms, including verbs, nouns, adjectives, not exhaustively, forming extracts of each class of forms with their frequencies of occurrence in the text or set of texts. This procedure is performed after lemmatization. Stemming is the technique of normalizing text words, transforming verb tense, gender and number. Normally, lemmatization takes verbs to the infinitive and nouns to the singular and masculine. When going through the lemmatization process, words are called forms when unique and occurrences when considering all repetitions, making it possible to compare forms within a set of texts. For purposes of operational definition in the area of innovation management, we argue in this research that the lexical level is the linguistic collection applied to the description of the new.

At the textual level, CHD descending hierarchical classification techniques and AFC correspondence factor analysis are applied. CHD is a procedure that involves a series of n-1 grouping decisions, where n is the number of shapes in a "tree" structure. The two basic types of grouping are agglomerative and divisive. In agglomerative methods, each observation starts as its own grouping and they are grouped two by two, consecutively. The divisive cluster type starts with a single cluster that splits into two, and each of the clusters formed into two more consecutively until each cluster is a unitary cluster. It is, therefore, a persistent iterative procedure until the stabilization of unitary groupings. The nature of the hierarchical algorithm that we have just explained, based on prof. Joseph Hair (Hair, 2009), makes it clearer that the classification of shapes is carried out ex post. It is based on the lemmatized words of the text or the set of texts that the so-called correlated forms are grouped, delimiting the thematic classes that are divided according to the significant hypothesis tests ($p<0.05$, h0 – the form belongs to the other class). Chi2 tests can be used to test this hypothesis, using the frequency of form i within the set of texts j. When dividing the set of texts into two parts, form i will have a higher Chi2 score in one of the two parts of the division of the set of texts and so on until form i stabilizes in a final grouping, called thematic class or semantic class. The procedure will be repeated with all the different forms i that make up the set of texts j. Therefore, thematic classes are not defined ex ante, thus remaining completely isolated from induction, either from the researcher's experience or from the theoretical interest itself.

We consider that Max Reinert's work represented a relevant evolution of hierarchical classification due to the classification approach that surpassed the frequency analysis of forms, but innovated with the idea of classifying ST text segments (Reinert, 1990a, 1990b, 1995, 2007; van Meter et al., 1991). Even though the hierarchical classification is the general statistical procedure for data classification, its application within the TS represents a relevant advance for the quality of the classification with repercussions for the content level and even making the idea of semantic analysis more robust. Text segments are uniformly sized clippings of each text. Instead of analyzing the frequency of words in the format i x j, the algorithm proposed by Reinert changes the level of analysis previously from the form to the ST text segment level, assuming that each piece of text that is written within a space of 40 or 50 letters, approximately 10 to 15 words, is likely to contain an idea. The hypothesis starts to be tested based on the Chi2 score of the ST, compared with the other ST of the set of texts. The clusters will stabilize, consisting of n contextually homogeneous STs, using the divisive hierarchical logic that is descending. This implementation of the TS-based view provided the metaphor used by Reinert

of the "lexical world", represented by the TS groupings of homogeneous context, making the connection between the lexical level and the textual level more consistent. It also facilitates the differentiation between algorithms that apply word frequency and those that apply additional techniques. In the case of Reinert's algorithm, the name given was Alceste, an acronym for Lexical Analysis of Context by Set of Text Segments. For purposes of operational definition in the area of innovation management, we argue in this research that the textual level is the frequent vocabulary applied to describe the new and a semantic class can be an innovative domain. The content of innovative domains can be composed of technologies, processes, functions, applications or practices, depending on the data source used in the analysis. For example, when analyzing full text of patents, the function, application or process of the technologies deposited in the patent metabases will be present in the innovative domains; when analyzing summaries of scientific articles on innovation, the research context, the research objective (reflection of part of the problem), the most used methods and discoveries will be present in the innovative domains and; when analyzing transcribed interviews on innovation, evidence of the cultural characteristics of the environment or the respondent's experience will be present in the innovative domains practices and innovation performance, in this case according to the fundamentals of the interview script. The innovative domain is not autonomous or self-explanatory. The innovative domain is a rational delimitation between the different innovative domains that are formed, after the descending hierarchical classification. Correspondence factor analysis is the technique popularized by the works of Benzecri and Cbois in the applied social sciences (Benzécri, 1973; Cibois & Jambu, 1981).

Application of content analysis in Innovation Management

Like many applied social science topics, innovation is broad in scope, as is innovation management. Research difficulties with a quantitative approach inherent to broad scope themes are recognized in the field of innovation. They are especially related to the difficulty of obtaining conceptual or operational definitions, delimitation of concepts, that is, the extent to which one concept reaches and another begins, and the problems of identifying flow in construct systems, which I can simplify by saying: difficulties in understanding which constructs are antecedents, which are intervening and which are consequent. As for qualitative approaches, the broad scope also poses challenges of conceptual delimitation and obtaining conceptual and operational definitions or even the definition and operationalization of qualitative variables (categorical or nominal, for example). The qualitative approach also

imposes the need for the researcher's great experience to make choices and make decisions about the "theoretical conversation" on which the research will be based, and which variables may involve propositions defined ex ante. Elementally, it is not necessary to define ex ante propositions for qualitative research, but researchers can choose to design the research in this way. If the researcher chooses to define ex ante propositions, the variables explained in these propositions can be included in the analysis algorithm and, therefore, can be interpreted and discussed in order to strengthen or eliminate the ex ante propositions.

In order to constitute some systematization of the innovation management approach, we turned to a classic book on innovation management by Joe Tidd and John Bessant, currently in its 5th edition. Edition (Tidd & Bessant, 2015). The choice is due to the recognition of this work by researchers in the field of innovation, as well as its intensive use in undergraduate and graduate courses. According to the compilation offered by Tidd and Bessant, there are four main steps involved in innovation management: search for innovations, selection of innovations, implementation of innovations and value capture. In this research, the idea is to present some suggestions for using content analysis to improve understanding of the intersection between the subareas of innovation implementation and value capture and, for this, we will use the patent information metabases.

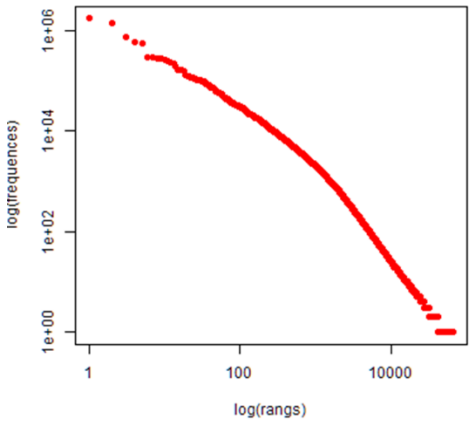
Patent metabase

Patents are industrial property titles limited by time and geographic location. There are more than 140 million patent applications and, by law, they must present the abstract in English or French, in addition to the language of origin. All patents receive an International Patent Classification (IPC) number, which is the identification of their patentability, that is, all patentable matter can be described by IPC numbers. In addition to the conveniences described above, patent metabases are publicly accessible and free of charge. Patent information has been used as a proxy for innovation efforts or as a result of innovation efforts. There is also great agreement that a patent is not an actual innovation, but a potential innovation. This is why we argue that by reaching a company's patent portfolio, we can analyze such portfolio as innovation management efforts that have already surpassed the search and selection phase. We argue that if a patent has been filed, such filing is a part of implementing the innovation, given the costs and resources involved in filing a patent. Ideas about the eligibility or otherwise of patent bases as part of the innovation strategy are based on the fact that small companies do not deposit patents or that certain less developed countries do not deposit patents. In the case of this

research, both restrictions do not apply due to the choice of the most innovative companies in the world. The 10 most innovative companies in the world concentrate around 560 thousand patents out of 140 million. Just for purposes of understanding, during this research, we accessed the 560,000 patents using the free Lens software (lens.org) chosen for its low cost and extreme simplicity of use by students and interested professionals, which makes it possible and simple the replicability of this study. The patents that had their due summaries were chosen and, in this case, there were 245 thousand patents. Other patent retrieval tools may increase the rate of retrieved abstracts but, in the case of this research, it will remain clear that we have a larger sample than the purpose of applying the method (245 thousand patents, 2.17 million occurrences, 66 thousand unique forms and 23 thousand hapax (0.11% of occurrences and 35.52% of forms). The average of 88.75 occurrences per text demonstrates the textual nature of the sample. In the next figure we present the relationship between rang (number of word categories) and frequency of the categories, both with the application of logarithm. Even with the application of the logarithm, we found close to 20 thousand rang at the end of the x-axis and 1 x 10^6 or frequency greater than 1 million. At these sampling levels, it is not easy to manipulate the data with computers and more domestic spreadsheets due to time (from a few hours to process to days for each review). Therefore, the cost of computation and manual intervention of data preparation may not be justified. wanted, we carried out the procedures to inform the need for reflection regarding the sample size to be considered.

Figure 1

Basic Lexical Analysis



Source: Authors, 2022.

In terms of minimum sample, there are authors who recommend 30 texts, so the approximate minimum limit is known. As for the maximum limit, those with more than 2 million occurrences are considered large corpus. The thing is, bigger databases don't mean better research. After analyzing more than 30 million occurrences between 2013 and 2022, some empirical conclusions can be shared in this article. There are textual analyzes in the microm, meso and macro context. By context of text analysis Micro is considered, empirically, the corpus that cannot be read or analyzed manually; from 50 texts. Less than 50 texts can achieve better results if analyzed manually for greater detail. The mathematical assumptions involved in the classification or factoring techniques consider matrices composed of the presence or absence of shapes and, although small samples such as those composed of 30 texts can work, in terms of the operation of the analysis software, they can frustrate due to the superficiality of the results it offers in semantic terms. Corpus formed with amounts between 50 and 20 thousand texts can be considered Meso text analysis contexts, which is when we can expose the content to be interpreted more effectively. From 500,000 texts or 2 million occurrences (as demonstrated in this article) the power of discrimination of the text segments is reduced, even making a direct analysis unfeasible, what we call Macro text analysis context. The analysis of corpus larger than 500,000 texts or 2 million words should use extensions of the techniques described in this article, such as prior clustering and testing subsampling techniques and learning subsampling when the hardware and software tooling is simple (personal computer processing). In the case of this research, when we processed 2.3 million occurrences, the Descending Hierarchical Classification was unable to classify at least 3 semantic classes, which eliminates the possibility of hierarchically analyzing two classes, having no practical sense or utility. The practical effect is that the analysis software does not reach the end, making the CHD analysis unfeasible. We emphasize that this behavior is specific to the current sample due to the vocabularies involved. The issue is that the larger the sample, the greater the chance of homogenizing the textual content and reducing the power to discriminate classes due to the mathematical principles involved in solving the presence and absence matrices of words, which ends up being a binary combination.

Methodological procedures

Next, obtaining a base with more than 2 million words will be detailed. The intention of this section is to make explicit the sample choices, the delimitation of the sample size and the

main parameters to be considered to perform the descending hierarchical classification and the correspondence factor analysis.

We know that the choice of a tool over others or a technique over others is always a reason for great debate. For this research, as for the others we conduct, we always have cost and access to tools and software as our main selection criteria. We are inclined to carry out research with the social responsibility of allowing a student or professional who is in a more remote region to be able to replicate and continue the study. As for accessibility, we consider two points. First, if the software or tool can be accessed over the internet, downloaded, installed and used without the need for great specialization. Second, if the code is open, it allows understanding the algorithms or even correcting or expanding the code. The tool that achieves these characteristics is Iramuteq (Camargo & Justo, 2013; Ratinaud & Marchand, 2012). Tropes is open source and free to download and install, but has a slower learning curve than Iramuteq. All the most famous software in Brazil, such as Atlas TI, Maxqda and Nvivo, are software that have a trial version at no cost and then paid versions are offered, which leads us to not choose such software for cost, in addition to, due to their commercial nature, they are not open source, a second point that does not meet the criteria established in our research. Obviously all are good products within their characteristics, but they do not meet the selection criteria established by this research.

The choice of companies to be analyzed was based on the ranking of the 50 most innovative companies by the Boston Consulting Group. The Boston Consulting Group analyzed 1500 companies based on Global MindShare, Industry Peer Review, Industry Disruption and Value Creation criteria. Global Mindshare, consists of the consistency of the response of the responding executives regarding innovation, responding about their own companies. Industry peer review consists of executives voting on the innovation of companies other than their own. Industry disruption is an index that measures votes on industries. Value creation is the total return on shares for the last three years, from December 2017 to December 2020 (Ang, 2021). Below is table 1 with the 50 most innovative companies in 2021, according to the Boston Consulting Group.

Table 1

10 Most Innovative Companies of 2021 According to the Boston Consulting Group

2021	Firm	Industry	Headquarter	2020
1	Apple	Technology	US U.S.	--
2	Alphabet	Technology	US U.S.	--
3	Amazon	Consumer Goods	US U.S.	--
4	Microsoft	Technology	US U.S.	--
5	Tesla	Transport & Energy	US U.S.	6
6	Samsung	Technology	KR South Korea	-1
7	IBM	Technology	US U.S.	1
8	Huawei	Technology	CN China	-2
9	Sony	Consumer Goods	JP Japan	--
10	Pfizer	Healthcare	US U.S.	Return

Source: Authors, 2022 based on Boston Consulting Group.

For this research, the 10 most innovative companies in the world, representing 560,000 patents, were analyzed. As previously explained, in this sampling, the descending hierarchical classification is unfeasible.

On the problem of time between invention and innovation and sample choice

Typically, innovation researchers have the additional challenge of establishing the distance between innovation efforts and innovation itself. There are many works that propose the most different ways to treat this distance.

In this research, we defined to consider the time of three years between effort and effective innovation. The innovation effort will be identified in the patent deposits of the most innovative companies in the world that occurred between 2016 and 2019 to be reflected in effective innovation in 2021 and 2022. Effective or accomplished innovation will not be the object of analysis in this research.

The criterion for choosing the sample must follow the theoretical alignment involved in the research. In the case of this research, the choice is made by judgment, established in the ranking of the 50 most innovative companies and specifically observing the 10 best placed. Of the 10 best placed, we decided to analyze the first three companies in the ranking - Apple, Alphabet/Google and Amazon, considering the limit of textual analysis Meso (20 thousand texts) the sum of the patents of these three companies between 2016 and 2019, which is approximately 19 thousand patents.

Search expression

Usually looking for words in Dictionaries, Thesaurus and Literature Review can help to find keywords from a linguistic etymological or synonymic point of view. The energy curve (word frequency) is inverse to the saturation curve (accumulated frequencies). Population, by definition, is unknown. The justification of the sample starts from the etymological linguistic analysis, synonymy, but it is still not enough because the applicant can use in the description of the fields of the patents targeted by the search other terms than those discovered in the definition phase of the search expression.

These terms may be related to language addictions or cultural variations and other semantic differences. In the case of this research, given the search for patents by certain holders or applicants (Apple, Google and Amazon), we have a thesaurus reduced to the names of the companies. However, problems with the way the company names were written or even companies with the same name but which are not the companies of interest. This research will not deal with the construction techniques of the search expressions, we just point out that care was taken to remove as much noise as possible.

For this research, the patents of interest were filtered in an electronic spreadsheet and patents that were unrelated to the body of expected inventors or international patent classification codes that were unrelated to the expected portfolio were identified. In the event of one or more of these inconsistencies, the patent was examined individually and a decision was made to delete it or keep it in the database that would later compose the textual corpus: Busca lens.org (applicant.name:apple).

Iramuteq analysis configuration plan

Step 1: Configuration of the textual corpus

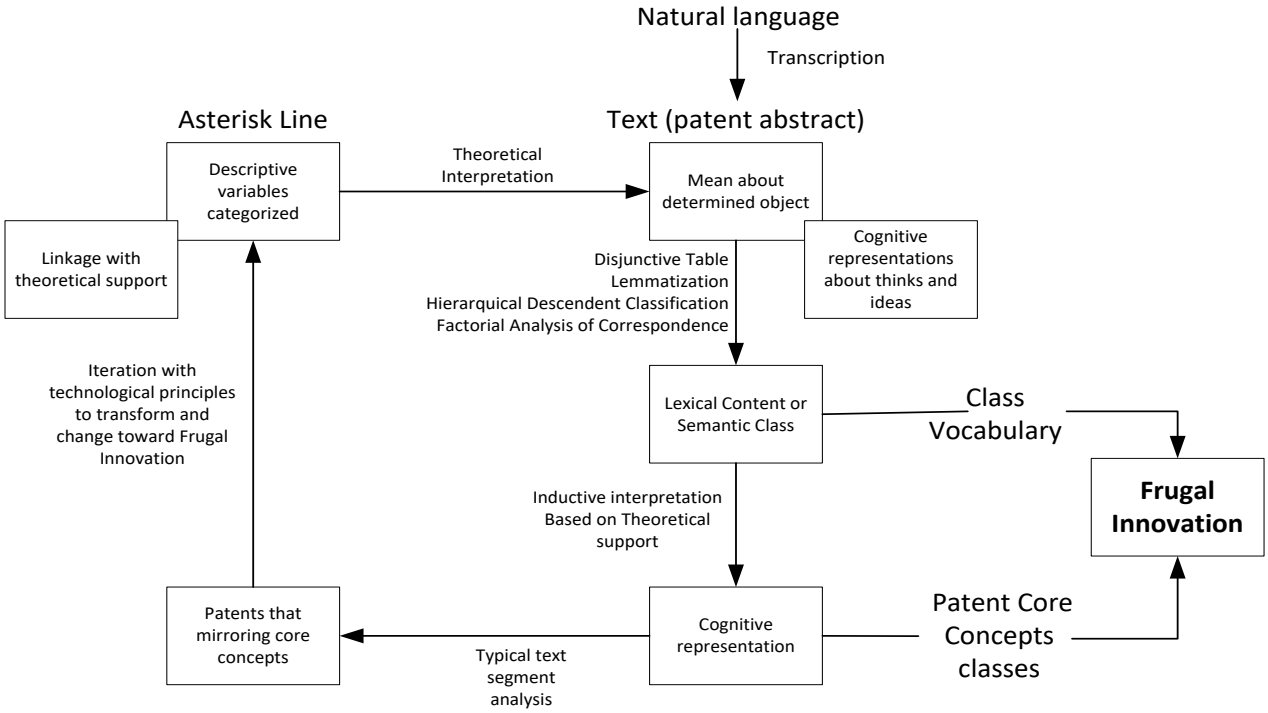
As main procedures for configuring the textual corpus, we highlight two main ones: first, all the texts to be analyzed must be placed in a single electronic file in the standard “txt” format “Unicode UTF-8”. The texts have different origins, such as articles, news, patents, interviews, books, mission statements or company vision, business documents, such as meeting minutes or innovation policies. Maintaining the homogeneity of the corpus is the most critical point of this step. We have noticed that little attention has been paid to corpus preparation; researchers invariably prepare interview scripts with questions related to theoretical topics that are distinct in terms of constructs or main ideas. Thus, when preparing the textual corpus of this

interview, for example, with five questions, each answer must be joined to a textual corpus, so that all the answers to question 1 will be gathered to form the textual corpus about question 1 that is of interest. in understanding the theoretical topic that the question represents. Obviously, if question 1 and 2 are related to the same construct or the same theoretical background (1 and 2), they can compose the same textual corpus. Another classic example is to intend to analyze complete technological or scientific articles and, for that, a textual corpus is created with all the sections of the articles. Articles, as well as other documents that are constructed according to established formats, have heterogeneity in content and writing style when we compare the abstract with the introduction or with the theoretical framework, for example. This means that, for content analysis purposes, it does not seem to make sense to mix all sections of an article into a single textual corpus. The recommendation is to choose a specific section of the article and in the same way when analyzing a patent. The safest and most traditional choice is the abstract, but there is no impediment to choosing other sections of interest, as long as the textual corpus formed maintains homogeneity in terms of the expected content.

The second important point in the configuration of the textual corpus is the organization of manifest variables, which are the variables related to the origins of each text that will compose the textual corpus. It is through these variables that it will be possible to reach more in-depth interpretations when in a data-driven paradigm or it is still the point of connection between textual data and the theoretical lens in use in the research. In the case of patents on frugal innovation, the analysis model that relates the descriptive or manifest variables with the textual corpus is presented below. The textual corpus represents the meaning of certain objects issued by the producer of the texts that compose it. Therefore, for each text that will compose the textual corpus, there is a set of manifest variables that link this text to theory, metadata, demography, psychographics, behavior or other attributes that explain the origin of the text (Mazieri, 2016).

