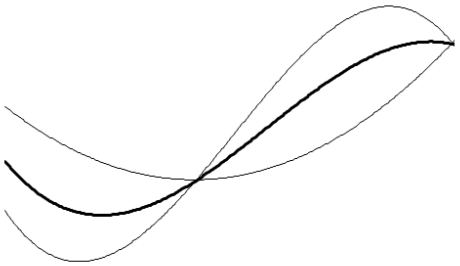
UNIVERSIDADE FEDERAL DE MINAS GERAIS
ESCOLA DE CIÊNCIA DA INFORMAÇÃO

LUIZ ANTÔNIO LOPES MESQUITA



SINTAGMAS NOMINAIS NA INDEXAÇÃO AUTOMÁTICA: uma análise estrutural da distribuição de termos relevantes em teses de doutorado da UFMG.

Belo Horizonte
2012

LUIZ ANTÔNIO LOPES MESQUITA

SINTAGMAS NOMINAIS NA INDEXAÇÃO AUTOMÁTICA: uma análise estrutural da distribuição de termos relevantes em teses de doutorado da UFMG.

Dissertação apresentada ao Programa de Pós-Graduação em Ciência da Informação da Escola de Ciência da Informação da Universidade Federal de Minas Gerais para obtenção do grau de Mestre em Ciência da Informação.

Linha de Pesquisa: Organização e Uso da Informação

Orientador: Prof. Dr. Renato Rocha Souza

Co-orientadora: Profª. Drª. Renata Maria Abrantes Baracho Porto

BELO HORIZONTE

2012

M582s Mesquita, Luiz Antônio Lopes.
Sintagmas nominais na indexação automática [manuscrito] : uma análise estrutural da distribuição de termos relevantes em teses de doutorado da UFMG / Luiz Antônio Lopes Mesquita. – 2012.
261 f., enc. : il.

Orientador: Renato Rocha Souza.
Coorientadora: Renata Maria Abrantes Baracho Porto.
Dissertação (Mestrado) – Universidade Federal de Minas Gerais, Escola de Ciência da Informação.
Referências: f. 116-127.
Apêndices: f. 128-261.

1. Ciência da informação – Teses. 2. Indexação automática – Teses. 3. Recuperação da informação – Teses. 4. Linguagens de indexação – Teses. I. Título. II. Souza, Renato Rocha. III. Porto, Renata Maria Abrantes Baracho. IV. Universidade Federal de Minas Gerais, Escola de Ciência da Informação.

CDU: 025.4

Ficha catalográfica: Biblioteca Profª Etelvina Lima, Escola de Ciência da Informação da UFMG.

UFMG

Universidade Federal de Minas Gerais
Escola de Ciência da Informação
Programa de Pós-Graduação em Ciência da Informação

FOLHA DE APROVAÇÃO

"SINTAGMAS NOMINAIS NA INDEXAÇÃO AUTOMÁTICA: UMA ANÁLISE ESTRUTURAL DA DISTRIBUIÇÃO DE TERMOS RELEVANTES EM TESES DE DOUTORADO DA UFMG"

Luiz Antônio Lopes Mesquita

Dissertação submetida à Banca Examinadora, designada pelo Colegiado do Programa de Pós-Graduação em Ciência da Informação da Universidade Federal de Minas Gerais, como parte dos requisitos à obtenção do título de **"Mestre em Ciência da Informação"**, Linha de Pesquisa: **"Organização e Uso da Informação - OUI"**.

Dissertação aprovada em: 19 de dezembro de 2012.

Por:

Prof. Dr. Renato Rocha Souza - FGV/RJ (Orientador)

Profa. Dra. Renata Maria Abrantes Baracho Porto - ECI/UFMG (Co-orientadora)

Prof. Dr. Wagner Meira Júnior - DCC/ICEX/UFMG

Profa. Dra. Gercina Ângela Borém de Oliveira Lima - ECI/UFMG

Aprovada pelo Colegiado do PPGCI

Profa. Renata Maria Abrantes Baracho Porto
Coordenadora

Versão final Aprovada por

Prof. Renato Rocha Souza
Orientador

UFMG

Universidade Federal de Minas Gerais
Escola de Ciência da Informação
Programa de Pós-Graduação em Ciência da Informação

