ISSN 1413-389X Temas em Psicologia – 2013, Vol. 21, nº 2, 513-518
DOI: 10.9788/TP2013.2-16

IRAMUTEQ: Um Software Gratuito para Análise de Dados Textuais

IRAMUTEQ: Interface de R pour les Analyses Multidimensionnelles de Textes et de Questionnaires

Resenha do *software*: Ratinaud, P. (2009). IRAMUTEQ: Interface de R pour les Analyses Multidimensionnelles de Textes et de Questionnaires [Computer software]. Retrieved from http://www.iramuteq.org

Brigido Vizeu Camargo[1]
Programa de Pós-Graduação em Psicologia da Universidade Federal de Santa Catarina, Florianópolis, Brasil
Coordenação do Laboratório de Psicologia Social da Comunicação e Cognição, Florianópolis, Brasil
Ana Maria Justo
Programa de Pós-Graduação em Psicologia da Universidade Federal de Santa Catarina, Florianópolis, Brasil

Resumo

Esta nota visa apresentar o *software* IRAMUTEQ (*Interface de R pour les Analyses Multidimensionnelles de Textes et de Questionnaires*), desenvolvido por Pierre Ratinaud (2009). Trata-se de um programa informático gratuito, que se ancora no *software* R e permite diferentes formas de análises estatísticas sobre corpus textuais e sobre tabelas de indivíduos por palavras. Desenvolvido inicialmente em língua francesa, este programa começou a ser utilizado no Brasil em 2013. O dicionário experimental em língua portuguesa encontra-se em fase de aprimoramento, embora já seja bastante adequado. O IRAMUTEQ possibilita os seguintes tipos de análises: estatísticas textuais clássicas; pesquisa de especificidades de grupos; classificação hierárquica descendente; análises de similitude e nuvem de palavras. Pelo seu rigor estatístico, pelas diferentes possibilidades de análise, interface simples e compreensível, e, sobretudo por seu acesso gratuito, o IRAMUTEQ pode trazer muitas contribuições aos estudos em ciências humanas e sociais, que têm o conteúdo simbólico proveniente dos materiais textuais como uma fonte importante de dados de pesquisa.

Palavras-chave*:* Análise textual, classificação hierárquica descendente, IRAMUTEQ.

IRAMUTEQ: A Free Software for Analysis of Textual Data

Abstract

This note aims to present the software IRAMUTEQ (*Interface de R pour les Analyses Multidimensionnelles de Textes et de Questionnaires*), developed by Pierre Ratinaud (2009). This is a free program,

[1] Endereço para correspondência: Centro de Filosofia e Ciências Humanas, Programa de Pós-Graduação em Psicologia, Universidade Federal de Santa Catarina, Campus Universitário Reitor João David Ferreira Lima, Trindade, Florianópolis, SC, Brasil 88040-900. E-mail: brigido.camargo@yahoo.com.br

anchored in R software; and it allows different means of textual statistics analysis in both textual material and tables (individuals by words). Developed originally in French, this software started to be used in Brazil in 2013. An experimental Portuguese dictionary is being improved, nevertheless it allows sufficiently accurate analyzes. The IRAMUTEQ enables different types of analysis: classical textual statistics; specificities of groups; descending hierarchical classification; analyzes of similarity and word cloud. Because of its statistical accuracy, of the distinct possibilities of analysis it allows us to carry out, of its simple and understandable interface, and especially because it is free; IRAMUTEQ can bring many contributions to humanities and social sciences, which are subject areas accustomed with working with symbolic content derived from textual materials as an important kind of research data.

Keywords: Textual analysis, descendant hierarchical classification, IRAMUTEQ.

IRAMUTEQ: Un *Software* Libre para el Análisis de Datos Textuales

Resumen

Esta nota presenta el *software* IRAMUTEQ (*Interface de R pour les Analyses Multidimensionnelles de Textes et de Questionnaires*), desarrollado por Pierre Ratinaud (2009). Este es un *software* gratuito que se basa en el *software* R y permite diferentes formas de análisis estadísticas de corpus textual y de tablas: individuos x palabras. Desarrollado originalmente en francés, este programa comenzó a ser utilizado en Brasil en 2013. El diccionario experimental de la lengua portuguesa se encuentra actualmente en la mejora, aunque es bastante adecuado. El IRAMUTEQ permite los siguientes tipos de análisis: estadísticas textuales clásicas; la investigación grupos específicos; clasificación jerárquica descendiente; análisis de similitud y la nube de palabras. Por su rigor estadístico, las diferentes posibilidades de análisis, su presentación simple y comprensible, y sobre todo por su acceso libre, el IRAMUTEQ puede traer muchas contribuciones a los estudios de humanidades y ciencias sociales, que tienen el contenido simbólico de los materiales textuales de una fuente de datos importantes de investigación.