Figure 2

Model of Integration Between Manifest Variables or Theoretical Elements and the Contents of The Textual Corpus



Source: Mazieri, 2016.

It is possible to identify that the connection between the theoretical support and the cognitive representation that will be analyzed in the texts is made in Iramuteq through the line of asterisks. Below is an example of the row of asterisks from the current search.

Table 2

Text Format Model for Construction of the Textual Corpus

****** *Applicants_APPLE INC *Kind_B2 *Publication Date_28/11/2017 *Application Number_US 201514656048 A *IPC_IPCABSENT**
Asset data streams are provided that facilitate the display of large numbers of media assets. Encoded asset data streams provide approximated aspect ratio information for media assets to be used in determining a position for each media asset in a dataset, thus being able to position all of the media assets in a media asset arrangement prior to being scrolled into view by the user. By communicating aspect ratio approximations as part of an encoded asset data stream of data to a web application, a user is able to scroll to any part of a dataset (e.g., a photo library) when presented in the web application without having to wait on the receipt of

****** *Applicants_APPLE INC *Kind_B2 *Publication Date_28/11/2017 *Application Number_US 201514656048 A *IPC_IPCABSENT**
information for all media assets. Encoded asset data streams may further include asset identification offsets that indicate a sequential ordering of the individual assets in a dataset.

Souce: Authors, 2022.

The entire line of asterisks starts with 4 asterisks, signifying the beginning of the text. Before each manifest variable, we use an asterisk followed by the name of the variable (applicant, for example, is the name of the patent holder). After naming the variable, an underscore (_) and the value for that variable (APPLE INC, in this case) are placed. After the declaration of the manifest variables, follows the text that will be one of the parts of the textual corpus. This procedure must be done for all the texts that will be grouped to compose the textual corpus.

General guidelines on text content

Content analysis software is based on vocabularies, so it is necessary to check that the text is free of typing errors. In the case of interviews, the researcher's notes and interventions should not appear in the transcribed text. We must not use bold, italics or other formatter. As a practical tip, due to difficulties in using punctuation, it is suggested to leave only paragraphs, without including commas or semicolons. Hyphenated words will be treated as two words within the text, just like compound words; if you want to treat santa casa as the place where patients are treated, you must write santa_casa, otherwise the algorithm will analyze the term santa separately from the term casa. Verbs with pronouns must be in proclisis form, because the Iramuteq dictionary does not provide for inflections; instead of inform me we should use inform me. Nowhere in the text can we use apostrophes, quotation marks, hyphens, dollar signs, percentages or asterisks.

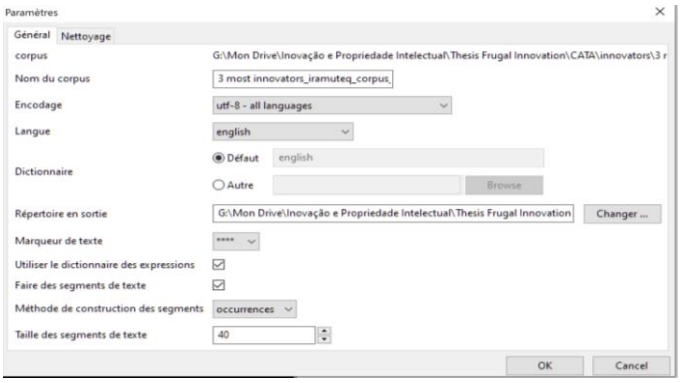
Step 2: Iramuteq configuration

In this section, some screens will be presented with comments on the configuration to be used, when applicable. The screens offered as guides are in French because Iramuteq was developed in French. There is an option to use the interface in English and some other languages, however, the more advanced parameterization menus are still in French. Therefore,

the idea is to present the screens in the language that has complete coverage of the menus and settings. For everyday use, you can choose the language of interest, knowing that the most advanced parameters will not be translated.

Figure 3

Parameters for Importing the Textual Corpus



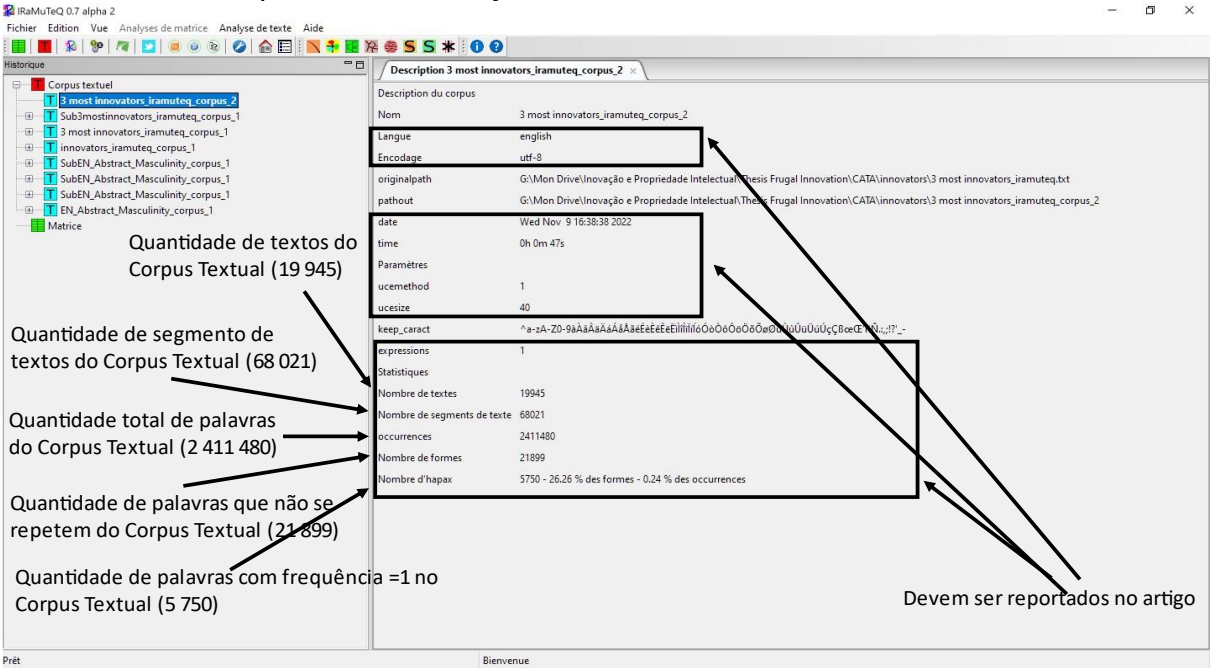
Source: Authors, 2022.

When requesting that Iramuteq load the textual corpus to be analyzed, the screen above is displayed for choosing interface parameters between Iramuteq and the electronic file where the textual corpus is located. The encoding to be used is UTF-8 all languages, the dictionary is the default, the text markup is the 4 asterisks (****) using the dictionary of expressions, using text segments constructed from occurrences whose size of the text segment will be 40 characters long. The construction mode of the text segment can be changed according to the analysis needs and needs to be reported in the scientific article where the research is being presented. The text segment construction mode options are: (i) number of characters, when the texts that make up the corpus are small, such as a title or a sentence, (ii) number of occurrences when the texts are composed of two or more sentences, as in the case of summaries or other parts of a text, or (iii) length and punctuation criteria; in this case, the scoring criteria are calculated by the size/score ratio with the priority ".", "?", "!" then ";", ":" and lastly "," and space. The idea of this way of constructing the text segment is to adapt as much as possible to the structure of the language. In textual data analyzes involving innovation management, when secondary data that have undergone spelling and grammatical review are analyzed, this mode can be useful. In the case of interview analysis it can be risky due to the need to control the grammatical and spelling quality of the transcript, therefore this mode is not recommended. The language to be chosen on this screen is the language in which the text transcription to be analyzed is written. In the

case of this research, the abstracts of the patents are in English, so the English language was selected.

Figure 4

Initial Parameters of the Content Analysis Plan



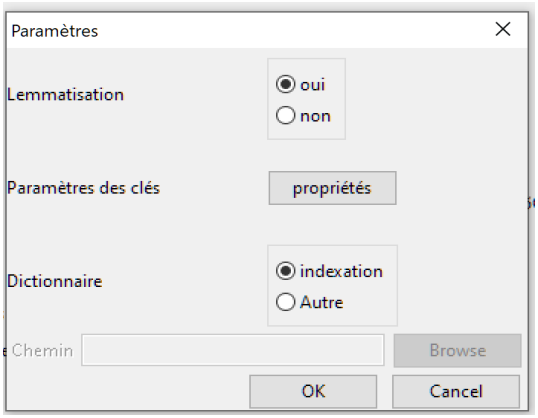
Source : Authors, 2022.

When loading the textual corpus, the load summary parameters are offered and these must be reported in the article that is using analysis with Iramuteq, both for the reader's understanding of the conditions of the electronic file where the textual corpus is located, its size, as well as the parameters chosen to load the corpus. There are also the basic lexical statistics that show how many occurrences, s,t,, forms and Hapax were identified in the textual corpus. When you reach this stage, it means that there are no coding errors and that the analyzes can be tried.

Any analysis that is started goes through the following parameterization steps. For example, in the last step of Iramuteq configuration, we requested a basic lexical analysis, and the parameters window will appear as shown below. Stemming should be selected and the dictionary by indexing too, except in cases of using external dictionaries.

Figure 5

Lemmatization and Dictionary

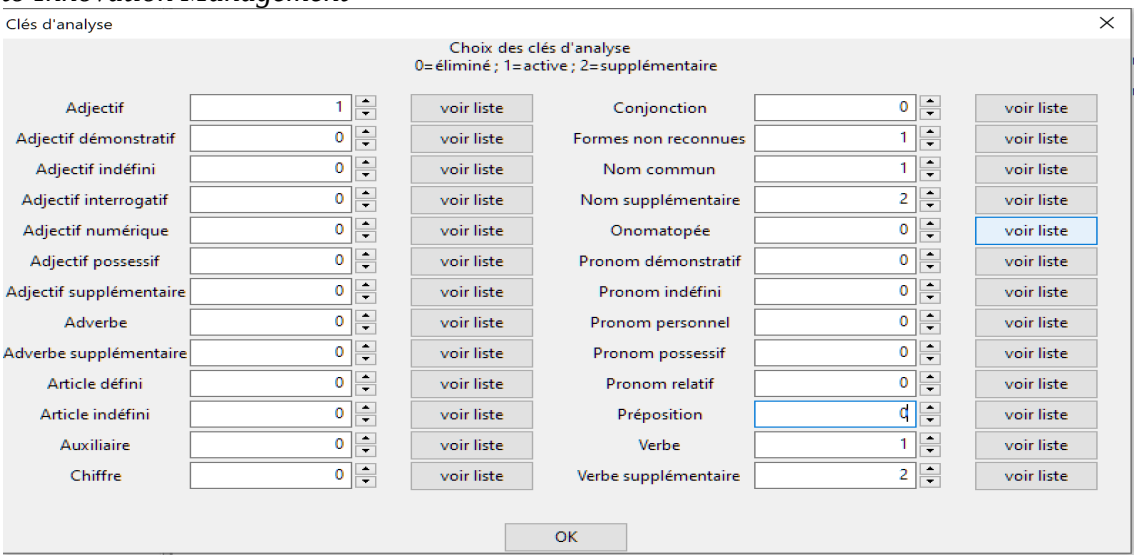


Source: Authors, 2022.

We must select the properties to define the treatment that the algorithm should give to the different groups of words, which we call analysis key. This step is quite neglected, which causes visualization and interpretation difficulties. The concept of content analysis is to achieve explicit cognitive representation in the text, therefore, it is not a study of modes of speech or writing. In this way, the parameters of interest are the adjectives, nouns and verbs that will be marked as active forms (1). The supplements of verbs and nouns, marked as supplements (2), it is recommended to mark unrecognized forms as active (1), because in innovation we are interested in something that we do not know how to classify and is present in the text. Next, the final screen of how Iramuteq should be parameterized to start the analyses. An important detail is that these parameters are kept only while Iramuteq is kept open. If for any reason the software is closed, when reopening, these parameters must be adjusted again to perform any new analysis. Obviously the analyzes already carried out are saved and their results keep the configuration parameters chosen before the analysis.

Figure 6

Presentation Parameters for Improved Visualization Purposes in Analysis of Topics Related to Innovation Management



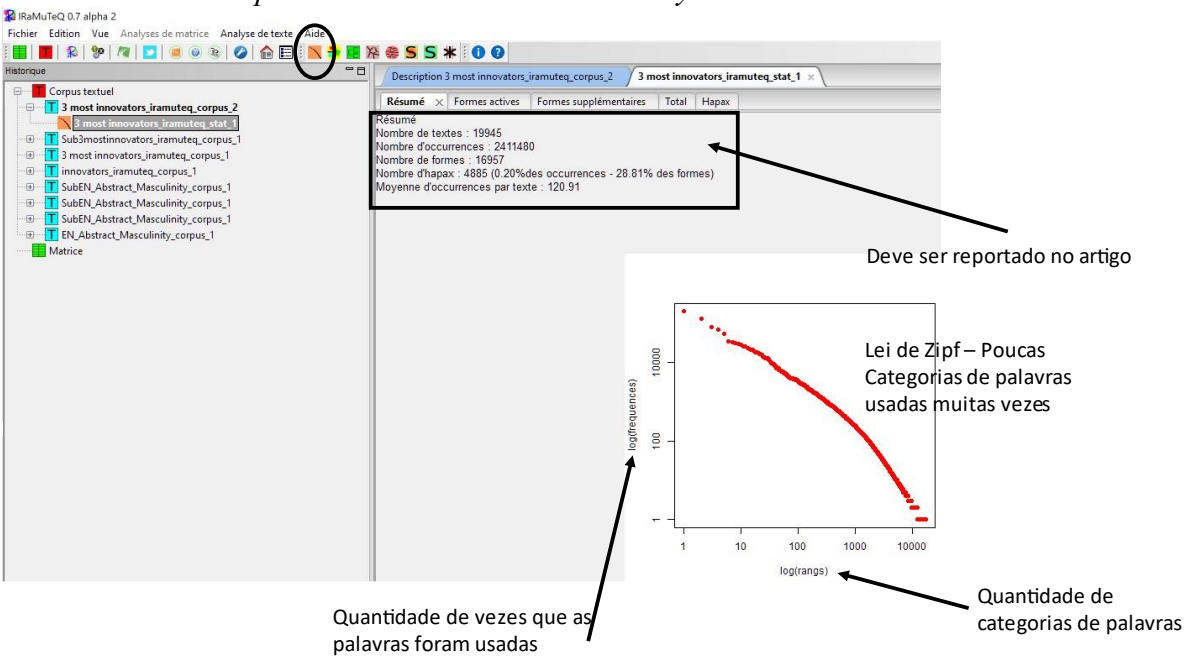
Source: Authors, 2022.

Step 3: Analysis Request

Start or First analysis: basic lexical analysis

Figure 7

Parameters to be reported in the article in the "Analysis Plan" section



Source: Authors, 2022.

The difference between the number of forms discovered in the statistics screen of the textual corpus that was loaded and the one that will be effectively analyzed in the basic lexicography occurs due to the exclusion parameters used to leave only the active and supplementary forms of interest.

Analyzes available in Iramuteq and their applications in innovation studies

Basic Lexical Analysis

Through a basic lexical analysis we can know the structure of the textual corpus in terms of size and lexical structure. In innovation, it is possible to identify the most common vocabularies used in texts and to know how a certain technology is usually described, what are the most used words and what are the most used grammatical classes, making it possible to perceive the pattern used to record knowledge in the researched area.

Specificity analysis and AFC

Analysis based on matrices whose lines are the words of the textual corpus and the columns are composed by the manifest variables chosen for analysis. It is possible to verify both the frequency of the shape in the variable (effect) and the chi-square of the shape in the variable (relative effect). Certain nouns can be important markers for understanding a particular technology. In our case, the word “antenna” was identified as a word present in Apple technologies, but not so present in Google or Amazon technologies. In this way, the word “antenna” will be classified in one of the semantic classes and this class will have a predominance of Apple technologies. It should be noted that we did not have this information before starting to analyze the content of the patents.

Similitude analysis

Analysis based on graph theory that presents words in co-occurrence, that is, they appear together in the textual corpus. This analysis allows you to know which words are most closely related. In our case, “antenna” is related to substrate and surface. We can understand that Apple's patents that describe antennas are patents related to electronic construction and not aesthetics, describing how the coupling of the transmission antenna of certain equipment happens.

Descending hierarchical classification

It is the classic descending hierarchical classification, using the Renert method which, instead of words, uses text segments. There are three main options: Double sur srt, which we normally don't use due to low retention of text segments; Simple sur text segments which is the standard form of analysis and is effective in longer texts and finally simple sur textes which is the direct analysis over the text, without considering the text segments, an alternative for short texts such as title analysis or keywords, for example.

Factorial correspondence analysis

Although the mathematical principles were explained in the introduction to this article, it is important to complement that the correspondence factorial analysis is formed by a matrix in the form of a contingency table that crosses the active forms with the variables (defined in the line with the asterisks).

Word cloud

Iramuteq also offers the word cloud. However, there is no effective use for the purposes of technical or scientific analysis because it is only visual, without presenting parameters. Usage is much more illustrative than useful, analytically. It consists of showing the words present in the corpus in font sizes that represent the frequency of their presence in the text.

Results and interpretation

Resuming, the research problem that is intended to be addressed using the textual analysis methods described, it is intended to analyze and discriminate the technological choices of the most innovative companies in the world. The CHD of Apple, Google and Amazon patents formed 4 semantic classes. Class 4 is hierarchically superior, which means that it can carry the most homogeneous vocabulary within the corpus (to confirm, it is necessary to check in the AFC if class 4 is positioned at 0.0 standard deviation of the factorial plane). In other words, the vocabularies present in class 4 are shared between Apple, Google and Amazon. Class 3 is the next class in the hierarchy and the fact that it is separate from class 4 demonstrates that there is a distinct vocabulary. The same interpretation for semantic classes 2 and 1.

Figure 8

Descending Hierarchical Classification Dendrogram with Apple, Google and Amazon Technology Choice Classes: Apple, Google and Amazon patent CHD from 2016 to 2021 With 19,000 patents

Source: Authors, 2022.

This initial analysis allows us to answer our first question: Do the world's most innovative companies make identical choices? The answer is that there are at least three semantic classes: class 4, class 2 and class 1, which show that the technological choices do not seem to be identical, since they present vocabularies that are hierarchically discriminated.

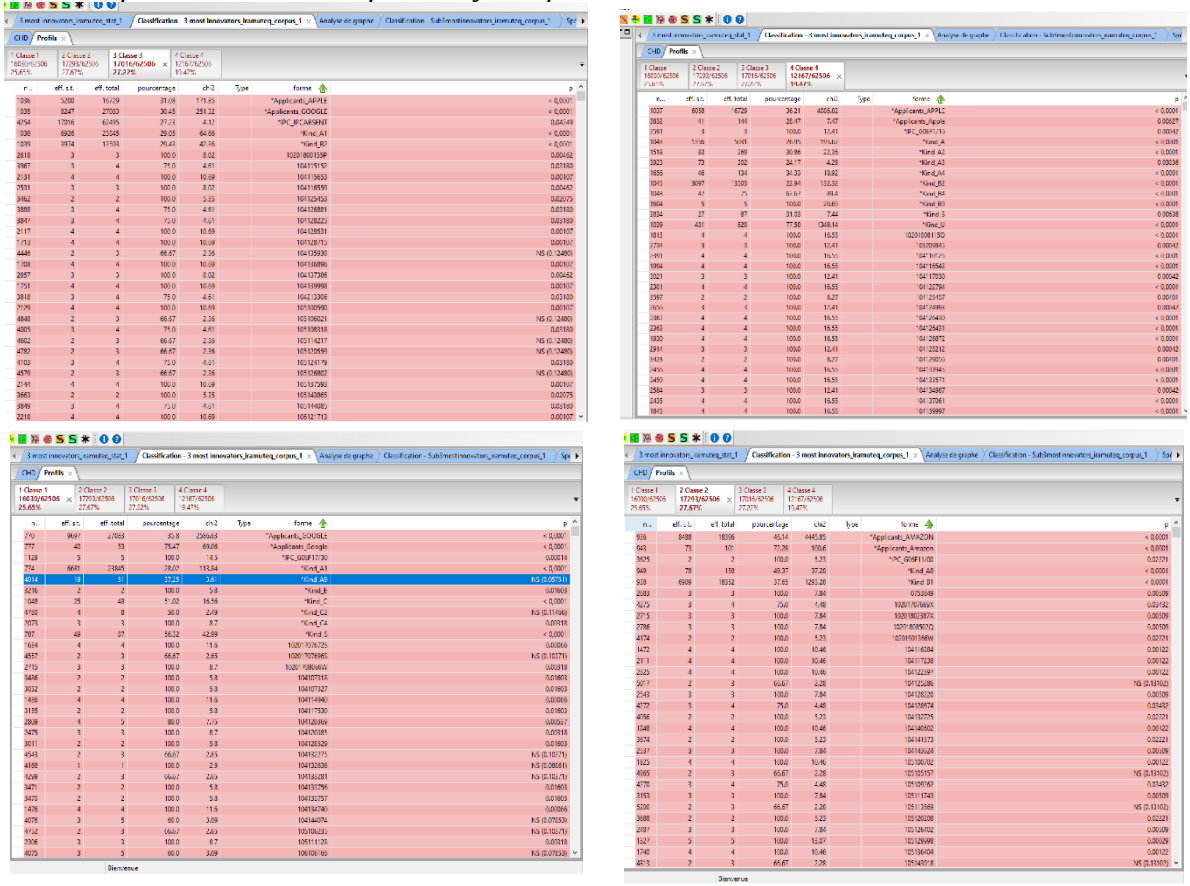
As for the second question: If the choices are identical or similar, how are they similar? We can answer that the technological choices are similar in relation to the content stored in semantic class 4, where the vocabularies are shared or common to the entire analyzed textual corpus. The practical phenomenon is simple to understand and is related to product portfolios in multiple verticals. Apple develops smartphones, but also services and even self-driving cars. Google develops quantum computing software, services and hardware. As for Amazon, there is a presence in retail sectors such as e-commerce, space and other areas. Therefore, even though there is discrimination regarding semantic classes, it is necessary to evolve in understanding each class before answering this question.

To rationalize the interpretation: Each class stores the active forms (words that form the vocabularies of the class, such as verbs and adjectives), the complementary forms (high frequency forms that, therefore, do not express content) and the manifest variables (originated

of the rows of asterisks). Once the analysis possibilities of each class have been established, the recommendation is to analyze the manifest variables that are marked in red and the active forms that are marked in gray. Complementary forms, marked in green, are normally not analyzed due to their low explanatory power. In figure 9, each class can be verified; as an example, when selecting class 4 and clicking on "form", the list is organized in descending order, showing that Apple is predominant in this class (Significant $p<0.0001$). That answers part of the question. Apple predominates in Class 4, but is also present in Class 3. Class 3 is shared by Apple and Google. Class 2 is predominantly Amazon and Class 1 is predominantly Google. It is interpreted that the technological choices of class 4 are predominantly Apple, but also carry the most common vocabulary of the entire textual corpus due to its position in the AFC (0,0); class 2 technology choices are predominantly Amazon; Class 1 technology choices are predominantly Google. Therefore, class 3 is different from class 2 which is different from class 1 and shows the part of technological choices that are different between Apple, Amazon and Google. Another conclusion is that Amazon does not share the technological choices of Apple or Google, since there is no simultaneous occurrence of Amazon in the same class as other companies. Finally, similar technological choices occur between Apple and Google and are located in Class 4.

Figure 9a, 9b, 9c, 9d

Details of Semantic Classes for Analysis of Associated Variables



Source: Authors, 2022.

The next analysis step is to access the content of the active forms and identify which technologies are stored in each class. To make it clearer, each semantic class has technological options from both Apple, Google and Amazon, since the class is the result of grouping the text segments that are fragments of the natural language present in the patent summaries, without distinction from the company to which they belong. Therefore, when analyzing the predominant manifest variables in each class, we are verifying whether or not there is a predominance of this variable in the class. In the case under analysis, the variable is the patent holder. After verifying the predominant variable in the class, the next step is to analyze the content of the class vocabulary.

To do this, starting with class 4, just click with the right button of the mouse on the form with the greatest Chi2 effect and check the agreement with the text segments (concordancier). Next are the reproductions of the "surface" form of class 4. It can be interpreted in a more basic way that part of the similar technological choices between Apple, Google and Amazon are the

hardware projects, mainly involving the modules where they are fixed the active elements (such as electronic circuits), and the passive elements (such as antennas, lenses and supports) are technologies that support the aesthetic (design) and functional aspects (connections and mounting systems). The interpretation is done by reading the segments of texts that were stabilized in the class under interpretation. The procedure can be repeated with the second most effective form in the class, then with the third and so on until a certain saturation occurs, that is, content repetition. The following is an example of the "concordancier" output.

****** *Applicants_APPLE INC *Kind_A1 *Publication Date_28/03/2019 *Application Number_US 201715717821 A *IPC_IPCABSENT**

score : 38678.42

the **slot antenna** may be **fed** via **near field coupling** using a **conductive patch** that is **located** within the **slot** at the **surface** of the **substrate** the **conductive layer rear housing wall** and **vertical portion** may **form** a **cavity** for the **slot antenna**

****** *Applicants_AMAZON TECH INC *Kind_B1 *Publication Date_10/07/2018 *Application Number_US 201514791708 A *IPC_IPCABSENT**

score : 35090.22

a **housing** for an **electronic** device **includes** a single **rear housing assembly coupled** to the **cover glass** of a **display assembly** the **rear housing assembly includes** a **metal rear chassis** with two **layers** of **injection molded material formed** on at least the **chassis side regions**

****** *Applicants_APPLE INC *Kind_B2 *Publication Date_26/04/2016 *Application Number_US 201414195130 A *IPC_IPCABSENT**

score : 34607.80

the **antenna structures** may **include conductive structures** such as **metal traces** on **printed circuits** or other **dielectric substrates internal metal housing structures** or other **conductive electronic** device **housing structures** a **main resonating element arm** may be **separated** from an **antenna ground** by an **opening**

****** *Applicants_APPLE INC *Kind_B2 *Publication Date_23/04/2019 *Application Number_US 201615008139 A *IPC_IPCABSENT**

score : 34044.12

flexible printed circuits with **ground traces** may bisect the **slot shaped opening** to **form** three **electrically isolated slots** each of which is **aligned** with a respective **cavity antenna** the **antennas** may have **antenna grounds formed** from **portions** of the **metal housing** and other **conductive structures**

****** *Applicants_APPLE INC *Kind_A1 *Publication Date_27/07/2017 *Application Number_US 201615008139 A *IPC_IPCABSENT**

score : 34044.12

flexible printed circuits with **ground traces** may bisect the **slot shaped opening** to **form** three **electrically isolated slots** each of which is **aligned** with a respective **cavity antenna** the **antennas** may have **antenna grounds formed** from **portions** of the **metal housing** and other **conductive structures**

****** *Applicants_APPLE INC *Kind_A1 *Publication Date_24/01/2019 *Application Number_US 201715655311 A *IPC_IPCABSENT**

score : 33652.01

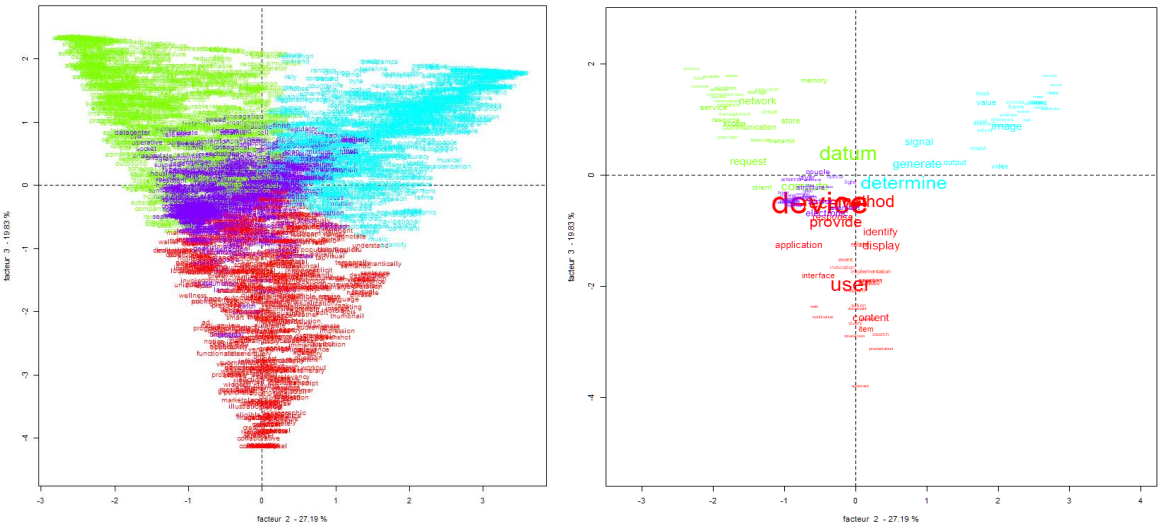
a **housing** made from a **circuit laminate includes** first and **second layers coupled** together each **includes** a **rigid electrically insulating** non **planar structural layer flexible conductive traces disposed** on **surfaces** of the **structural layer** and **flexible connector layers contacting** to the **flexible conductive traces**

After analyzing class 4, the same procedure must be performed for the other classes. When carrying out the analysis procedure highlighted above, it is verified that class 3 deals with image recognition and image transmission technologies, class 2 deals with technologies of the layers of connectivity services (data network) and services, such as components of software applications and interface with customers, such as back-end layers and class 1 deals with technologies related to content management and user experience in a social network context.

In fact, when observing the factorial analysis of correspondence, it can be seen that the center of the factorial axes (0,0) is where class 4 is located (common part of the technological choices of Apple, Google and Amazon). In CFA, the distances are Euclidean, so that the metric distances between shapes or classes are representative. The various observations are in figures 10, 11, 12 and 13.

Figure 10 a and 10 b

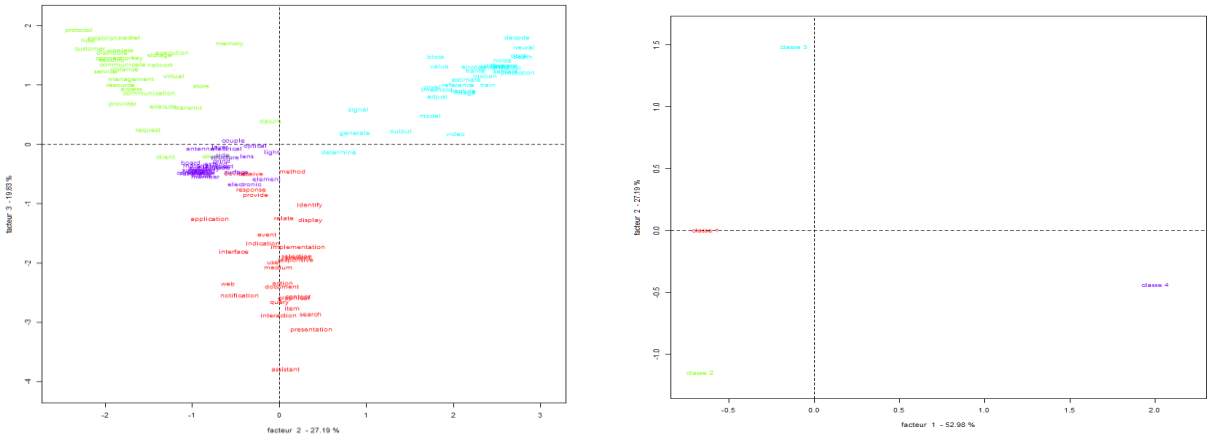
Visualization Refinement (Factor 1 and 2, 30 First Class Points, With Presentation Size Weighting Based on chi2): Factor Analysis of Correspondence Apple, Google and Amazon from 2016 to 2021. 19 Thousand Patents



Souce: Authors, 2022.

Figure 11 a and 11 b

Visualization refinement (Factor 1 and 2, 30 first points of the class, without frequency weighting and Euclidean distances between classes): Factorial analysis of correspondence Apple, Google and Amazon from 2016 to 2021. 19 thousand patents

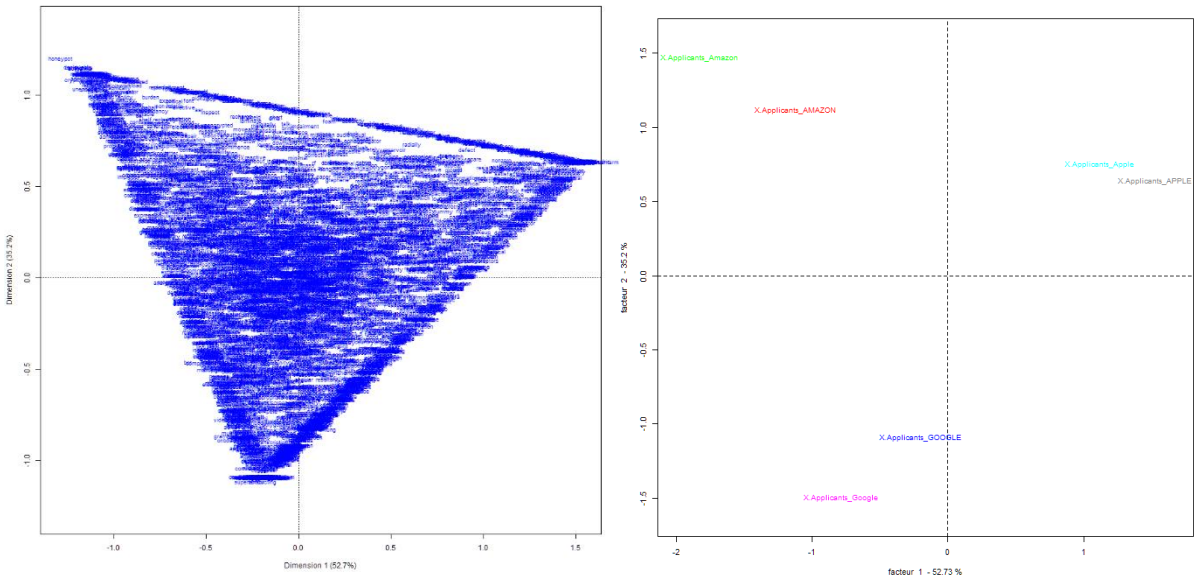


Source: Authors, 2022.

Specificity analyzes

Figure 12 a and 12 b

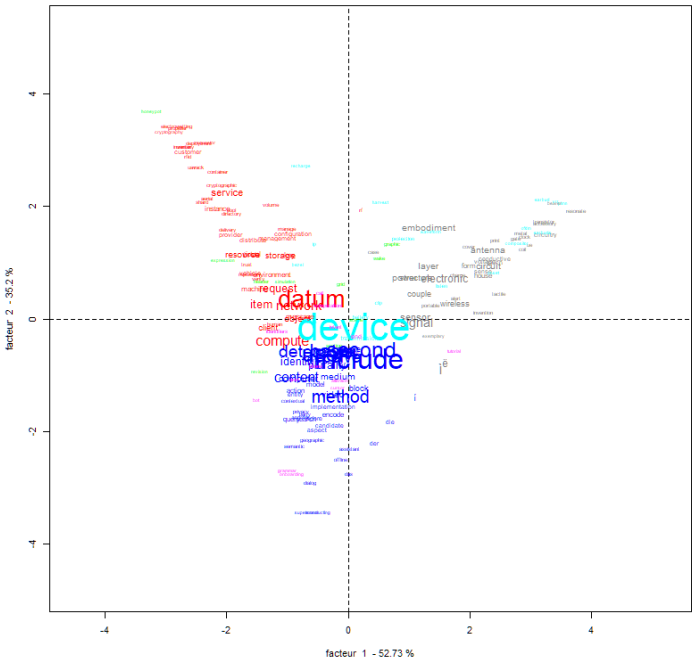
Specifics: Target variable "Holder". On the Left is The Visualization of The Row Data and on The Left The Column Data



Source: Authors, 2022.

Figure 13

Refinement of The Visualization (Factor 1 and 2, 30 First Class Points, With Presentation Size Weighting Based on Chi2)



Source : Authors, 2022.

Conclusion

At the beginning of this article we said that it is not possible to know intuitively whether the most innovative companies in the world make identical choices. This statement led us to analyze and discriminate the technological choices of the most innovative companies in the world. We asked some leading questions at the beginning of the survey to see how well we can respond using the CHD and CFA method. Do the world's most innovative companies make identical technology choices? It turned out not to be. Although there are common choices in class 4 between the three companies, there are more specific technological choices, as they appear in class 3, class 2 and class 1. As for the second question: If the choices are identical or similar, in what part are they similar? ? It was evident in this sampling that the technological choices are similar in relation to the content stored in semantic class 4, as they are core technologies shared by all sample holders:

> *"hardware projects, mainly involving modules where active elements (such as electronic circuits) and passive elements (such as antennas, lenses and supports) are fixed are technologies that support aesthetic (design) and functional (connections and systems of communication) aspects. mounting)"*

As previously explained, it makes sense that there is a common part, if we think that they are companies that compete in several business verticals, and this effect is related to the product portfolios in multiple verticals. Apple develops smartphones, but also services and even self-driving cars. Google develops quantum computing software, services and hardware. As for Amazon, there is a presence in retail sectors such as e-commerce, space and other areas. Therefore, even though there is discrimination regarding semantic classes, it is necessary to evolve in understanding each class before answering this question. As for the third question: If they make different choices, what are the core technologies and how are the different choices characterized? The distinct core technologies uncovered by the analysis can be structured into two groups. The group of core technologies not shared between holders in terms of dominance, which is the case of Google and Amazon:

> *"Google: Deals with technologies related to content management and user experience in a social network context"*

> *"Amazon: Deals with the technologies of the layers of connectivity services (data network) and services as components of software applications, in this case the processes involved in the software applications both at the interface with the customers and the back-end layers"*

The group of core technologies partially shared between holders in terms of predominance, which is the case of Apple and Google:

> *"Google and Apple: Addressing Image Recognition and Image Streaming Technologies"*

As for the last question: Is there technological overlap? It was evident that there is technological overlap, clearly observed in class 4. After analysis, it can be argued that the most innovative companies in the world have made different technological choices (class 1, 2 and 3). The common technology choices between Apple, Google and Amazon (class 4) indicate hardware and device technologies as the common target with Apple's predominance. It is known that Apple was the most involved in the business model based on the sale of hardware and this is evident when observing its predominance in the class that stores the choices common to all the holders analyzed. As a final suggestion, it is recommended that the articles that will

be written considering the demonstrated methodologies present one or more tables that summarize the analyses, as shown in table 3.

Decision-making for innovation management can be reviewed according to the findings of these analyzes that demonstrate the technological choices presented. If Apple and Google are interested in a particular market vertical, it will be important to observe the construction of the competitive advantage that has distinctiveness by nature. It is not necessarily a problem to master the same technologies, that is, companies can have similar technological domains and still have different marketing approaches. The issue is to reach awareness about where Research and Development resources are being invested, when we analyze the portfolio of patents that are the proxy for the technological choices of these companies. Does investment in hardware still make sense, that is, will allocating resources to build proprietary hardware technologies have an effect on competitive advantage? The analyzed choices are from the period from 2016 to 2019, therefore, recent. On the other hand, can reaching proprietary technologies in hardware be a strategy to change the business model of selling devices for rent or even devices in the form of a service? As we can see, by knowing the technological choices, we brought to the discussion only a small part of the possible contributions to the discussions of innovation management. This is certainly just the beginning.

The following are two summary tables; table 3 summarizes the step-by-step analysis and interpretation of the descending hierarchical classification and the correspondence factor analysis and table 4 presents the summary table model of the analysis of the textual data used in this article.