Palabras clave: Análisis textual, clasificación jerárquica descendiente, IRAMUTEQ.

A Análise Textual com Auxílio de Programas Informáticos

A análise textual consiste num tipo específico de análise de dados, que se trata especificamente da análise de material verbal transcrito, ou seja, de textos produzidos em diferentes condições tais como: textos originalmente escritos, entrevistas, documentos, redações etc., fontes usadas tradicionalmente em Ciências Humanas e Sociais (Nascimento & Menandro, 2006). Por tratar-se de dados que são compostos essencialmente pela linguagem, os mesmos mostram-se relevantes aos estudos sobre pensamentos, crenças, opiniões – conteúdo simbólico produzido em relação a determinado fenômeno.

A análise de dados textuais, ou análise lexical, conforme *Lahlou* (1994) propõe que se supere a dicotomia clássica entre quantitativo e qualitativo na análise de dados, na medida em que possibilita que se quantifique e empregue cálculos estatísticos sobre variáveis essencialmente qualitativas – os textos. Torna-se possível, a partir da análise textual, descrever um material produzido por determinado produtor, seja individual ou coletivamente (um indivíduo ou um grupo), como também pode ser utilizada a análise textual com a finalidade comparativa, relacional, comparando produções diferentes em função de variáveis específicas que descrevem quem produziu o texto.

O uso *softwares* específicos para análise de dados textuais tem sido cada vez mais presente em estudos na área de Ciências Humanas e Sociais, especialmente naqueles estudos em que o corpus a ser analisado é bastante volumoso (Chartier & Meunier, 2011; Lahlou, 2012; Nascimento & Meandro, 2006). No Brasil, já desde a década de 1990 são utilizados alguns *softwares* para análises de textos, tais como o *Ethnogra-*

ph, o *Nudist* e o *Atlas TI*, os quais, ao organizarem os dados, facilitam a realização de análises de conteúdo. Nesta época, na França, o uso de programas informáticos para análises de dados textuais já era mais voltado para cálculos estatísticos (análise quantitativa de dados textuais). Alguns dos *softwares* pioneiros foram o *Tri Deux Mots*, desenvolvido por P. Cibois (1990); o SPAD (*Système Portable pour l'Analyse des Données*), desenvolvido por L. Lebart (Lebart & Salem, 1994; SPAD, 2008); *Evocation* e *Similitude*, desenvolvidos por P. Vergès (Vergès, Junique, Barbry, Scano, & Zeliger, 2002; Vergès, Scano, & Junique, 2002), os quais realizam tanto análises estatísticas clássicas, quanto multivariadas, sobre dados textuais, e possibilitam que se relacione as palavras encontradas na produção textual, com variáveis categoriais caracterizadoras dos produtores do texto.

Enquanto nos *softwares* até então mencionados a unidade de análise era obrigatoriamente a palavra, um programa informático inovador desenvolvido por M. Reinert (*Analyse Lexicale par Context d'un Ensemble de Segments de Texte* [ALCESTE], 2009; Reinert, 1990) se diferenciou dos demais, pois possibilitou que se recuperasse o contexto em que as palavras ocorriam. O ALCESTE apresenta um interesse particular, pois possibilita a execução de uma análise do tipo Classificação Hierárquica Descendente (CHD), que, além de permitir uma análise lexical do material textual, oferece contextos (classes lexicais), caracterizados por um vocabulário específico e pelos segmentos de textos que compartilham este vocabulário (Camargo, 2005). Ele foi introduzido no Brasil em 1998 (Veloz, Nascimento-Schulze, & Camargo, 1999), e passou a ser utilizado, sobretudo entre os pesquisadores da área de Representações Sociais.