Table 3

Summary of the Analysis Plan

Preparation of the textual corpus	**(1)** Check for typing errors; **(2)** Exclude bold, italic, apostrophe, quotation mark, dollar sign, percentage, asterisk, comma and semi-colon; **(3)** Verbs with pronouns must be in proclisis form; **(4)** Replace hyphen with underscore.
Organization of manifest variables to be associated with the textual corpus	Line starts with four asterisks (****) followed by the variables introduced by an asterisk and the variable name, including an underscore (_) and the variable value. After the declaration of all manifest variables, include the text that will compose the textual corpus. Ex: **** *manifest variable name1_variable value *manifest variable name 2_variable value
Save the file in standard Unicode UTF-8 txt format	Save the file in standard Unicode UTF-8 txt format
Definition of Iramuteq interface parameters and the electronic file of the textual corpus to be imported	**(1)** Select "UTF-8 all languages encoding"; **(2)** Select the language in which the textual corpus is presented; **(3)** Select "default dictionary"; **(4)** Select text marking "****" **(5)** Select "use the dictionary of expressions"; **(6)** Select "Use text segments constructed through occurrences"; **(7)** Keep text segment length to 40 characters; **(8)** Select segment construction mode (number of characters or number of occurrences or size and punctuation criteria).
Import the file to Iramuteq	
Parameterization of analyzes	**(1)** Select lemmatization; **(2)** Select "Dictionary by indexing", except when using external dictionaries.
Selection of properties to define the treatment of different groups of words (analysis key)	**(1)** Mark active shapes as "1"; **(2)** Mark supplementary forms as "2"; **(3)** Mark shapes excluded from analysis as "zero".
Analyzes from Iramuteq	**(1)** Basic lexical analysis – verifies the frequency distribution of words in the corpus; **(2)** Similarity analysis – allows inferring the construction structure of the text from the co-occurrence between words; **(3)** Descending Hierarchical Sort (CHD) – "double sur srt", "simples sur segments de texte" or "simple sur textes" must be selected; **(4)** Factor Correspondence Analysis (CFA); **(5)** Specificity Analysis; **(6)** Word Cloud

Source: Authors, 2022.

Table 4

Summary of textual analysis

Applied Methodological Nomenclature	Class	Manifest Variable	text segments	AFC central position (0.0)	Summary of Content Analysis		
Textual analysis nomenclature	Semantic Class	Holder	Predominant Content	Corpus Common Vocabulary	Recombination		
Research objectives	Grouping of Technological Choices	Predominant Holder	Technological Choices		**Apple**	**Google**	**Amazon**
Analysis of the technological choices of the most innovative companies	1	Google	Deals with technologies related to content management and user experience in a social network context	No		Deals with technologies related to content management and user experience in a social network context	
	2	Amazon	It deals with the technologies of the layers of connectivity services (data network) and services as components of software applications, in this case the processes involved in the software applications both at the interface with the clients and the back-end layers	No			It deals with the technologies of the layers of connectivity services (data network) and services as components of software applications, in this case the processes involved in the software applications both at the interface with the clients and the back-end layers
	3	Apple e Google	Deals with image recognition and image transmission technologies	No	Deals with image recognition and image transmission technologies	Deals with image recognition and image transmission technologies	
	4	Apple	Hardware projects, mainly involving modules where active elements (such as electronic circuits) and passive elements (such as antennas, lenses and supports) are fixed are technologies that support aesthetic (design) and functional (connections and mounting systems) aspects)	Yes	Hardware projects, mainly involving modules where active elements (such as electronic circuits) and passive elements (such as antennas, lenses and supports) are fixed are technologies that support aesthetic (design) and functional (connections and mounting systems) aspects)	Hardware projects, mainly involving modules where active elements (such as electronic circuits) and passive elements (such as antennas, lenses and supports) are fixed are technologies that support aesthetic (design) and functional (connections and mounting systems) aspects)	Hardware projects, mainly involving modules where active elements (such as electronic circuits) and passive elements (such as antennas, lenses and supports) are fixed are technologies that support aesthetic (design) and functional (connections and mounting systems) aspects)

Source: Authors, 2022.

References

Ang, C. (2021, julho 19). *Ranked: The Most Innovative Companies in 2021*. Visual Capitalist. https://www.visualcapitalist.com/ranked-the-most-innovative-companies-in-2021/

Bardin, L. (1977). Content analysis. *São Paulo: Livraria Martins Fontes*.

Benzécri, J.-P. (1973). L'analyse des données, vol. 2. *Paris: Dunod*.

Camargo, B. V., & Justo, A. M. (2013). IRAMUTEQ: Um software gratuito para análisede dados textuais. *Temas em Psicologia*, *21*(2), 513–518. https://doi.org/10.9788/TP2013.2-16

Campion, E. D., & Campion, M. A. (2020). Using Computer-assisted Text Analysis (CATA) to Inform Employment Decisions: Approaches, Software, and Findings. *Research in Personnel and Human Resources Management*.

Cibois, P., & Jambu, M. (1981). Analyse des données et sociologie. *L'Année sociologique (1940/1948-)*, *31*, 333–348.

Hair, J. F. (2009). *Multivariate data analysis*.

Hirschfeld, H. O. (1935). A Connection between Correlation and Contingency. *Mathematical Proceedings of the Cambridge Philosophical Society*, *31*(4), 520–524. https://doi.org/10.1017/S0305004100013517

Mazieri, M. R. (2016). *Patentes e inovação frugal em uma perspectiva contributiva*. http://bibliotecatede.uninove.br/handle/tede/1600

Miraballes, M., & Gámbaro, A. (2018). Influence of Images on the Evaluation of Jams Using Conjoint Analysis Combined with Check-All-That-Apply (CATA) Questions. *Journal of food science*, *83*(1), 167–174.

Miraballes, M., Hodos, N., & Gámbaro, A. (2018). Application of a pivot profile variant using CATA questions in the development of a whey-based fermented beverage. *Beverages*, *4*(1), 11.

Ratinaud, P., & Marchand, P. (2012). Application de la méthode ALCESTE à de" gros" corpus et stabilité des" mondes lexicaux": Analyse du" Cable-Gate" avec IraMuTeQ. *Actes des 11e Journées internationales d'Analyse statistique des Données Textuelles. JADT 2012*.

Reinert, M. (1990a). Une méthode de classification des énoncés d'un corpus présentée à l'aide d'une application. *Les cahiers de l'analyse des données*, *15*(1), 21–36.

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

Reinert, M. (1995). Quelques aspects de choix des unités d'analyse et de leur contrôle dans la méthode Alceste. *JADT1995*, *1*, 27–34.

Reinert, M. (2007). Postures énonciatives et mondes lexicaux stabilisés en analyse statistique de discours. *Langage et société*, *3*, 189–202.

Short, J. C., Broberg, J. C., Cogliser, C. C., & Brigham, K. H. (2010). Construct validation using computer-aided text analysis (CATA) an illustration using entrepreneurial orientation. *Organizational Research Methods*, *13*(2), 320–347.

Tidd, J., & Bessant, J. (2015). *Gestão da inovação-5*. Bookman Editora.

van Meter, K. M., Mounier, L., Chartron, G., & Reinert, M. (1991). Multimethod Analysis: Official Biographies of Members of the Central Committee of the Soviet Union Communist Party. *Bulletin of Sociological Methodology/Bulletin de Méthodologie Sociologique*, *33*(1), 20–37. https://doi.org/10.1177/075910639103300102

Recebido: 13 dez. 2022/**Aprovado**: 29 dez. 2022
Editores responsáveis: Julio Araujo Carneiro da Cunha & André Torres Urdan
Processo de Avaliação: *Double Blind Review*
e-ISSN: 2177-5184
https://doi.org/10.5585/remark.v21i5.21290

ARTIGO TUTORIAL

USO DO IRAMUTEQ PARA ANÁLISE DE CONTEÚDO BASEADA EM CLASSIFICAÇÃO HIERARQUICA DESCENDENTE E ANÁLISE FATORIAL DE CORRESPONDÊNCIA

Marcos Rogério Mazieri
Universidade Nove de Julho (Uninove)
São Paulo, SP - Brasil.
marcosmazzzieri@gmail.com

Luc Marie Quoniam
Université du Sud Toulon-Var
La Garde, Provence-Alpes-Côte d'Azur, FR - França
mail@quoniam.info

David Reymond
Université de Toulon - Var
La Garde, Provence-Alpes-Côte d'Azur, FR - França
dreymond@univ-tln.fr

Katia Cinara Tregnago Cunha
Universidade Nove de Julho (Uninove)
São Paulo, SP - Brasil.
katiapatentes@gmail.com

Objetivo: Apresentar a análise de conteúdo baseada em classificação hierárquica descendente e a análise fatorial de correspondência como técnicas complementares e sequenciais, com possível aplicação na área de gestão da inovação

Método: Análise de conteúdo baseada em técnicas de *Computer-Aided Text Analisys* (CATA) por meio do software IRAMUTEQ.

Originalidade/Relevância: O paradigma de pesquisa, normalmente influencia as próprias bases de conhecimento e metodologias de domínio do pesquisador. A análise de textos manual pode eventualmente ser atribuída ao paradigma interpretativista e as técnicas CATA podem ser eventualmente atribuídas ao paradigma pós-positivista, no entanto, não parece haver razão para fazer essa distinção.

Resultados: Apresentação de framework que demonstra as escolhas tecnológicas das empresas mais inovadoras do mundo (Google, Apple e Amazon), as escolhas comuns e as distintas, entre elas.

Contribuições teóricas/metodológicas: Desenvolvimento de método de análise de escolhas tecnológicas das empresas mais inovadoras do mundo, aplicando a análise de conteúdo com base na classificação hierárquica descendente e a análise fatorial de correspondência de forma sequencial e complementar.

Contribuições sociais / para a gestão: As tomadas de decisões para a gestão da inovação podem ser revistas de acordo com as escolhas tecnológicas apresentadas. A vantagem competitiva tem por natureza a distintividade. Demonstramos que não é um problema dominar as mesmas tecnologias, ou seja, empresas podem ter domínios tecnológicos similares e ainda assim terem abordagens mercadológicas diferentes.

Palavras-chave: Análise de conteúdo. Classificação hierárquica descendente. Análise fatorial de correspondência. Patentes. Gestão da inovação. Transformação digital. Iramuteq.

Como citar

American Psychological Association (APA)

Mazieri, M. R., Quoniam, L. M., Reymond, D., & Cunha, K. C. T. (2022, out./dez.). Uso do iramuteq para análise de conteúdo baseada em classificação hierarquica descendente e análise fatorial de correspondência. *Brazilian Journal of Marketing*. 21(5) 2012-2048. https://doi.org/10.5585/remark.v21i5.21290.

Introdução

O objetivo geral desse artigo é apresentar a análise de conteúdo baseada em classificação hierárquica descendente e a análise fatorial de correspondência como técnicas complementares e sequenciais, com possível aplicação na área de gestão da inovação. O problema de pesquisa que se pretende usar para exemplificação do uso dos métodos é baseado nos desafios envolvidos na compreensão das escolhas tecnológicas das empresas consideradas como as mais inovadoras do mundo. Não é possível saber intuitivamente se as empresas mais inovadoras do mundo fazem escolhas idênticas caso façam, em que parte elas são parecidas. Caso façam escolhas diferentes, quais são as tecnologias centrais e como se caracterizam as diferentes escolhas tecnológicas? Há sobreposição tecnológica? O objetivo de pesquisa que pode ser declarado para esses problemas é: Analisar e discriminar as escolhas tecnológicas das empresas mais inovadoras do mundo. A Análise de conteúdo é a técnica usada para discriminar grupos temáticos homogêneos num texto, que passam a ser organizados em classes. Cada classe absorve os temas mais similares, de forma que, ao final do procedimento haverá tantas classes quanto a quantidade de temáticas discriminadas dentro de um certo texto ou conjunto de textos. Os procedimentos de formação das classes temáticas para análise de dados textuais podem ser realizados manualmente (Bardin, 1977) ou com a ajuda de computadores (Short et al., 2010; Miraballes et al., 2018; Miraballes & Gámbaro, 2018; Campion & Campion, 2020). Para avaliar a análise de conteúdo baseada em técnicas de *Computer-Aided Text Analisys* (CATA) pretendemos explicar como podem ser matematicamente realizados os processamentos de dados textuais, com vistas à área de ciências sociais aplicadas na área de gestão da inovação para facilitar as etapas posteriores de interpretação de informações a ser realizada pelo pesquisador. Há alguma controvérsia entre a visão interpretativista e a visão pós-positivista sobre o uso de técnicas CATA que iremos comentar, em boa parte usando o pressuposto de que a análise de texto é um processo sistemático que não é doutrinal nem normativo. Análise de conteúdo não é sobre encontrar algo encoberto pelo texto, mas sim o que o texto expressa e essa é uma informação explicita (Bardin, 2009). Estamos argumentando nesse artigo que as técnicas de CATA podem contribuir tanto para pesquisas interpretativistas quanto com pesquisas pós-positivistas. Para isso, passaremos a apresentar o funcionamento de alguns dos algoritmos de análise de texto mais comuns.

As técnicas responsáveis pelo processamento automatizado são inúmeras. No entanto, foi na escola de análise de dados francesa que as técnicas de análise de texto, como um tipo de

dado, se evidenciaram e por isso tais técnicas foram escolhidas como parte do objeto de análise desse artigo. As principais técnicas usadas na escola francesa de análise de dados são originárias da matemática aplicada, especialmente a classificação hierárquica descendente (CHD) e a análise fatorial de correspondência (AFC). A AFC é conhecida desde a década de 1930, verificada no trabalho de Hirschfield (Hirschfeld, 1935), quando era chamada de "tratamento de dados sem média". A técnica da AFC passou a ser usada especialmente por pesquisadores franceses para análise de dados textuais depois que o trabalho de Jean-Paul Benzécri de 1973 (Benzécri, 1973) foi publicado, cujo principal objetivo era obter representações gráficas das linhas e das colunas de uma tabela de contingência 2 x 2. O uso da técnica de AFC na área de ciências sociais foi intensificado na década de 1980 por trabalhos que mostravam casos de uso em sociologia, antropologia e psicologia (Cibois & Jambu, 1981). A Classificação Hierárquica Descendente também é uma técnica estatística conhecida desde 1950, porém, foi a partir do trabalho de Máx Reinert (Reinert, 1990a, 1990b, 1995, 2007) que o uso técnica de análise usando CHD foi intensificada nas áreas das ciências sociais.

Na área das ciências sociais aplicadas, especialmente na área de gestão, há projetos de pesquisa que necessitam da análise de dados textuais para analisar o conteúdo de determinado conjunto de dados. Obviamente como em todas as áreas do conhecimento, o aumento da capacidade de processamento, redução de custos de armazenamento e a disponibilidade de aplicativos de software facilitaram o uso destas técnicas e removeram as barreiras técnicas, de certa forma incentivaram os pesquisadores e praticantes à usar análise de conteúdo automatizada, no entanto, o que numa mão trouxe acesso e facilidade para aplicar as técnicas de análise em outra mão tornou alguns resultados de análise herméticos, de compreensão parcial e em alguns casos de compreensão equivocada. A questão é que pesquisadores das áreas de ciências sociais naturalmente não são especialistas léxicos, linguísticos, nem profissionais da área de matemática, tampouco engenheiros de computação, ainda que precisem discutir sobre essas áreas de conhecimentos mencionadas para poder projetar e executar uma análise de conteúdo auxiliada por computador, cujo termo é conhecido por CATA. Há ainda as diversas vertentes de estudos que podem ser originadas na análise textual; análise de conteúdo, análise semântica e análise de discurso. Análise de discurso, consiste em examinar as construções ideológicas dos produtores dos textos, onde cada texto traz refletida a visão de mundo dos seus produtores e, portanto, esse artigo não trata de análise de discurso. Análise semântica consiste em examinar as características léxicas dos textos para buscar o entendimento sistemático de conceitos e informações com base na constituição da semântica e do léxico do idioma e,

portanto, esse artigo trata de análise semântica. A análise de conteúdo, consiste em examinar as classes temáticas que podem ser sistematicamente agrupadas com interpretação induzida por teoria, por experiência ou ambas e esse artigo tratará de análise de conteúdo no campo da gestão da inovação, especificamente na transformação digital.

Devo destacar que análise textual não é sinônimo de análise de conteúdo diretamente. Análise textual é a organização dos dados contidos no texto. O pressuposto da análise textual é de que os textos são formados por dados, simbolizados por toda a ortografia, gramática e demais estruturas linguísticas no nível mais baixo, que chamamos de nível léxico. A combinação dos dados do texto produz as informações contidas no texto que não apresentam significado, situando-se no nível médio, que chamamos de nível textual. O encadeamento das informações presentes no texto segue uma lógica que produz sentido ou significado, de acordo com a experiência do pesquisador, num nível alto que chamamos de nível de conteúdo. Textos são representações léxicas da linguagem natural verbal transcrita, como também são a representação léxica da cognição humana em forma de texto. A análise léxica de um texto pode ser obtida por meio do desdobramento das formas do texto (palavras), com o objetivo de permitir classificar tais formas que o compõe para gerar informação derivada dos dados do texto, mas que não estão visíveis na forma natural da leitura humana. Nesse artigo, palavra e forma foram consideradas sinônimas, uma vez que o algoritmo usado na etapa empírica dessa pesquisa usa essa nomenclatura. A análise de conteúdo é a interpretação das informações encontradas no texto, orientadas por teoria ou conhecimento empírico, portanto, sob a abordagem indutiva e *data-driven* ou guiadas por proposições *ex ante*.

Devido à multidisciplinariedade envolvida em uma análise textual auxiliada por computador, compreender melhor os procedimentos matemáticos específicos que são a base dos algoritmos de processamento dos dados textuais pode tornar o nível de análise de conteúdo mais familiar aos pesquisadores, tanto interpretativistas como os pós-positivistas. Na área de ciências sociais aplicadas, a análise de conteúdo tem sido realizada com o auxílio de consagrados e competentes protocolos de análise de conteúdo manual, descritos por diversos pesquisadores, e em especial da Bardin (Bardin, 1977). A análise de conteúdo manual segue o pressuposto de que o pesquisador tem a habilidade de encontrar os conteúdos nos textos, realizando manualmente a classificação das formas textuais, procurando conjunções e disjunções de partes do texto. A técnica é tradicional e usada há muitos anos e é dependente da experiência e da energia do pesquisador para classificar todas as informações textuais. Para quantidades maiores de texto, como cem ou duzentos textos de uma página formato A4, pode

ser ainda mais desafiador para o pesquisador, tanto pelo tempo e energia a ser dispendido na análise, como pela necessidade de manter a estabilidade dos critérios de julgamento de classificação estabelecido pelo pesquisador no protocolo de análise de conteúdo.

O paradigma de pesquisa, normalmente influencia as próprias bases de conhecimento e metodologias de domínio do pesquisador. A análise de textos manual pode eventualmente ser atribuída ao paradigma interpretativista e as técnicas CATA podem ser eventualmente atribuídas ao paradigma pós-positivista, no entanto, não parece haver razão para fazer essa distinção. Esse pensamento pode ser um dos mal-entendidos que afastam os pesquisadores que são interpretativistas das técnicas CATA, ao mesmo tempo que induzem os pesquisadores que são pós-positivistas a análises parciais, superficiais ou equivocadas quando usam as técnicas CATA. Na análise textual auxiliada por algoritmos, todo o processamento da classificação das formas, envolvendo a obtenção de tabelas conjuntivas, disjuntivas são realizadas por meio de técnicas quantitativas, usando testes de Chi2. Evidentemente, essa análise alcança apenas os níveis léxico e textual, mas não o nível de conteúdo. O nível de conteúdo é perceptual e nesse caso, é mais útil para a produção de conhecimento quando realizado pelo pesquisador. Argumentamos que técnicas CATA podem ser usadas tanto no paradigma de pesquisa interpretativista quanto no paradigma pós-positivista, uma vez que a automação envolvida na análise por técnicas CATA interfere nas camadas do nível léxico e textual, sem prejuízo ao nível de conteúdo.

Referencial teórico

Nível léxico, textual e de conteúdo

Léxico é o nome do conjunto de palavras em um determinado idioma. No caso de uma análise textual, refere-se ao conjunto de formas presentes no texto ou no conjunto de textos em análise. Os desafios linguísticos de deslocamento do sentido das formas ao longo do tempo devem ser considerados, visto que é uma propriedade inerente ao acervo linguístico transmitido entre gerações. Portanto, análise de textos produzidos em intervalo de tempo maior que 15 ou 20 anos, entre o texto mais antigo e o texto mais novo do conjunto de textos, precisam ser realizados de forma a controlar os períodos, repartindo os textos por subperíodos, por exemplo. Léxico e vocabulário não são sinônimos, sendo o léxico o formador dos vocabulários de um idioma. O léxico é composto de formas semânticas agrupadas em classes de formas, de acordo com a gramática de cada idioma. Substantivos, adjetivos e verbos são classes de formas que

existem em muitos idiomas. Análise de nível léxico consiste em separar as classes de formas, entre verbos, substantivos, adjetivos, não exaustivamente, formando extratos de cada classe de formas com suas frequências de ocorrências no texto ou conjunto de textos. Esse procedimento é realizado após a lematização. Lematização é a técnica de normalização das palavras do texto, transformando tempo verbal, gênero e número. Normalmente a lematização leva os verbos para o infinitivo e os substantivos para o singular e masculino. Ao passar pelo processo de lematização, as palavras passam a ser chamadas de formas quando únicas e ocorrências quando consideradas todas as repetições tornando possível comparar as formas dentro de um conjunto de textos. Para fins de definição operacional na área de gestão da inovação, argumentamos nesta pesquisa que o nível léxico é o acervo linguístico aplicado para a descrição do novo.

No nível textual, são aplicadas as técnicas de classificação hierárquica descendente CHD e a análise fatorial de correspondência AFC. A CHD é um procedimento que envolve uma série *n*-1 decisões de agrupamento, sendo *n* o número de formas em uma estrutura do tipo "árvore". Os dois tipos básicos de agrupamento são os aglomerativos e os divisivos. Nos métodos aglomerativos, cada observação começa como seu próprio agrupamento e vão agrupando de dois em dois, consecutivamente. O tipo de agrupamento divisivo começa por um único agrupamento que se divide em dois, e cada um dos agrupamentos formados em mais dois consecutivamente até que cada agrupamento seja um agrupamento unitário. Trata-se, portanto, de um procedimento iterativo persistente até a estabilização dos agrupamentos unitários. A natureza do algoritmo hierárquico que acabamos de explicar, com base no prof. Joseph Hair (Hair, 2009), torna mais claro que a classificação das formas é realizada *ex post*. É a partir das palavras lematizadas do texto ou do conjunto de textos que são realizados os agrupamentos das chamadas formas correlacionadas, delimitando as classes temáticas que são divididas de acordo com os testes de hipótese significativos ($p<0,05$, *h0 – a forma pertence a outra classe*). Testes de Chi2 podem ser usados para testar essa hipótese, usando a frequência da forma *i* dentro do conjunto de textos *j*. Ao dividir o conjunto de textos em duas partes, a forma *i*, terá maior score Chi2 em uma das duas partes da divisão do conjunto de textos e assim sucessivamente até que a forma *i* estabilize em um agrupamento final, chamada classe temática ou classe semântica. O procedimento será repedido com todas as diferentes formas *i* que compõe o conjunto de textos *j*. Portanto, as classes temáticas não são definidas *ex ante*, permanecendo dessa forma completamente isoladas de indução, seja da experiência do pesquisador, seja do próprio interesse teórico.

Consideramos que o trabalho de Max Reinert representou relevante evolução da classificação hierárquica devido a abordagem de classificação que superou a análise de frequência das formas, mas inovou com a ideia de classificar os segmentos de texto ST (Reinert, 1990a, 1990b, 1995, 2007; van Meter et al., 1991). Ainda que a classificação hierárquica seja o procedimento estatístico geral para classificação de dados, a aplicação dentro dos ST representa o avanço relevante para a qualidade da classificação com repercussões para o nível de conteúdo e até mesmo tornando mais robusta a ideia de análise semântica. Segmentos de texto são recortes de cada texto de tamanho uniforme. Ao invés de analisar a frequência das palavras no formato *i x j*, o algoritmo proposto por Reinert muda o nível de análise anteriormente a partir da forma para o nível de segmento de texto ST, partindo do pressuposto que cada pedaço de texto que é escrito dentro de um espaço de 40 ou 50 letras, aproximadamente 10 a 15 palavras, tem a probabilidade de conter uma ideia. A hipótese passa a ser testada com base no score Chi2 do ST, comparados com os demais ST do conjunto de textos. Os agrupamentos irão se estabilizar, constituídos de *n* ST homogêneos contextualmente, usando a lógica hierárquica divisiva que é descendente. Essa implementação da visão baseada em ST propiciou a metáfora usada por Reinert do "mundo léxico", representado pelos agrupamentos dos ST de contexto homogêneo, tornando a ligação entre o nível léxico e o nível textual mais consistente. Facilita também a diferenciação entre os algoritmos que aplicam a frequência de palavras daqueles que aplicam técnicas adicionais. No caso do algoritmo de Reinert, o nome dado foi Alceste, acrônimo para Análise Léxica de Contexto por Conjunto de Segmento de Texto. Para fins de definição operacional na área de gestão da inovação, argumentamos nesta pesquisa que o nível textual é o vocabulário frequente aplicado para a descrição do novo e uma classe semântica pode ser um domínio inovativo. O conteúdo dos domínios inovativos pode ser composto por tecnologias, processos, funções, aplicações ou práticas, dependendo da fonte de dados usadas na análise. Por exemplo, ao analisar *full text* de patentes, estarão presentes nos domínios inovativos a função, aplicação ou o processo das tecnologias depositadas nas metabases de patentes; ao analisar resumos de artigos científicos sobre inovação, estarão presentes nos domínios inovativos o contexto de pesquisa, o objetivo da pesquisa (reflexo de parte da problemática), os métodos mais usados e as descobertas e; ao analisar entrevistas transcritas sobre inovação, estarão presentes nos domínios inovativos práticas e desempenho da inovação, evidências das características culturais do ambiente ou da experiência do respondente, neste caso de acordo com os fundamentos do roteiro de entrevistas. O domínio inovativo não é autônomo ou autoexplicativo. O domínio inovativo é uma

delimitação racional entre os diversos domínios inovativos que se formam, depois da classificação hierárquica descendente. A análise fatorial de correspondência é a técnica popularizada pelos trabalhos de Benzecri e Cbois nas ciências sociais aplicadas (Benzécri, 1973; Cibois & Jambu, 1981).

Aplicação de análise de conteúdo em gestão da inovação

Como muitos temas das ciências sociais aplicadas, inovação é de escopo amplo, bem como a gestão da inovação. As dificuldades de pesquisa com abordagem quantitativa inerente aos temas de escopo amplo são reconhecidas no campo da inovação. Estão relacionadas especialmente à dificuldade de obter definições conceituais ou operacionais, delimitação de conceitos, ou seja, até que ponto alcança um conceito e começa outro e os problemas de identificar fluxo nos sistemas de construtos, que posso simplificar dizendo: dificuldades em compreender quais construtos são antecedentes, quais são intervenientes e quais são consequentes. Quanto às abordagens qualitativas, o escopo amplo também impõe desafios de delimitação conceitual e de obtenção de definições conceituais e operacionais ou mesmo na definição e operacionalização de variáveis qualitativas (categóricas ou nominais, por exemplo). Também na abordagem qualitativa impõe-se a necessidade de grande experiência do pesquisador para fazer as escolhas e tomar as decisões sobre a "conversação teórica" sobre a qual a pesquisa será assentada, e quais as variáveis podem envolver proposições definidas *ex ante*. Elementarmente não é necessária a definição de proposições *ex ante* para que haja uma pesquisa qualitativa, mas os pesquisadores podem optar em projetar a pesquisa desta forma. Se a opção do pesquisador for pela definição de proposições *ex ante*, as variáveis explicitadas nestas proposições podem ser incluídas no algoritmo de análise e, portanto, poderão ser interpretadas e discutidas para fins de fortalecer ou eliminar as proposições *ex ante*.

Para fins de constituir alguma sistematização da abordagem da gestão da inovação, recorremos a um livro clássico sobre gestão da inovação de Joe Tidd e John Bessant, neste momento em sua 5ª. Edição (Tidd & Bessant, 2015). A escolha deve-se ao reconhecimento dessa obra pelos pesquisadores da área de inovação, bem como o seu uso intensivo em cursos de graduação e pós-graduação. Segundo a compilação oferecida por Tidd e Bessant, há quatro grandes etapas envolvidas na gestão da inovação: busca por inovações, seleção de inovações, implementação de inovações e captura de valor. Nessa pesquisa, a ideia é apresentar algumas sugestões para usar análise de conteúdo para melhorar a compreensão da intersecção entre as

subáreas da implementação de inovações e captura de valor e, para isso, usaremos as metabases de informações sobre patentes.

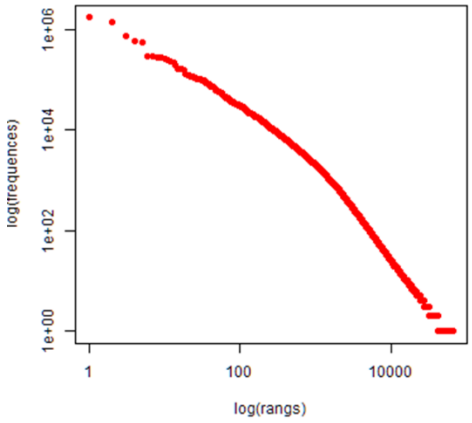
Metabase de patentes

Patentes são títulos de propriedade industrial limitadas ao tempo e a localização geográfica. Há mais de 140 milhões de pedidos de patentes e, por força de lei, devem apresentar o resumo em inglês ou francês, além da língua de origem. Todas as patentes recebem uma numeração de classificação internacional de patentes (IPC) que é a identificação de sua patenteabilidade, ou seja, toda a matéria patenteável pode ser descrita por números IPC. Além das conveniências descritas acima, as metabases de patentes são de acesso público e gratuito. As informações de patentes vêm sendo usadas como proxy dos esforços de inovação ou ainda como resultado dos esforços de inovação. Também há grande concordância de que a patente não é uma inovação realizada, mas uma inovação potencial. Por isso argumentamos que ao alcançar o portfólio de patentes de uma empresa, podemos analisar tal portfólio como esforços de gestão da inovação que já ultrapassaram a fase de busca e seleção. Argumentamos que se foi depositada uma patente, tal depósito é uma parte da implementação da inovação, haja visto os custos e recursos empenhados no depósito de uma patente. As ideias sobre a elegibilidade ou não das bases de patentes como parte da estratégia de inovação baseia-se no fato de pequenas empresas não depositarem patentes ou de determinados países menos desenvolvidos não depositarem patentes. No caso desta pesquisa, ambas restrições não se aplicam devido a escolha das empresas mais inovadoras do mundo. As 10 empresas mais inovadoras do mundo concentram cerca de 560 mil patentes das 140 milhões. Apenas para fins de compreensão, durante a realização desta pesquisa, acessamos as 560 mil patentes usando o software free Lens (lens.org) escolhido por não ter custo e pela extrema simplicidade de uso por alunos e profissionais interessados, o que torna possível e simples a replicabilidade deste estudo. Foram escolhidas as patentes que tinham os seus devidos resumos e, neste caso, foram 245 mil patentes. Outras ferramentas de recuperação de patentes podem aumentar a taxa de resumos recuperados mas, no caso desta pesquisa, restará claro que temos amostra maior do que o propósito de aplicação do método (245mil patentes, 2,17 milhões ocorrências, 66mil formas únicas e 23mil hapax (0,11% das ocorrências e 35,52% das formas). A média de 88,75 ocorrências por texto demonstra a natureza textual da amostragem. Na próxima figura apresentamos a relação entre rang (quantidade de categorias de palavras) e a frequência das

categorias, ambas com aplicação de logaritmo. Mesmo com a aplicação do logaritmo, verificamos perto de 20mil rang no extremo do eixo x e 1 x 10^6 ou frequência superior a 1 milhão. Nesses patamares de amostragem não é simples manipular os dados com computadores e planilhas mais domésticas devido ao tempo (de algumas horas para processar até dias para cada revisão). Portanto, o custo de computação e de intervenção manual de preparação de dados pode não se justificar. Para esta pesquisa, realizamos os procedimentos para informar a necessidade de reflexão quanto ao tamanho de amostra a ser considerado.

Figura 1

Análise léxica básica



Fonte: Autores, 2022.

Em termos de amostra mínima, há autores que recomendam 30 textos, portanto, o limite mínimo aproximado é conhecido. Quanto ao limite máximo, são considerados corpus de grande porte os que tem mais de 2 milhões de ocorrências. A questão é que não são bases de dados maiores não significam melhor pesquisa. Após análise de mais de 30 milhões de ocorrências entre 2013 e 2022, algumas conclusões empíricas podem ser compartilhadas neste artigo. Há análises textuais no contexto microm, meso e macro. Por contexto de análise de texto Micro considera-se, empiricamente, o corpus que não pode ser lido ou analisado manualmente; a partir de 50 textos. Menos do que 50 textos, podem alcançar melhores resultados se analisados manualmente para maior detalhamento. Os pressupostos matemáticos envolvidos nas técnicas de classificação ou de fatoração consideram matrizes compostas pela presença ou ausência das formas e, ainda que amostras pequenas como as compostas por 30 textos possam funcionar, em

termos de operação do software de análise, podem frustrar pela superficialidade dos resultados que oferece em termos semânticos. Corpus formados com quantidades entre 50 e 20 mil textos podem ser considerados contextos de análise de texto Meso que é quando podemos expor o conteúdo a ser interpretado de forma mais eficaz. A partir de 500 mil textos ou 2 milhões de ocorrências (caso demonstrado neste artigo) o poder de discriminação dos segmentos de texto é reduzido, chegando até mesmo a inviabilizar uma análise direta, o que chamamos de contexto de análise de texto Macro. A análise de corpus maiores que 500mil textos ou 2 milhões de palavras deve usar extensões das técnicas descritas neste artigo, como a clusterização prévia e técnicas de subamostragens de testes e subamostragem de aprendizagem quando o ferramental de hardware e software for simples (processamento de computador pessoal). No caso desta pesquisa, quando processamos 2,3 milhões de ocorrências, a Classificação Hierárquica Descendente não conseguiu classificar ao menos 3 classes semânticas, o que elimina a possibilidade de analisar hierarquicamente duas classes, não tendo sentido ou utilidade prática. O efeito prático é que os softwares de análise não chegam ao final, inviabilizando a análise por CHD. Destacamos que esse comportamento é específico para a amostragem atual devido aos vocabulários envolvidos. A questão é que quanto maior a amostra, maior a chance de homogeneização do conteúdo textual e redução do poder de discriminar as classes devido aos princípios matemáticos envolvidos para resolver as matrizes de presença e ausência das palavras, o que acaba sendo uma combinação binária.

Procedimentos metodológicos

A seguir, será detalhado a obtenção de uma base com mais de 2 milhões de palavras. A pretensão desta seção é explicitar as escolhas de amostra, a delimitação do tamanho de amostra e os principais parâmetros a serem considerados para realizar a classificação hierárquica descendente e a análise fatorial de correspondência.

Sabemos que a escolha de uma ferramenta em detrimento às outras ou de uma técnica em detrimento às outras é sempre motivo de grande e bom debate. Para essa pesquisa, como para as demais que conduzimos, temos sempre como critério de escolha principal o custo e o acesso às ferramentas e softwares. Estamos inclinados a realizar pesquisas com a responsabilidade social de permitir que um aluno ou profissional que esteja em uma região mais remota possa conseguir replicar e dar continuidade ao estudo. Quanto à acessibilidade, consideramos dois pontos. Primeiro se o software ou ferramenta pode ser acessada pela internet,

baixada, instalada e usada sem necessidade de grande especialização. Segundo, se o código é aberto, que permite compreender os algoritmos ou até mesmo corrigir ou ampliar o código. A ferramenta que alcança estas características é Iramuteq (Camargo & Justo, 2013; Ratinaud & Marchand, 2012). O Tropes é de código aberto e sem custos para baixar e instalar, porém tem uma curva de aprendizagem mais lenta que o Iramuteq. Já todos os softwares mais famosos no Brasil, como o Atlas TI, Maxqda e Nvivo são softwares que têm uma versão de testes sem custo e depois são oferecidas versões pagas, o que nos leva a não escolher tais softwares por custo, além do que, pela natureza comercial, não são de código aberto, segundo ponto que não atende aos critérios estabelecidos em nossas pesquisas. Obviamente todos são bons produtos dentro de suas características, mas não atendem aos critérios de escolha estabelecidos por essa pesquisa.

A escolha das empresas a serem analisadas foi baseada no Ranking das 50 empresas mais inovadoras do Boston Consulting Group. O Boston Consulting Group analisou 1500 empresas com base nos critérios de Global MindShare, Industry Peer Review, Industry Disruption e Value Creation. Global Mindshare, consiste na consistência de resposta dos executivos respondentes a respeito da inovação, respondendo sobre suas próprias empresas. Industry peer review consiste no voto dos executivos a respeito da inovação das empresas que não as deles próprios. Industry disruption consiste num índice que mensura votos sobre as indústrias. Criação de valor é o total de retorno sobre ações dos últimos três anos, de dezembro de 2017 à dezembro de 2020 (Ang, 2021). A seguir, é apresentada a tabela 1 com as 50 empresas mais inovadoras de 2021, segundo a Boston Consulting Group.

Tabela 1

10 Empresas Mais inovadoras de 2021 Segundo a Boston Consulting Group

2021	Empresa	Industria	Headquarter	2020
1	Apple	Technology	US U.S.	--
2	Alphabet	Technology	US U.S.	--
3	Amazon	Consumer Goods	US U.S.	--
4	Microsoft	Technology	US U.S.	--
5	Tesla	Transport & Energy	US U.S.	6
6	Samsung	Technology	KR South Korea	-1
7	IBM	Technology	US U.S.	1
8	Huawei	Technology	CN China	-2
9	Sony	Consumer Goods	JP Japan	--
10	Pfizer	Healthcare	US U.S.	Return

Fonte: Autores, 2022 baseado em Boston Consulting Group.

Para essa pesquisa, foram analisadas as 10 empresas mais inovadoras do mundo, que representam 560 mil patentes. Conforme exposto anteriormente, nessa amostragem, é inviabilizada a classificação hierárquica descendente.

Sobre o problema do "Leg Temporal" e escolha de amostra

Normalmente, os pesquisadores da área de inovação têm como desafio adicional estabelecer a distância entre os esforços de inovação e a inovação propriamente dita. Há muitos trabalhos que propõe as mais diferentes maneiras de tratar essa distância.

Nessa pesquisa, definimos considerar o tempo de três anos entre esforço e inovação efetiva. O esforço de inovação será identificado nos depósitos de patentes das empresas mais inovadoras do mundo que tenham ocorrido entre 2016 e 2019 para estar refletido em inovação efetiva em 2021 e 2022. A inovação efetiva ou realizada não será objeto de análise nesta pesquisa.

O critério de escolha da amostra deve seguir o alinhamento teórico envolvido na pesquisa. No caso desta pesquisa, a escolha é por julgamento, estabelecido no ranking das 50 empresas mais inovadoras e especificamente observando as 10 mais bem colocadas. Das 10 mais bem colocadas, decidimos analisar as três primeiras empresas do ranking - Apple, Alphabet/Google e Amazon, considerando o limite de análise textual Meso (20 mil textos) a soma das patentes destas três empresas entre 2016 e 2019 que é de aproximadamente 19 mil patentes.

Expressão de busca

Normalmente buscar as palavras em Dicionários, Thesaurus e Revisão de Literatura podem ajudar a encontrar palavras-chave do ponto de vista linguístico etimológico ou sinonímico. A curva de energia (frequência de palavras) é inversa à curva da saturação (frequências acumuladas). A População, por definição, é desconhecida. A justificativa da amostra parte da análise linguística etimológica, sinonímia, mas ainda não é suficiente pois o depositante pode usar na descrição dos campos das patentes alvo da busca outros termos que não os descobertos na fase de definição da expressão de busca.

Esses termos podem estar relacionados a vícios de linguagem ou variações culturais e demais diferenças semânticas. No caso desta pesquisa, dada a busca de patentes de determinados titulares ou aplicantes (Apple, Google e Amazon), temos um thesaurus reduzido

ao nome das empresas. Porém, problemas com a forma como os nomes das empresas foram escritos ou ainda empresas com mesmo nome mas que não são as empresas de interesse. Essa pesquisa não tratará das técnicas de construção das expressões de busca, apenas sinalizamos que os cuidados para remover a maior quantidade possível de ruído foram tomadas.

Para esta pesquisa, foram filtradas em planilha eletrônica as patentes de interesse e identificadas patentes que não tinham relação com o corpo de inventores esperados ou códigos de classificação internacional de patentes que não estavam relacionadas com o portfólio esperado. Na ocorrência de uma ou mais destas inconsistências, a patente foi examinada individualmente e foi tomada a decisão de excluir ou manter na base de dados que iria mais tarde compor o corpus textual: Busca lens.org (applicant.name:apple).