Recentemente surgiu uma alternativa para realização de análises textuais tão ou mais sofisticadas que o *software* ALCESTE. Em 2011, o Laboratório de Psicologia Social da Comunicação e Cognição da Universidade Federal de Santa Catarina (LACCOS/UFSC) obteve informação de um *software* gratuito e com fonte aberta, desenvolvido pelo pesquisador francês Pierre Ratinaud (2009), que utiliza-se do mesmo algorítmico do ALCESTE (Reinert, 1990) para realizar análises estatísticas de textos. Tal informação já foi publicada por Lahlou (2012), o qual salienta o profundo conhecimento de Ratinaud na área e seu brilhante trabalho no desenvolvimento do *software* IRAMUTEQ (*Interface de R pour les Analyses Multidimensionnelles de Textes et de Questionnaires*), que incorpora, além da CHD proposta por Reinert (1990), outras análises lexicais que não são realizadas pelo *software* ALCESTE.

O *Software* IRAMUTEQ

O IRAMUTEQ é um *software* gratuito e desenvolvido sob a lógica da *open source,* licenciado por GNU GPL (v2). Ele ancora-se no ambiente estatístico do *software* R e na linguagem *python* (www.python.org).

Este programa informático viabiliza diferentes tipos de análise de dados textuais, desde aquelas bem simples, como a lexicografia básica (cálculo de frequência de palavras), até análises multivariadas (classificação hierárquica descendente, análises de similitude). Ele organiza a distribuição do vocabulário de forma facilmente compreensível e visualmente clara (análise de similitude e nuvem de palavras).

Nas *análises lexicais clássicas,* o programa identifica e reformata as unidades de texto, transformando *Unidades de Contexto Iniciais* (UCI) em *Unidades de Contexto Elementares* (UCE); identifica a quantidade de palavras, frequência média e número de *hapax* (palavras com frequência um); pesquisa o vocabulário e reduz das palavras com base em suas raízes (lematização); cria dicionário de formas reduzidas, identifica formas ativas e suplementares.

Na *análise de especificidades*, é possível associar diretamente os textos do banco de dados com variáveis descritoras dos seus produtores; é possível analisar a produção textual em função das variáveis de caracterização. Trata-se de uma análise de contrastes, na qual o *corpus* é dividido em função de uma variável escolhida pelo pesquisador. Por exemplo, é possível comparar a produção textual de homens e mulheres em relação a determinado tema.

O *método da Classificação Hierárquica Descendente* (CHD) proposto por Reinert (1990) e utilizado pelo *software* ALCESTE classifica os segmentos de texto em função dos seus respectivos vocabulários, e o conjunto deles é repartido com base na frequência das formas reduzidas (palavras já lematizadas). Esta análise visa obter classes de UCE que, ao mesmo tempo, apresentam vocabulário semelhante entre si, e vocabulário diferente das UCE das outras classes. O IRAMUTEQ também fornece outra forma de apresentação dos resultados, por meio de uma análise fatorial de correspondência feita a partir da CHD (Análise Pós-Fatorial) que representa num plano cartesiano as diferentes palavras e variáveis associadas a cada uma das classes da CHD. A interface possibilita que se recuperem, no *corpus* original, os segmentos de texto associados a cada classe, momento em que se obtém o contexto das palavras estatisticamente significativas, possibilitando uma análise mais qualitativa dos dados.

A *análise de similitude* se baseia na teoria dos grafos, possibilita identificar as coocorrências entre as palavras e seu resultado traz indicações da conexidade entre as palavras, auxiliando na identificação da estrutura de um *corpus* textual, distinguindo também as partes comuns e as especificidades em função das variáveis ilustrativas (descritivas) identificadas na análise (Marchand & Ratinaud, 2012).

A *nuvem de palavras* as agrupa e as organiza graficamente em função da sua frequência. É uma análise lexical mais simples, porém graficamente bastante interessante, na medida em que possibilita rápida identificação das palavras-chave de um *corpus*.

Estas análises podem ser realizadas tanto a partir de um grupo de textos a respeito de uma determinada temática (*corpus*) reunidos em um único arquivo de texto; como a partir de tabelas com indivíduos em linha e palavras em coluna, organizadas em planilhas, como é o caso dos bancos de dados construídos a partir de testes de evocações livres. Os textos ou tabelas devem ser preferencialmente gerados pelos softwares OpenOffice.org ou LibreOffice, para evitar bugs relativos a codificação.

Para instalar o *software* gratuitamente basta fazer o *download* do *software* R em www.r-project.org e instalá-lo; e em seguida fazer o *download* do *software* IRAMUTEQ em www.iramuteq.org, e também instalá-lo. É necessário que antes de instalar o IRAMUTEQ se instale o R, pois o IRAMUTEQ se utilizará do *software* R para processar suas análises.

IRAMUTEQ e Análise de Dados em Língua Portuguesa

O *software* IRAMUTEQ foi desenvolvido inicialmente em língua francesa, onde estudos já o empregam como ferramenta de análise de dados (Marchand & Ratinaud, 2012; Ratinaud & Marchand, 2012) e também já possui os dicionários completos nas línguas inglesa e italiana. Ele começou a ser utilizado no Brasil em 2013. Neste momento a equipe do LACCOS (UFSC) em parceria com o Centro Internacional de Estudos em Representações Sociais e Subjetividade – Educação, da Fundação Carlos Chagas (CIERS-ed/FCC); e com o grupo de pesquisa Valores, Educação e Formação de Professores da Universidade Estadual Paulista Júlio de Mesquita Filho (UNESP); estão aprimorando o dicionário experimental em língua portuguesa, o qual deverá ser concluído até o final deste ano, garantindo análises mais estáveis. Mesmo assim, nos processamentos de dados já realizados nessa fase experimental, observou-se que o atual dicionário já está bem aprimorado, permitindo realização de análises suficientemente precisas, o que torna o *software* IRAMUTEQ útil para análises de dados em língua portuguesa. Encontram-se também em fase experimental os dicionários nas línguas alemã, sueca, espanhola e grega.

Conclusões

O *software* IRAMUTEQ apresenta rigor estatístico e permite aos pesquisadores utilizarem diferentes recursos técnicos de análise lexical. Além disso, sua interface é simples e facilmente compreensível, e, sobretudo seu acesso é gratuito e é do tipo *open source.* Por estas características acredita-se que o mesmo possa trazer muitas

contribuições ao campo de estudo das ciências humanas e sociais, em diversos países do mundo, e em especial nos de língua portuguesa.

O uso de *softwares* para análise de textos tem recebido algumas críticas, como mencionam Chartier e Meunier (2011) ao salientarem que o uso de programas informáticos, por facilitar o processamento de grandes volumes ou número de textos, abre a possibilidade do pesquisador negligenciar seu papel na análise dos dados textuais. Nestes casos ocorre certo esvaziamento das relações do material textual com o contexto, além de descrições mecânicas do conteúdo estudado. Conforme Lahlou (2012), em muitos casos, confunde-se o *software* utilizado com um método, o que se deu especialmente nas publicações que envolviam o uso do ALCESTE. Concorda-se com os autores citados, que além do manejo do *software* é importante que o pesquisador conheça as técnicas de processamento dos dados empregadas, a forma de recuperação deste material analisado e o método de pesquisa usado no estudo que utiliza este recurso.

Lahlou (2012) aponta ainda que por alguns anos houve certa carência de publicações em língua inglesa sobre as análises envolvendo estatísticas textuais; sendo a maior parte delas publicadas exclusivamente em língua francesa, portanto de difícil acesso. No Brasil podemos observar fenômeno semelhante, onde a carência de referências em língua portuguesa, somada à agilidade no tratamento dos textos, e a certo "fascínio" que os *softwares* de análise textual exercem nos pesquisadores, resultam em inúmeras publicações que citam o próprio software como se fosse a técnica de análise dos dados, e ainda, como se fosse o método da pesquisa. Observa-se também que há trabalhos que restringem a análise dos dados às informações presentes nos outputs dos *softwares*, o que resulta muito aquém do exercício necessário ao pesquisador, que consiste em explorar o material de texto, interpretar os resultados apresentados pelo *software*, considerando inclusive aqueles dados que não foram diretamente expressos pelo processamento informático.

Considera-se que IRAMUTEQ pode trazer importantes contribuições aos estudos que envolvam dados textuais. O processamento de dados permitido pelo *software* viabiliza o aprimoramento das análises, inclusive em grandes volumes de texto. Pode-se utilizar das análises lexicais, sem que se perca o contexto em que a palavra aparece, tornando possível integrar níveis quantitativos e qualitativos na análise, trazendo maior objetividade e avanços às interpretações dos dados de texto. Entretanto, lembramos a ressalva apontada por Chartier e Meunier (2011) e reiterada por Lahlou (2012) de que um *software* não é um método, e os relatórios gerados pelo *software* não são, em si, a análise dos dados. O IRAMUTEQ pode ser muito útil se acompanhado de um estudo sobre o significado das análises lexicais e do emprego de análises multivariadas, além de um bom domínio do estado da arte que envolve o tema específico de cada pesquisa.