Plano de Configuração da análise Iramuteq

Etapa 1: Configuração do corpus textual

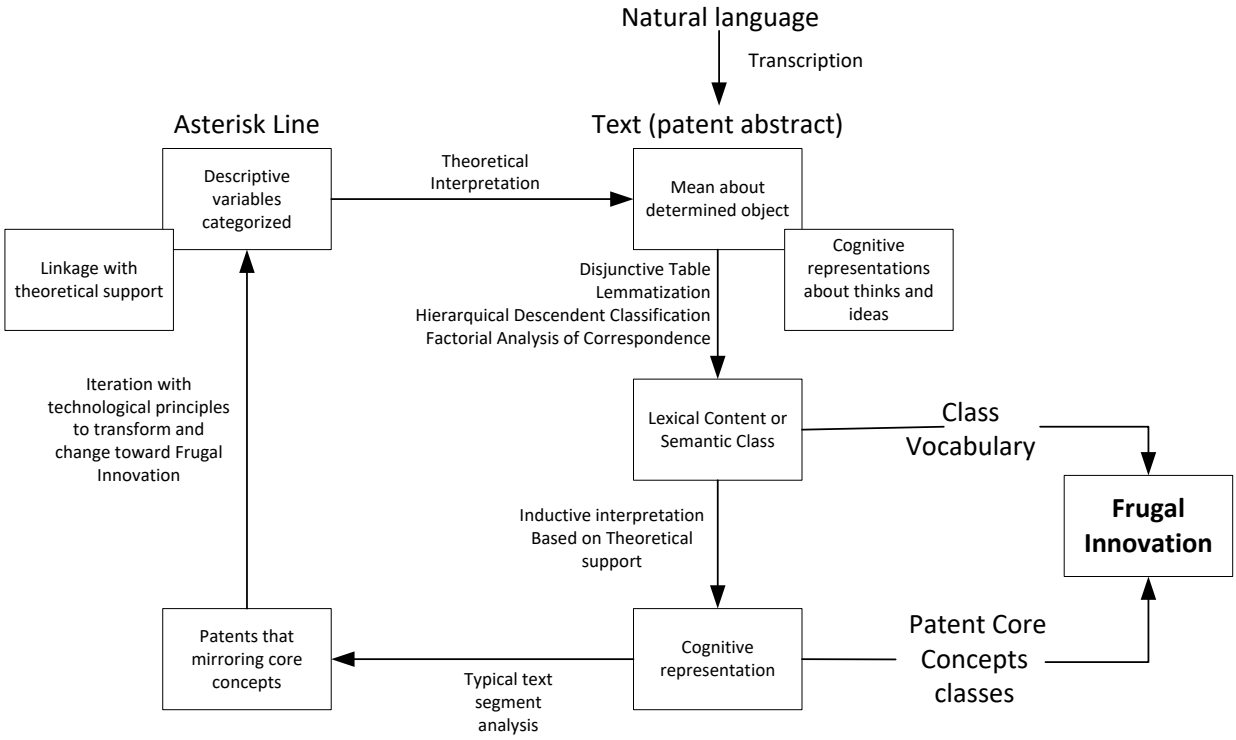
Como procedimentos principais para configuração do corpus textual, destacamos dois principais: primeiro, devem ser colocados todos os textos a serem analisados em um único arquivo eletrônico no formato “txt” padrão “Unicode UTF-8”. Os textos têm origens diversas, como artigos, notícias, patentes, entrevistas, livros, declaração de missão ou visão de empresas, documentos empresariais, como atas de reunião ou políticas de inovação. Manter a homogeneidade do corpus é o ponto mais crítico desta etapa. Temos percebido que pouca atenção tem sido dada para a preparação do corpus; invariavelmente pesquisadores preparam roteiros de entrevistas com perguntas relacionadas a tópicos teóricos que são distintos em termos de construtos ou de ideias principais. Assim, ao preparar o corpus textual desta entrevista, por exemplo, com cinco perguntas, cada resposta deve ser unida a um corpus textual, de forma que todas as respostas à pergunta 1 serão reunidas para formar o corpus textual sobre a pergunta 1 que está interessada em compreender o tópico teórico que a pergunta representa. Obviamente, se a pergunta 1 e 2 são relacionadas ao mesmo construto ou a mesma formação teórica (1 e 2), podem compor o mesmo corpus textual. Outro exemplo clássico é pretender analisar artigos tecnológicos ou científicos completos e para tanto, é criado um corpus textual com todas as seções dos artigos. Artigos, assim como demais documentos que são construídos de acordo com formatos estabelecidos, tem heterogeneidade de conteúdo e estilo de escrita quando comparamos o resumo com a introdução ou com o referencial teórico, por exemplo. Isso significa que, para fins de análise de conteúdo, não parece fazer sentido misturar todas as

seções de um artigo num único corpus textual. A recomendação é escolher uma seção específica do artigo e da mesma forma quando da análise de uma patente. A escolha mais segura e tradicional é a do resumo, mas não há impedimento em escolher outras seções de interesse, desde que o corpus textual formado mantenha a homogeneidade quanto ao conteúdo esperado.

O segundo ponto importante na configuração do corpus textual é a organização de variáveis manifestas, que são as variáveis relacionadas as origens de cada texto que irá compor o corpus textual. É por meio destas variáveis que será possível alcançar interpretações mais aprofundadas quando em paradigma *data-driven* ou ainda é o ponto de ligação entre os dados textuais e a lente teórica em uso na pesquisa. No caso de patentes sobre inovação frugal, o modelo de análise que relaciona as variáveis descritivas ou manifestas com o corpus textual, está apresentado a seguir. O corpus textual representa o significado de determinados objetos emitidos pelo produtor dos textos que o compõe. Portanto, para cada texto que irá compor o corpus textual, há um conjunto de variáveis manifestas que ligam esse texto à teoria, aos meta-dados, a demografia, psicografia, comportamento ou demais atributos que explicam a origem do texto (Mazieri, 2016).

Figura 2

Modelo de Integração Entre as Variáveis Manifestas ou Elementos Teóricos e os Conteúdos do Corpus Textual



Fonte: Mazieri, 2016.

É possível identificar que a ligação entre o suporte teórico e a representação cognitiva que será analisada nos textos é feita no Iramuteq por meio da linha de asteriscos. A seguir, o exemplo da linha de asteriscos da atual pesquisa.

Tabela 2

Modelo de Formato do Texto para Construção do Corpus Textual

****** *Applicants_APPLE INC *Kind_B2 *Publication Date_28/11/2017 *Application Number_US 201514656048 A *IPC_IPCABSENT**
Asset data streams are provided that facilitate the display of large numbers of media assets. Encoded asset data streams provide approximated aspect ratio information for media assets to be used in determining a position for each media asset in a dataset, thus being able to position all of the media assets in a media asset arrangement prior to being scrolled into view by the user. By communicating aspect ratio approximations as part of an encoded asset data stream of data to a web application, a user is able to scroll to any part of a dataset (e.g., a photo library) when presented in the web application without having to wait on the receipt of information for all media assets. Encoded asset data streams may further include asset identification offsets that indicate a sequential ordering of the individual assets in a dataset.

Fonte: Autores, 2022.

Toda a linha de asteriscos começa com 4 asteriscos, significando o começo do texto. Antes de cada variável manifesta usamos um asterisco seguido do nome da varível (*applicant*, por exemplo, é o nome do titular da patente). Após nomear a variável, é colocado o *underscore* (_) e o valor para essa variável (APPLE INC, neste caso). Depois da declaração das variáveis manifestas, segue o texto que será uma das partes do corpus textual. Esse procedimento deve ser feito para todos os textos que serão agrupados para compor o corpus textual.

Orientações gerais sobre o conteúdo do texto

Os softwares de análise de conteúdo baseiam-se em vocabulários, então é necessário verificar se o texto está livre de erros de digitação. No caso de entrevistas, as anotações e intervenções do pesquisador não devem constar do texto transcrito. Não devemos usar negrito, itálico ou outro formatador. Como dica prática, devido a dificuldades no uso de pontuação, sugere-se deixar apenas parágrafos, sem incluir vírgulas ou ponto e vírgula. Palavras hifenizadas serão tratadas como duas palavras dentro do texto, assim como palavras compostas; se deseja tratar santa casa como o local onde são atendidos os pacientes, deve escrever santa_casa, caso contrário, o algoritmo analisará o termo santa separado do termo casa. Verbos com pronomes devem estar na forma de próclise, porque o dicionário do Iramuteq não prevê

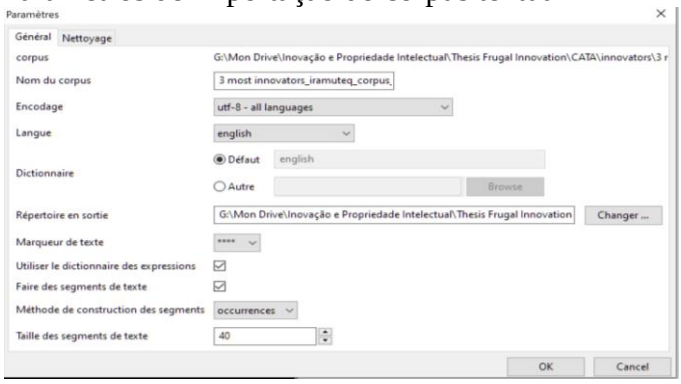
flexões; no lugar de informei-me devemos usar me informei. Em nenhuma parte do texto podemos usar, apóstrofo, aspas, hífen, cifrão, porcentagem ou asterisco.

Etapa 2: Configuração do Iramuteq

Nesta seção serão apresentadas algumas telas com comentários sobre a configuração a ser usada, quando pertinente. As telas oferecidas como guias estão em francês porque o Iramuteq foi desenvolvido em francês. Há opção de uso da interface em inglês e algumas outras línguas, no entanto, os menus de parametrização mais avançados continuam em francês. Por isso, a ideia é apresentar as telas na língua que tem a cobertura completa dos menus e configurações. Para o uso do dia a dia, pode-se escolher a língua de interesse, sabendo que os parâmetros mais avançados não estarão traduzidos.

Figura 3

Parâmetros de importação do corpus textual



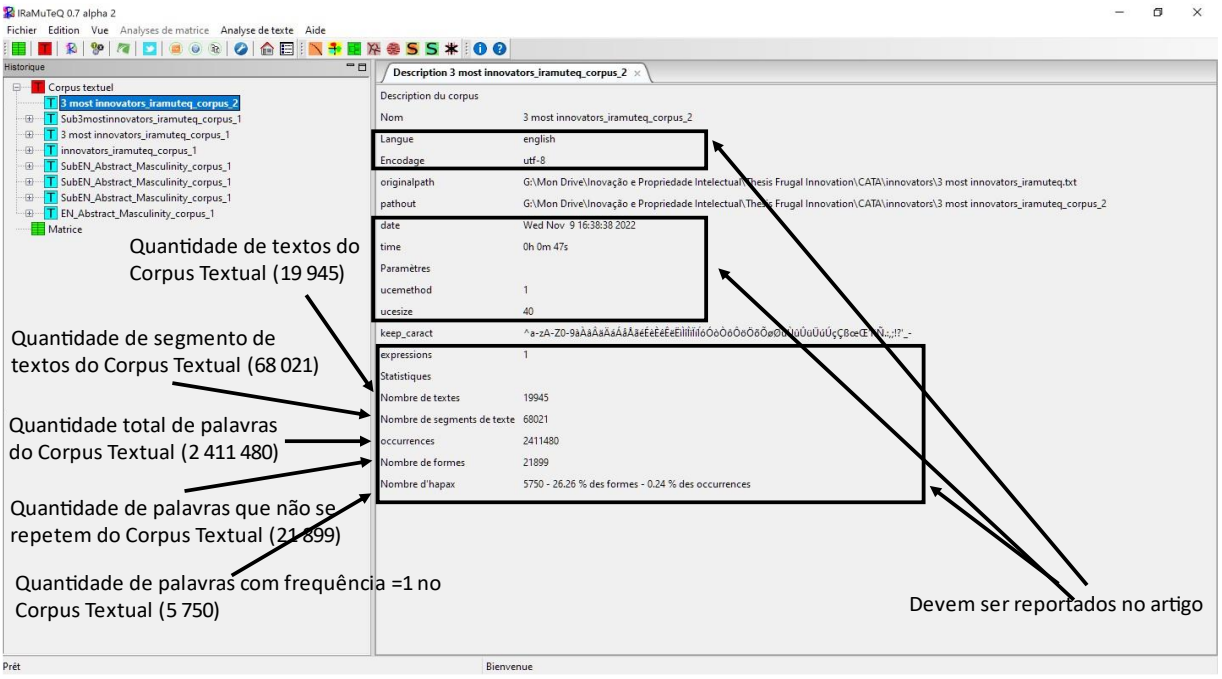
Fonte: Autores, 2022.

Ao solicitar que o Iramuteq carregue o corpus textual que será analisado, a tela acima é apresentada para escolha de parâmetros de interface entre o Iramuteq e o arquivo eletrônico onde está o corpus textual. A codificação a ser usada é o UTF-8 *all languages*, o dicionário é o default, a marcação de texto é dos 4 asteriscos (****) usando o dicionário de expressões, usando segmentos de texto construídos por meio de ocorrências cujo tamanho do segmento de texto será de 40 caracteres. O modo de construção do segmento de texto pode ser alterado de acordo com as necessidades de análise e precisa ser reportado no artigo científico onde a pesquisa está sendo apresentada. As opções de modo de construção do segmento de texto são: (i) quantidade de caracteres, quando os textos que compõe os *corpus* são pequenos, como um título ou uma frase, (ii) quantidade de ocorrências quando os textos são compostos por duas ou mais frases,

como o caso de resumos ou demais partes de um texto, ou (iii) critério de tamanho e pontuação; nesse caso, os critérios de pontuação são calculados pela razão tamanho/pontuação com a prioridade ".", "?", "!" depois ";", ":" e por último "," e espaço. A ideia desse modo de construção do segmento de texto é adequar-se o máximo possível à estrutura da língua. Nas análises de dados textuais envolvendo gestão da inovação, quando são analisados dados secundários que passaram por revisão ortográfica e gramatical, esse modo pode ser útil. No caso de análise de entrevistas pode ser arriscado devido à necessidade de controlar a qualidade gramatical e ortográfica da transcrição, portanto, não é recomendado esse modo. A língua a ser escolhida nesta tela é a língua em que está a transcrição de texto que será analisada. No caso desta pesquisa, os resumos das patentes estão em inglês, por isso, ficou selecionado a língua inglesa.

Figura 4

Parâmetros Iniciais o Plano de Análise de Conteúdo



Fonte: Autores, 2022.

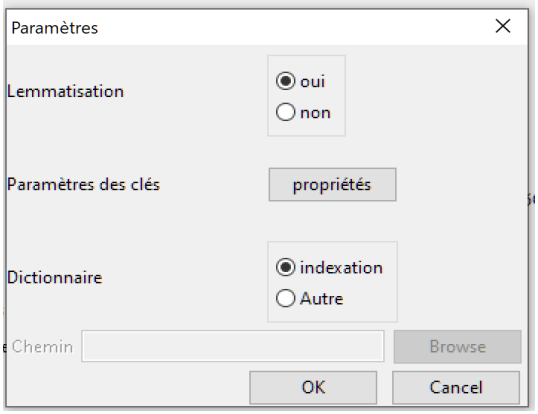
Ao carregar o corpus textual são oferecidos os parâmetros de resumo da carga e essas devem ser reportadas no artigo que está usando análise com Iramuteq, tanto para compreensão do leitor sobre as condições do arquivo eletrônico onde está o corpus textual, seu tamanho, quanto aos parâmetros escolhidos para realizar a carga do corpus. Há também as estatísticas léxicas básicas que mostram quantas ocorrências, s,t,, formas e Hapax foram identificadas no

corpus textual. Ao chegar nessa etapa, significa que não há erros de codificação e que as análises podem ser tentadas.

Qualquer análise que for iniciada passa pelas etapas de parametrização a seguir. Para exemplificar, na última etapa de configuração do Iramuteq solicitamos a análise léxica básica, e a janela de parâmetros aparecerá conforme indicado a seguir. A lematização deve ficar selecionada e o dicionário por indexação também, exceto em casos de uso de dicionários externos.

Figura 5

Parâmetros de Lematização e Dicionário



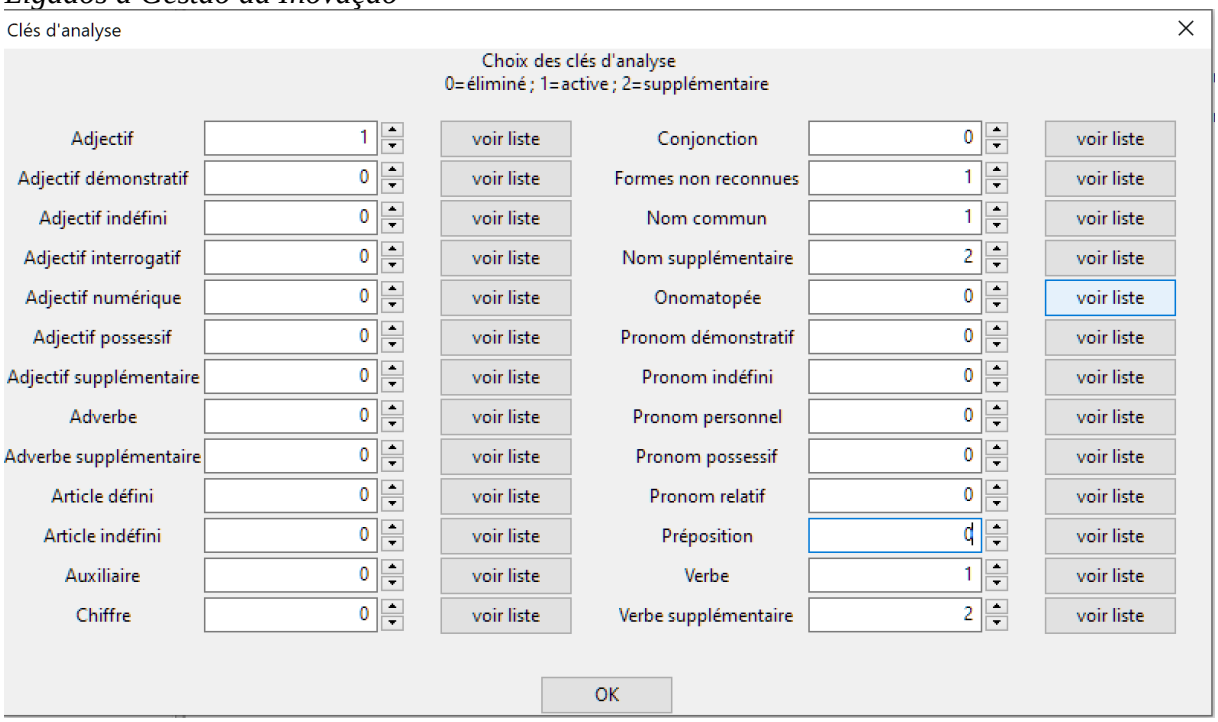
Fonte: Autores, 2022.

Devemos selecionar as propriedades para definir o tratamento que o algoritmo deverá dar aos diversos grupos de palavras, o que chamamos de chave de análise. Essa etapa é bastante negligenciada, o que acarreta dificuldades de visualização e interpretação. O conceito da análise de conteúdo é alcançar a representação cognitiva explícita no texto, portanto, não é um estudo sobre os modos da fala ou da escrita. Desta forma, os parâmetros de interesse são os adjetivos, substantivos e verbos que serão marcados como formas ativas (1). Os suplementos dos verbos e substantivos, marcados como suplementos (2), é recomendado marcar as formas não reconhecidas como ativas (1), pois em inovação nos interessa algo que não sabemos classificar e esteja presente no texto. A seguir, a tela final de como deverá estar parametrizado o Iramuteq para iniciar as análises. Um detalhe importante é que esses parâmetros são mantidos apenas enquanto o Iramuteq for mantido aberto. Caso por qualquer motivo o software seja fechado, ao reabrir, esses parâmetros deverão ser novamente ajustados para realizar qualquer nova análise.

Obviamente as análises já realizadas estão gravadas e seus resultados mantêm os parâmetros de configuração escolhidos antes da análise.

Figura 6

Parâmetros de Apresentação Para Fins de Visualização Melhorada em Análises de Temas Ligados a Gestão da Inovação



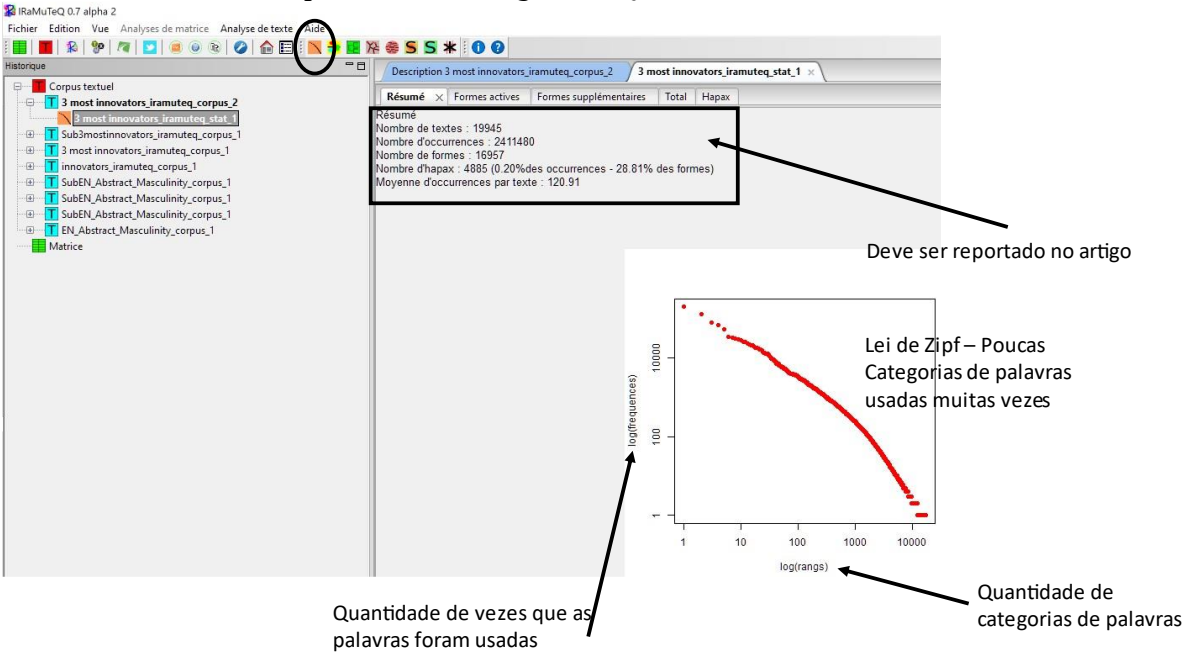
Fonte: Autores, 2022

Etapa 3: Solicitação de Análise

Início ou Primeira análise: análise léxica básica

Figura 7

Parâmetros a Serem Reportados no Artigo na Seção de " Plano de Analise"



Fonte: Autores, 2022.

A diferença entre a quantidade de formas descobertas na tela de estatísticas do corpus textual que foi carregado e a que será efetivamente analisada na lexicografia básica ocorre devido aos parâmetros de exclusão usados para deixar apenas as formas ativas e suplementares de interesse.

Análises disponíveis no Iramuteq e suas aplicações em estudos sobre inovação

Análise Léxica básica

Por meio de uma análise léxica básica podemos saber qual a estrutura do corpus textual em termos de tamanho e de estrutura léxica. Em inovação, é possível identificar os vocabulários mais comuns usados nos textos e saber como determinada tecnologia costuma ser descrita, quais são as palavras mais usadas e quais são as classes gramaticais mais usadas, sendo possível perceber o padrão usado para registar o conhecimento na área pesquisada.

Análise de especificidade e AFC

Análise baseada em matrizes cujas linhas são as palavras do corpus textual e as colunas são compostas pelas variáveis manifestas escolhidas para análise. É possível verificar tanto a frequência da forna na variável (efeito) quanto o qui-quadrado da forma na variável (efeito relativo). Certos substantivos podem ser marcadores importantes para entender uma determinada tecnologia. No nosso caso, a palavra "antena" foi identificada como uma palavra presente em tecnologias da Apple, mas não tão presente nas tecnologias do Google ou da Amazon. Desta forma, a palavra "antena" será classificada em uma das classes semânticas e essa classe terá predominância de tecnologias da Apple. Cabe destacar que não tínhamos essa informação antes de começar a analisar o conteúdo das patentes.

Análise de similitude

Análise baseada na teoria dos grafos que apresenta as palavras em coocorrência, ou seja, aparecem juntas no corpus textual. Essa análise permite conhecer quais palavras estão mais relacionadas. No nosso caso, "antena" está relacionada com substrato e com superfície. Podemos compreender que as patentes da Apple que descrevem antenas são patentes relacionadas à construção eletrônica e não estética, descrevendo a forma como acontece o acoplamento da antena de transmissão de determinado equipamento.

Classificação hierárquica descendente

É a classificação hierárquica descendente clássica, usando o método Renert que, ao invés de palavras, usa os segmentos de textos. São três opções principais: *Double sur srt*, que normalmente não usamos devido a baixa retenção de segmentos de texto; *Simple sur segments de texte* que é a forma padrão de análise e é eficaz em textos mais longos e finalmente *simple sur textes* que é a análise direta sobre o texto, sem considerar os segmentos de texto, uma alternativa para textos curtos como análise de títulos ou de palavras-chave, por exemplo.

Análise fatorial de correspondência

Embora os princípios matemáticos tenham sido explicados na introdução deste artigo, é importante complementar que a análise fatorial de correspondência é formada por uma matriz

em forma de tabela de contingência que cruza as formas ativas com as variáveis (definidas na linha com os asteriscos).

Nuvem de palavras

O Iramuteq oferece ainda a nuvem de palavras. No entanto, não há efetivo aproveitamento para fins de análise técnica ou científica devido a ser apenas visual, sem apresentação de parâmetros. O uso é muito mais ilustrativo do que útil, analiticamente. Consiste em mostrar as palavras presentes no *corpus* em tamanhos de fonte que representam a frequência de presença no texto.

Resultados e interpretação

Retomando, o problema de pesquisa que se pretende enfrentar usando os métodos de análise textual descritos, pretende-se analisar e discriminar as escolhas tecnológicas das empresas mais inovadoras do mundo. A CHD das patentes da Apple, Google e Amazon formaram 4 classes semânticas. A classe 4 é hierarquicamente superior, o que significa que pode carregar o vocabulário mais homogêneo dentro do corpus (para confirmar, é necessário verificar na AFC se a classe 4 posiciona-se em 0,0 desvio padrão do plano fatorial). Em outras palavras, os vocabulários presentes na classe 4 são compartilhados entre Apple, Google e Amazon. A classe 3 é a próxima classe na hierarquia e o fato de estar separada da classe 4 demonstra que há um vocabulário distinto. A mesma interpretação para as classes semânticas 2 e 1.

Figura 8

Dendrograma da Classificação Hierárquica Descendente Com as Classes de Escolhas Tecnológicas da Apple, Google e Amazon: CHD das patentes Apple, Google e Amazon de 2016 a 2021 Com 19Mil Patentes

Fonte: Autores, 2022.

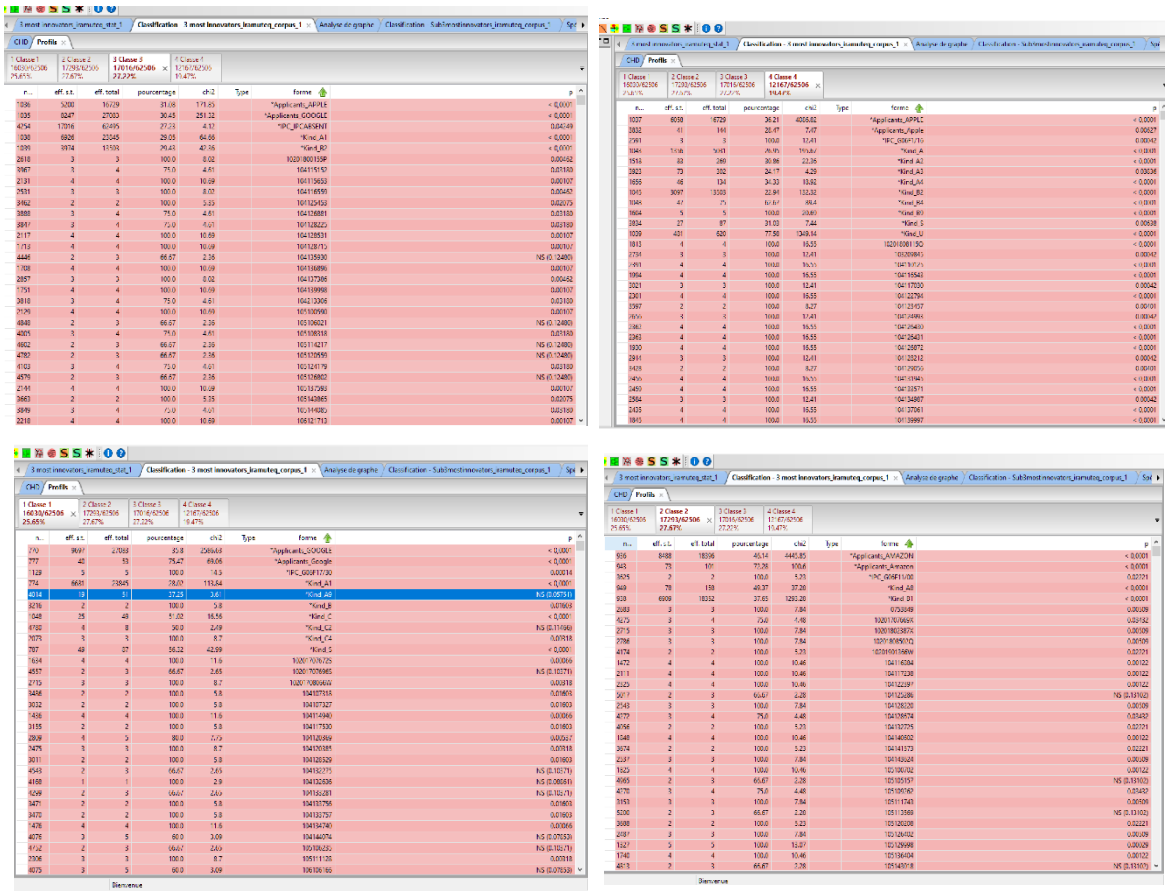
Essa análise inicial nos permite responder nossa primeira pergunta: As empresas mais inovadoras do mundo fazem escolhas idênticas? A resposta é que há pelo menos três classes semânticas: a classe 4, a classe 2 e a classe 1 que mostram que as escolhas tecnológicas não parecem ser idênticas, uma vez que apresentam vocabulários que se discriminam hierarquicamente.

Quanto a segunda pergunta: Se as escolhas forem idênticas ou semelhantes, em que parte elas são parecidas? Podemos responder que as escolhas tecnológicas são parecidas em relação ao conteúdo armazenado na classe semântica 4, onde os vocabulários são compartilhados ou comuns a todo o corpus textual analisado. O fenômeno prático é simples de compreender e está relacionado com os portfólios de produtos em múltiplas verticais. A Apple desenvolve smartphones, mas também serviços e até mesmo carros autônomos. A Google desenvolve softwares, serviços e hardware de computação quântica. Quanto à Amazon, há presença em setores de varejo, como o *e-commerce*, área espacial e demais áreas. Portanto, ainda que haja uma discriminação quanto às classes semânticas, é necessário evoluir na compreensão de cada classe antes de responder essa pergunta.

Para raciocinar a interpretação: Cada classe armazena as formas ativas (palavras que formam os vocabulários da classe, como os verbos e adjetivos), as formas complementares (formas de alta frequência e que, portanto, não expressam conteúdo) e as variáveis manifestas (originadas das linhas de asteriscos). Uma vez estabelecidas as possibilidades de análise de cada classe, a recomendação é analisar as variáveis manifestas que ficam marcadas em vermelho e as formas ativas que ficam marcadas em cinza. As formas complementares, marcadas em verde, normalmente não são analisadas pelo baixo poder de explicação. Na figura 9, pode-se verificar cada classe; como exemplo, ao selecionar a classe 4 e clicar na "forme", a lista se organiza em ordem decrescente, mostrando que a Apple é predominante nessa classe (Significante $p<0,0001$). Isso responde uma parte da questão. A Apple predomina na Classe 4, mas está presente também na classe 3. A classe 3 é compartilhada pela Apple e Google. A Classe 2 tem predominância da Amazon e a classe 1 é predominantemente Google. Interpreta-se que as escolhas tecnológicas da classe 4 são predominantemente Apple, mas também carregam o vocabulário mais comum de todo o corpus textual devido a sua posição na AFC (0,0); as escolhas tecnológicas da classe 2 são predominantemente Amazon; as escolhas tecnológicas da classe 1 são predominantemente Google. Portanto, a classe 3 é diferente da classe 2 que é diferente da classe 1 e mostra a parte das escolhas tecnológicas que são distintas entre a Apple, Amazon e Google. Outra conclusão é que a Amazon não compartilha das escolhas tecnológicas da Apple ou da Google, uma vez que não há ocorrência simultânea da Amazon na mesma classe das demais empresas. Por fim, as escolhas tecnológicas similares ocorrem entre a Apple e a Google e estão localizadas na Classe 4.

Figura 9a, 9b, 9c, 9d

Detalhamento das Classes Semânticas Para Análise de Variáveis Associadas



Fonte: Autores, 2022.

A próxima etapa de análise é acessar o conteúdo das formas ativas e identificar quais são as tecnologias armazenadas em cada classe. Para que fique mais claro, cada classe semântica tem opções tecnológicas tanto da Apple quanto da Google quanto da Amazon, uma vez que a classe é resultado do agrupamento dos segmentos de texto que os são fragmentos da linguagem natural presente nos resumos das patentes, indistintamente da empresa a que pertencem. Portanto, ao analisar as variáveis manifestas predominantes em cada classe, estamos verificando se há ou não a predominância dessa variável na classe. No caso em análise, a variável é o titular da patente. Após verificar a variável predominante na classe, segue-se para a etapa de análise do conteúdo do vocabulário da classe.

Para isso, iniciando pela classe 4, basta clicar com o botão direito do mouse na forma de maior efeito Chi2 e verificar a concordância com os segmentos de texto (*concordancier*). A seguir estão as reproduções da forma “surface” da classe 4. Pode-se interpretar de maneira mais

básica que a parte das escolhas tecnológicas semelhantes entre a Apple, a Google e a Amazon são os projetos de hardware, envolvendo principalmente os módulos onde são fixados os elementos ativos (como circuitos eletrônicos), e os elementos passivos (como antenas, lentes e suportes) são tecnologias que dão suporte aos aspectos estéticos (design) e funcionais (conexões e sistemas de montagem). A interpretação faz-se pela leitura dos segmentos de textos que ficaram estabilizados na classe sob interpretação. O procedimento pode ser repetido com a segunda forma de maior efeito na classe, depois com a terceira e assim por diante até que ocorra determinada saturação, ou seja, repetição de conteúdo. A seguir é apresentado um exemplo da saída "concordancier".

****** *Applicants_APPLE INC *Kind_A1 *Publication Date_28/03/2019 *Application Number_US 201715717821 A *IPC_IPCABSENT**

score : 38678.42

the **slot antenna** may be **fed** via **near field coupling** using a **conductive patch** that is **located** within the **slot** at the **surface** of the **substrate** the **conductive layer rear housing wall** and **vertical portion** may **form** a **cavity** for the **slot antenna**

****** *Applicants_AMAZON TECH INC *Kind_B1 *Publication Date_10/07/2018 *Application Number_US 201514791708 A *IPC_IPCABSENT**

score : 35090.22

a **housing** for an **electronic** device **includes** a single **rear housing assembly coupled** to the **cover glass** of a **display assembly** the **rear housing assembly includes** a **metal rear chassis** with two **layers** of **injection molded material formed** on at least the **chassis side regions**

****** *Applicants_APPLE INC *Kind_B2 *Publication Date_26/04/2016 *Application Number_US 201414195130 A *IPC_IPCABSENT**

score : 34607.80

the **antenna structures** may **include conductive structures** such as **metal traces** on **printed circuits** or other **dielectric substrates internal metal housing structures** or other **conductive electronic** device **housing structures** a **main resonating element arm** may be **separated** from an **antenna ground** by an **opening**

****** *Applicants_APPLE INC *Kind_B2 *Publication Date_23/04/2019 *Application Number_US 201615008139 A *IPC_IPCABSENT**

score : 34044.12

flexible printed circuits with **ground traces** may bisect the **slot shaped opening** to **form** three **electrically isolated slots** each of which is **aligned** with a respective **cavity antenna** the **antennas** may have **antenna grounds formed** from **portions** of the **metal housing** and other **conductive structures**

****** *Applicants_APPLE INC *Kind_A1 *Publication Date_27/07/2017 *Application Number_US 201615008139 A *IPC_IPCABSENT**

score : 34044.12

flexible printed circuits with **ground traces** may bisect the **slot shaped opening** to **form** three **electrically isolated slots** each of which is **aligned** with a respective **cavity antenna** the **antennas** may have **antenna grounds formed** from **portions** of the **metal housing** and other **conductive structures**

****** *Applicants_APPLE INC *Kind_A1 *Publication Date_24/01/2019 *Application Number_US 201715655311 A *IPC_IPCABSENT**

score : 33652.01

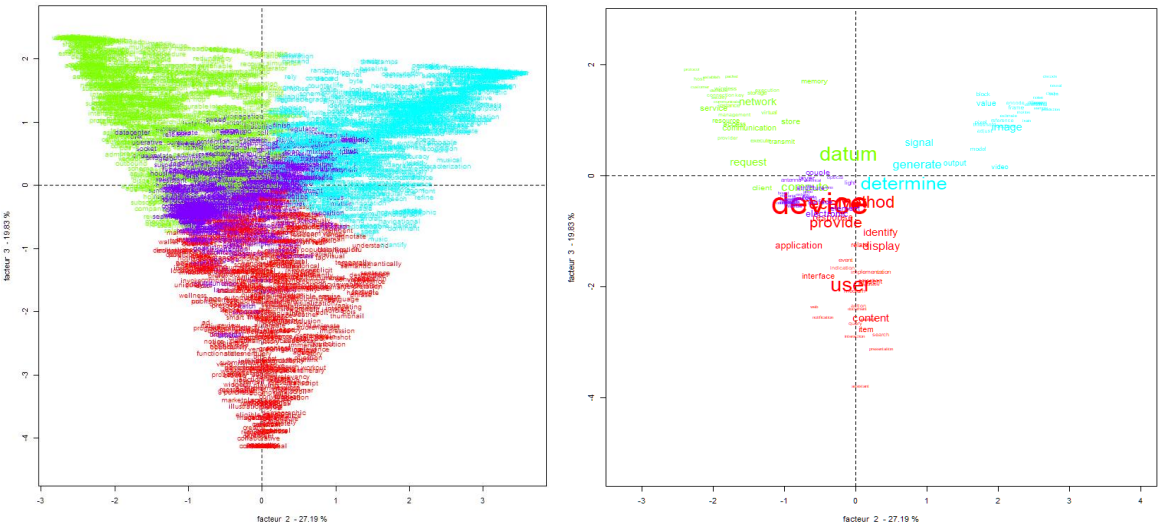
a **housing** made from a **circuit laminate includes** first and **second layers coupled** together each **includes** a **rigid electrically insulating** non **planar structural layer flexible conductive traces disposed** on **surfaces** of the **structural layer** and **flexible connector layers contacting** to the **flexible conductive traces**

Após analisar a classe 4, deve ser feito o mesmo procedimento para as demais classes. Ao realizar o procedimento de análise destacado anteriormente, verifica-se que a classe 3 trata das tecnologias sobre reconhecimento de imagem e de transmissão de imagens, a classe 2 trata das tecnologias das camadas de serviços de conectividade (rede de dados) e serviços, como componentes de aplicações de software e interface com os clientes, como as camadas de retaguarda e a classe 1 trata das tecnologias relacionadas à gestão de conteúdo e experiência do usuário em contexto de rede social.

De fato, ao observar a análise fatorial de correspondência, pode-se verificar que o centro dos eixos fatoriais (0,0) é onde está localizada a classe 4 (parte comum às escolhas tecnológicas da Apple, Google e Amazon). Na AFC, as distâncias são euclidianas, de forma a serem representativas as distâncias métricas entre os formas ou classes. As diversas observações estão nas figuras 10, 11, 12 e 13.