Referências

Analyse Lexicale par Context d'un Ensemble de Segments de Texte 2009: un logiciel d'analyse de données textuelles. Manuel d'utilisateur [Computer and manual software]. (2009). Toulouse, France: Sociétè Image.

Camargo, B. V. (2005). ALCESTE: Um programa informático de análise quantitativa de dados textuais. In A. S. P. Moreira, B. V. Camargo, J. C. Jesuíno, & S. M. Nóbrega (Eds.), *Perspectivas teórico-metodológicas em representações sociais* (pp. 511-539). João Pessoa, PB: Editora da Universidade Federal da Paraíba.

Chartier, J.-F., & Meunier, J.-G. (2011). Text mining methods for social representation analysis in Large Corpora. *Papers on Social Representations*, *20*(37), 1-47.

Cibois, P. (1990). *L'analyse des données en sociologie.* Paris: Presses Universitaires de France.

Lahlou, S. (1994). L'analyse lexicale. *Variances*, (3), 13-24.

Lahlou, S. (2012). Text mining methods: An answer to Chartier and Meunier. *Papers on Social Representations*, *20*(38), 1-7.

Lebart, L., & Salem, A. (1994). *Statistique textuelle.* Paris: Dunod.

Marchand, P., & Ratinaud, P. (2012). L'analyse de similitude appliqueé aux corpus textueles: les primaires socialistes pour l'election présiden-

tielle française. In *Actes des 11eme Journées internationales d'Analyse statistique des Données Textuelles. JADT 2012* (pp. 687-699). Liège, Belgique. Retrieved April 13, 2013, from http://lexicometrica.univ-paris3.fr/jadt/jadt2012/Communications/Marchand,%20Pascal%20et%20al.%20-%20L%27analyse%20de%20similitude%20appliquee%20aux%20corpus%20textuels.pdf

Nascimento, A. R. A., & Menandro, P. R. M. (2006). Análise lexical e análise de conteúdo: Uma proposta de utilização conjugada. *Estudos e Pesquisas em Psicologia, 6*(2), 72-88.

Ratinaud, P. (2009). IRAMUTEQ: Interface de R pour les Analyses Multidimensionnelles de Textes et de Questionnaires [Computer software]. Retrieved from http://www.iramuteq.org

Ratinaud, P., & Marchand, P. (2012). Application de la méthode ALCESTE à de "gros" corpus et stabilité des "mondes lexicaux": analyse du "CableGate" avec IraMuTeQ. In: *Actes des 11eme Journées internationales d'Analyse statistique des Données Textuelles* (pp. 835-844). Liège, Belgique. Retrieved April 13, 2013, from http://lexicometrica.univ-paris3.fr/jadt/jadt2012/Communications/Ratinaud,%20Pierre%20et%20al.%20-%20Application%20de%20la%20methode%20Alceste

Reinert, M. (1990). ALCESTE, une méthodologie d'analyse des données textuelles et une application: Aurélia de G. de Nerval. *Bulletin de Méthodologie Sociologique*, (28), 24-54.

Système Portable pour l'Analyse des Données. Guide de l'utilisateur [Computer and manual software]. (2008). Courvoie, France: Coheris SPAD.

Veloz, M. C. T., Nascimento-Schulze, C. M., & Camargo, B. V. (1999). Representações sociais do envelhecimento. *Psicologia: Reflexão e Crítica*, *12*(2), 479-501.

Vergès, P., Junique, C., Barbry, W., Scano, S., & Zeliger, R. (2002). *Ensembles de programmes permettant l'analyse de similitude de questionnaires et de données numeriques*. Aix en Provence, France: Université Aix en Provence.

Vergès, P., Scano, S., & Junique, C. (2002). *Ensembles de programmes permettant l'analyse des evocation*s. Aix en Provence, France: Université Aix en Provence.

Recebido: 17/04/2012
Aceite final: 02/05/2013