Figura 10 a e 10 b

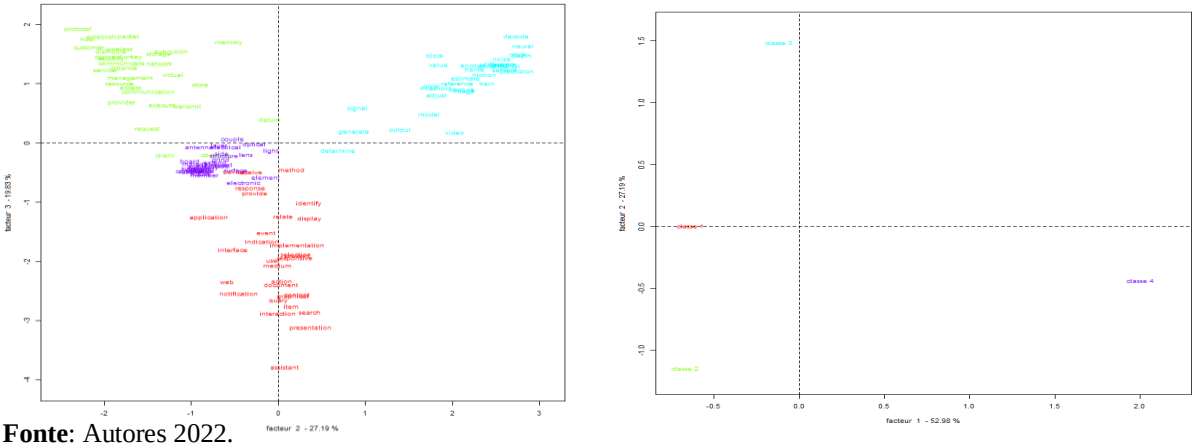
Refinamento da Visualização (Fator 1 e 2, 30 Primeiros Pontos da Classe, Com Ponderação de Tamanho de Apresentação Baseada em chi2): Análise Fatorial de Correspondência Apple, Google e Amazon de 2016 a 2021. 19mil patentes



Fonte: Autores, 2022.

Figura 11 a e 11 b

Refinamento da visualização (Fator 1 e 2, 30 Primeiros Pontos da Classe, Sem Ponderação de Frequência e as Distâncias Euclidianas Entre as Classes): Análise Fatorial de Correspondência Apple, Google e Amazon de 2016 a 2021. 19Mil patentes

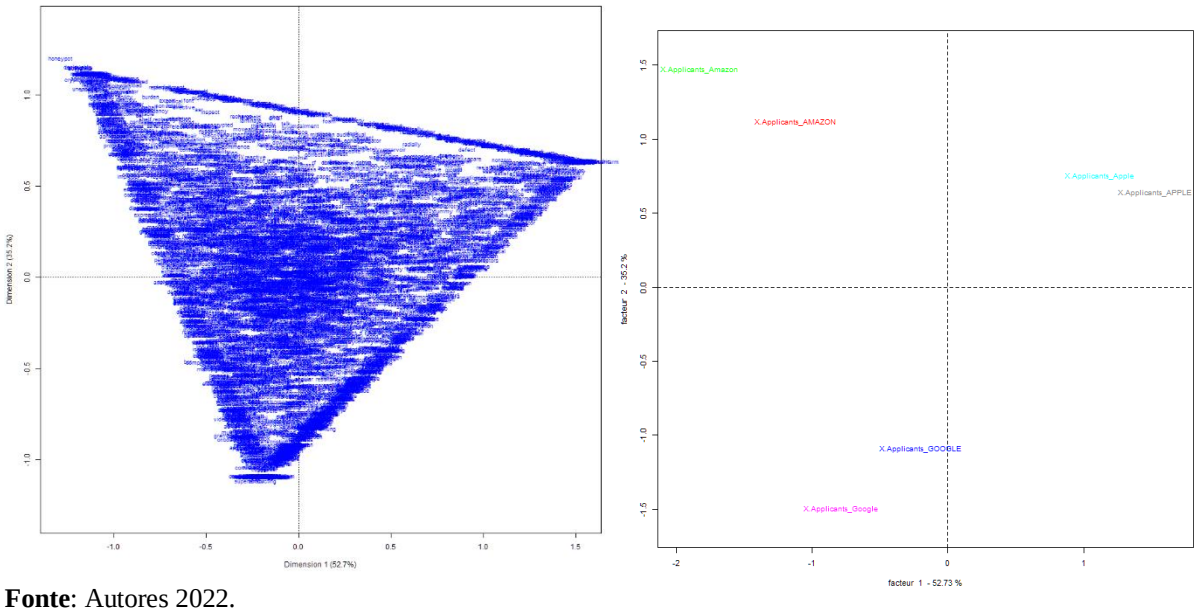


Fonte: Autores 2022.

Análises de especificidade

Figura 12 a e 12 b

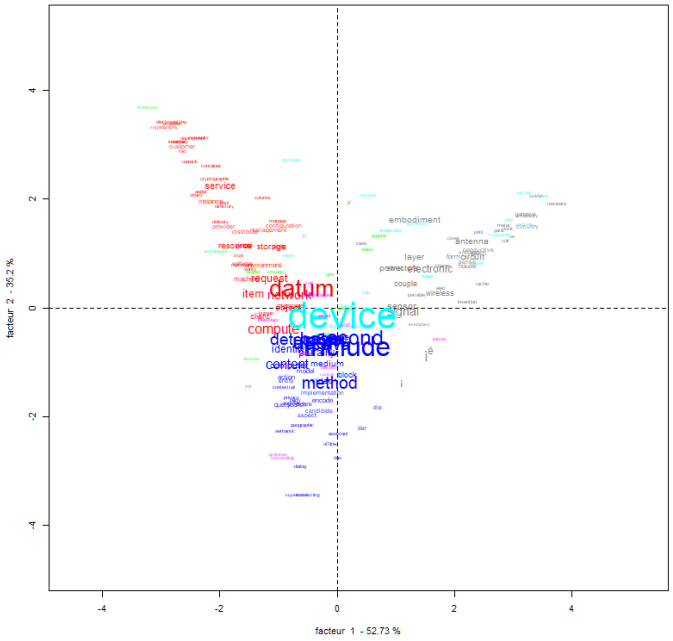
Especificidades: Variável alvo "Titular". A Esquerda a Viasualização dos Dados de Linhas e a Esquerda os Dados de Colunas



Fonte: Autores 2022.

Figura 13

Refinamento da Visualização (Fator 1 e 2, 30 Primeiros Pontos da Classe, Com Ponderação de Tamanho de Apresentação Baseada em chi2)



Fonte: Autores, 2022.

Conclusão

No início deste artigo dissemos que não é possível saber intuitivamente se as empresas mais inovadoras do mundo fazem escolhas idênticas. Tal afirmação nos levou a analisar e discriminar as escolhas tecnológicas das empresas mais inovadoras do mundo. Fizemos algumas perguntas orientativas no início da pesquisa para fins de verificar em que medida podemos responder usando o método da CHD e da AFC. As empresas mais inovadoras do mundo fazem escolhas tecnológicas idênticas? Ficou evidenciado que não. Ainda que existam na classe 4 escolhas comuns entre as três empresas, existem escolhas tecnológicas mais específicas, como aparecem na classe 3, classe 2 e classe 1. Quanto a segunda pergunta: Se as escolhas forem idênticas ou semelhantes, em que parte elas são parecidas? Ficou evidenciado nessa amostragem que as escolhas tecnológicas são parecidas em relação ao conteúdo armazenado na classe semântica 4, pois são tecnologias centrais compartilhadas por todos os titulares da amostra:

> *"projetos de hardware, envolvendo principalmente os módulos onde são fixados os elementos ativos (como circuitos eletrônicos) e os elementos passivos (como antenas, lentes e suportes) são tecnologias que dão suporte aos aspectos estéticos (design) e funcionais (conexões e sistemas de montagem)"*

Como explicado anteriormente, faz sentido que haja uma parte comum, se pensarmos que são empresas que concorrem em várias verticais de negócios, e tal efeito está relacionado com os portfólios de produtos em múltiplas verticais. A Apple desenvolve smartphones, mas também serviços e até mesmo carros autônomos. A Google desenvolve softwares, serviços e hardware de computação quântica. Quanto à Amazon, há presença em setores de varejo, como o *e-commerce*, área espacial e demais áreas. Portanto, ainda que haja uma discriminação quanto às classes semânticas, é necessário evoluir na compreensão de cada classe antes de responder essa pergunta. Quanto a terceira pergunta: Caso façam escolhas diferentes, quais são as tecnologias centrais e como se caracterizam as diferentes escolhas? As tecnologias centrais distintas descobertas pela análise podem ser estruturadas em dois grupos. O grupo das tecnologias centrais não compartilhadas entre os titulares em termos de predominância, que é o caso do Google e Amazon:

> *"Google: Trata das tecnologias relacionadas à gestão de conteúdo e experiência de usuários em contexto de rede social"*
>
> *"Amazon: Trata das tecnologias das camadas de serviços de conectividade (rede de dados) e serviços como componentes de aplicações de software, nesse caso os processos envolvidos nas aplicações dos softwares tanto de interface com os clientes como as camadas de retaguarda"*

O grupo das tecnologias centrais parcialmente compartilhadas entre os titulares em termos de predominância, que é o caso da Apple e Google:

> *"Google e Aple: Trata das tecnologias sobre reconhecimento de imagem e de transmissão de imagens"*

Quanto a última pergunta: Há sobreposição tecnológica? Evidenciou-se que há sobreposição tecnológica, claramente observada na classe 4.

Após análise, pode-se argumentar que as empresas mais inovadoras do mundo têm realizado escolhas tecnológicas distintas (classe1, 2 e 3). As escolhas tecnológicas comuns entre a Apple, Google e Amazon (classe 4) indicam as tecnologias de hardware e dispositivos como o alvo comum com a predominância da Apple. Sabidamente, a Apple foi a mais envolvida no modelo de negócios baseados em venda de hardware e isso fica evidente ao observar sua predominância na classe que armazena as escolhas comuns a todos os titulares analisados.

Como sugestão final, recomenda-se que os artigos que venham a ser escritos considerando as metodologias demonstradas apresentem uma ou mais tabelas que resumam as análises, como o demonstrado na tabela 3.

As tomadas de decisões para a gestão da inovação podem ser revistas de acordo com os achados dessas análises que demonstram as escolhas tecnológicas apresentadas. Caso Apple e Google tenham interesse em determinada vertical de mercado, será importante observar a construção da vantagem competitiva que têm por natureza a distintividade. Não é necessariamente um problema dominar as mesmas tecnologias, ou seja, empresas podem ter domínios tecnológicos similares e ainda assim terem abordagens mercadológicas diferentes. A questão é alcançar consciência sobre onde estão sendo investidos os recursos de Pesquisa e Desenvolvimento, quando analisamos o portfólio de patentes que são a *proxy* das escolhas tecnológicas dessas empresas. Investimento em hardware ainda faz sentido, ou seja, alocar recursos para construir tecnologias proprietárias de hardware terão efeito sobre a vantagem competitiva? As escolhas analisadas são do período de 2016 a 2019, portanto, recentes. Por outro lado, alcançar tecnologias proprietárias em hardware pode ser uma estratégia para alterar o modelo de negócios de venda de dispositivos para aluguel ou ainda para dispositivos em forma de serviço? Como podemos verificar, ao conhecer as escolhas tecnológicas, trouxemos à discussão apenas uma pequena parte das contribuições possíveis para as discussões da gestão da inovação. Certamente este é só o começo.

A seguir, duas tabelas resumo; a tabela 3 resume o passo a passo para análise e interpretação da classificação hierárquica descendente e da análise fatorial de correspondência e a tabela 4 apresenta o modelo de tabela resumo da análise dos dados textuais usados nesse artigo.

Tabela 3

Resumo do Plano de Análise

Preparação do corpus textual	**(1)** verificar erros de digitação; **(2)** excluir negrito, itálico, apóstrofo, aspas, cifrão, porcentagem, asterisco, vírgula e ponto e vírgula; **(3)** verbos com pronome devem estar na forma de próclise; **(4)** substituir hífen por underscore
Organização das variáveis manifestas a serem associadas ao corpus textual	linha começa com quatro asteriscos (****) seguido das variáveis manifestas introduzidas por um asterisco e o nome da variável, incluído underscore (_) e o valor da variável. Após a declaração das variáveis, incluir o texto quer irá compor o corpus textual. Ex: **** *nome da variável 1_ valor da variável * nome da variável 2 _ valor da variável Texto
salvar o arquivo no formato txt padrão unicode UTF-8	
Definição dos parâmetros de interface do Iramuteq e o arquivo eletrônico do corpus textual a ser importado	**(1)** selecionar "codificação UTF-8 all languages"; **(2)** selecionar o idioma em que o corpus textual é apresentado; **(3)** selecionar "dicionário default"; **(4)** selecionar marcação do texto "****"; **(5)** selecionar "usar o dicionário de expressões"; **(6)** selecionar "usar segmentos de texto construído por meio de ocorrências"; **(7)** manter tamanho do segmento de texto de 40 caracteres; **(8)** selecionar modo de construção do segmento (quantidade de caracteres ou quantidade de ocorrências ou critério de tamanho e pontuação)
importar o arquivo para o Iramuteq	
Parametrização das análises	**(1)** selecionar lematização; **(2)** selecionar "dicionário por indexação", exceto em caso de uso de dicionários externos
Seleção das propriedades para definir o tratamento aos diversos grupos de palavras (chave de análise)	**(1)** marcar formas ativas como "1"; **(2)** marcar formas suplementares como "2"; **(3)** marcar formas excluídas da análise como "zero"
Análises do Iramuteq	**(1)** Análise léxica básica - verifica a distribuição de frequência das palavras no corpus; **(2)** Análise de similitude - permite inferir a estrutura de construção do texto a partir da coocorrência entre as palavras; **(3)** Classificação Hierárquica Descendente (CHD) - selecionar "double sur srt", "simples sur segments de texte" ou "simple sur textes"; **(4)** Análise Fatorial de Correspondência (AFC) **(5)** Análise de Especificidade e Análise Fatorial de Correspondência (AFC) **(6)** Nuvem de Palavras

Fonte: Autores, 2022.

Tabela 4

Resumo de Análise Textual

Nomenclatura Metodológica Aplicada	Classe	Variável Manifesta	Segmentos de textos	Posição central da AFC (0,0)	Resumo de Análise de Conteúdo		
Nomenclatura de análise textual	Classe Semantica	Titular	Conteúdo Predominante	Vocabulário Comum do Corpus	Recombinação		
Objetivos de Pesquisa	Agrupamento das escolhas Tecnológicas	Titular Predominante	Escolhas Tecnológicas		**Apple**	**Google**	**Amazon**
Análise das escolhas tecnológicas das empresas mais inovadoras	1	Google	Trata das tecnologias relacionadas à gestão de conteúdo e experiência de usuários em contexto de rede social	Não		Trata das tecnologias relacionadas à gestão de conteúdo e experiência de usuários em contexto de rede social	
	2	Amazon	Trata das tecnologias das camadas de serviços de conectividade (rede de dados) e serviços como componentes de aplicações de software, nesse caso os processos envolvidos nas aplicações dos softwares tanto de interface com os clientes como as camadas de retaguarda	Não			Trata das tecnologias das camadas de serviços de conectividade (rede de dados) e serviços como componentes de aplicações de software, nesse caso os processos envolvidos nas aplicações dos softwares tanto de interface com os clientes como as camadas de retaguarda
	3	Apple e Google	Trata das tecnologias sobre reconhecimento de imagem e de transmissão de imagens	Não	Trata das tecnologias sobre reconhecimento de imagem e de transmissão de imagens	Trata das tecnologias sobre reconhecimento de imagem e de transmissão de imagens	
	4	Apple	Projetos de hardware, envolvendo principalmente os módulos onde são fixados os elementos ativos (como circuitos eletrônicos) e os elementos passivos (como antenas, lentes e suportes) são tecnologias que dão suporte aos aspectos estéticos (design) e funcionais (conexões e sistemas de montagem)	Sim	Projetos de hardware, envolvendo principalmente os módulos onde são fixados os elementos ativos (como circuitos eletrônicos) e os elementos passivos (como antenas, lentes e suportes) são tecnologias que dão suporte aos aspectos estéticos (design) e funcionais (conexões e sistemas de montagem)	Projetos de hardware, envolvendo principalmente os módulos onde são fixados os elementos ativos (como circuitos eletrônicos) e os elementos passivos (como antenas, lentes e suportes) são tecnologias que dão suporte aos aspectos estéticos (design) e funcionais (conexões e sistemas de montagem)	Projetos de hardware, envolvendo principalmente os módulos onde são fixados os elementos ativos (como circuitos eletrônicos) e os elementos passivos (como antenas, lentes e suportes) são tecnologias que dão suporte aos aspectos estéticos (design) e funcionais (conexões e sistemas de montagem)

Fonte: Autores, 2022.

Referências

Ang, C. (2021, julho 19). *Ranked: The Most Innovative Companies in 2021*. Visual Capitalist. https://www.visualcapitalist.com/ranked-the-most-innovative-companies-in-2021/

Bardin, L. (1977). Content analysis. *São Paulo: Livraria Martins Fontes*.

Benzécri, J.-P. (1973). L'analyse des données, vol. 2. *Paris: Dunod*.

Camargo, B. V., & Justo, A. M. (2013). IRAMUTEQ: Um software gratuito para análisede dados textuais. *Temas em Psicologia*, *21*(2), 513–518. https://doi.org/10.9788/TP2013.2-16

Campion, E. D., & Campion, M. A. (2020). Using Computer-assisted Text Analysis (CATA) to Inform Employment Decisions: Approaches, Software, and Findings. *Research in Personnel and Human Resources Management*.

Cibois, P., & Jambu, M. (1981). Analyse des données et sociologie. *L'Année sociologique (1940/1948-)*, *31*, 333–348.

Hair, J. F. (2009). *Multivariate data analysis*.

Hirschfeld, H. O. (1935). A Connection between Correlation and Contingency. *Mathematical Proceedings of the Cambridge Philosophical Society*, *31*(4), 520–524. https://doi.org/10.1017/S0305004100013517

Mazieri, M. R. (2016). *Patentes e inovação frugal em uma perspectiva contributiva*. http://bibliotecatede.uninove.br/handle/tede/1600

Miraballes, M., & Gámbaro, A. (2018). Influence of Images on the Evaluation of Jams Using Conjoint Analysis Combined with Check-All-That-Apply (CATA) Questions. *Journal of food science*, *83*(1), 167–174.

Miraballes, M., Hodos, N., & Gámbaro, A. (2018). Application of a pivot profile variant using CATA questions in the development of a whey-based fermented beverage. *Beverages*, *4*(1), 11.

Ratinaud, P., & Marchand, P. (2012). Application de la méthode ALCESTE à de" gros" corpus et stabilité des" mondes lexicaux": Analyse du" Cable-Gate" avec IraMuTeQ. *Actes des 11e Journées internationales d'Analyse statistique des Données Textuelles. JADT 2012*.

Reinert, M. (1990a). Une méthode de classification des énoncés d'un corpus présentée à l'aide d'une application. *Les cahiers de l'analyse des données*, *15*(1), 21–36.

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

Reinert, M. (1995). Quelques aspects de choix des unités d'analyse et de leur contrôle dans la méthode Alceste. *JADT1995*, *1*, 27–34.

Reinert, M. (2007). Postures énonciatives et mondes lexicaux stabilisés en analyse statistique de discours. *Langage et société*, *3*, 189–202.

Short, J. C., Broberg, J. C., Cogliser, C. C., & Brigham, K. H. (2010). Construct validation using computer-aided text analysis (CATA) an illustration using entrepreneurial orientation. *Organizational Research Methods*, *13*(2), 320–347.

Tidd, J., & Bessant, J. (2015). *Gestão da inovação-5*. Bookman Editora.

van Meter, K. M., Mounier, L., Chartron, G., & Reinert, M. (1991). Multimethod Analysis: Official Biographies of Members of the Central Committee of the Soviet Union Communist Party. *Bulletin of Sociological Methodology/Bulletin de Méthodologie Sociologique*, *33*(1), 20–37. https://doi.org/10.1177/075910639103300102