ATA DA DEFESA DE DISSERTAÇÃO DE **LUIZ ANTÔNIO LOPES MESQUITA**, matrícula: 2010654603

Às 9:00 horas do dia 19 de dezembro de 2012, reuniu-se na Escola de Ciência da Informação da UFMG a Comissão Examinadora aprovada pelo Colegiado do Programa de Pós-Graduação em Ciência da Informação em 04/12/2012, para julgar, em exame final, o trabalho intitulado ***Sintagmas nominais na indexação automática: uma análise estrutural da distribuição de termos relevantes em teses de doutorado da UFMG***, requisito final para obtenção do Grau de MESTRE em CIÊNCIA DA INFORMAÇÃO, Área de Concentração: Produção, Organização e Utilização da Informação, Linha de Pesquisa: Organização e Uso da Informação - OUI. Abrindo a sessão, o Presidente da Comissão, Prof. Dr. Renato Rocha Souza, após dar conhecimento aos presentes do teor das Normas Regulamentares do Trabalho Final, passou a palavra ao candidato para apresentação de seu trabalho. Seguiu-se a argüição pelos examinadores com a respectiva defesa do candidato. Logo após, a Comissão se reuniu sem a presença do candidato e do público, para julgamento e expedição do resultado final. Foram atribuídas as seguintes indicações:

Prof. Dr. Renato Rocha Souza - Orientador	APROVADO
Profa. Dra. Renata Maria Abrantes Baracho Porto - Co-orientadora	APROVADO
Prof. Dr. Wagner Meira Júnior	APROVADO
Profa. Dra. Gercina Ângela Borém de Oliveira Lima	APROVADO

Pelas indicações, o candidato foi considerado APROVADO.

O resultado final foi comunicado publicamente ao candidato pelo Presidente da Comissão. Nada mais havendo a tratar, o Presidente encerrou a sessão, da qual foi lavrada a presente ATA que será assinada por todos os membros participantes da Comissão Examinadora.

Belo Horizonte, 19 de dezembro de 2012

Prof. Dr. Renato Rocha Souza
FGV/RJ (orientador)

Profa. Dra. Renata Maria Abrantes Baracho Porto
ECI/UFMG (co-orientadora)

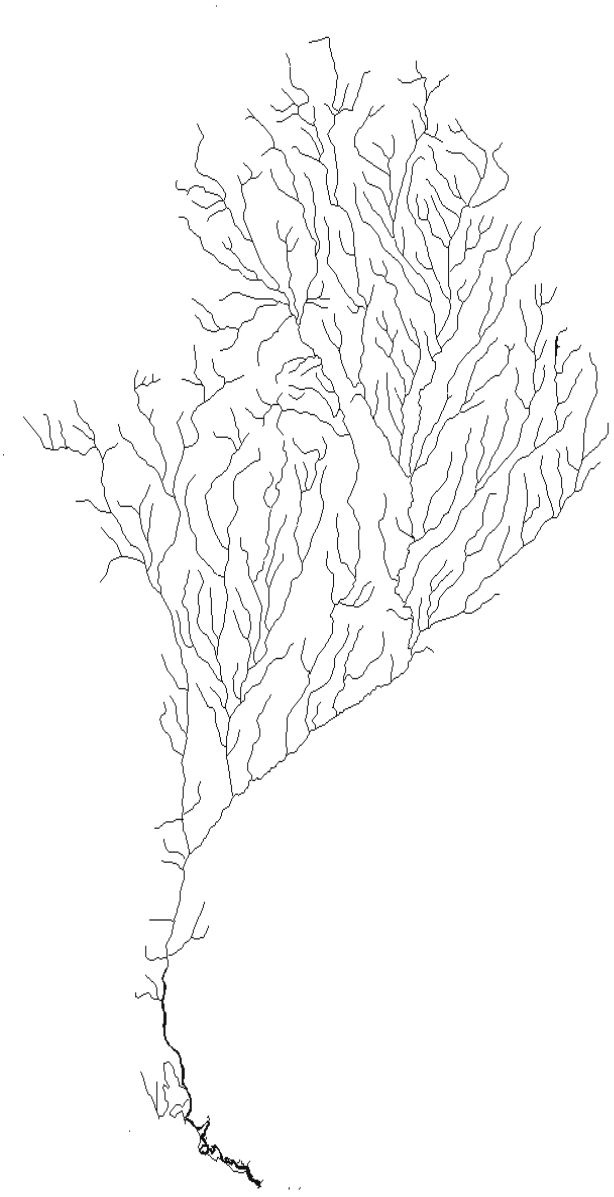
Prof. Dr. Wagner Meira Júnior
DCC/ICEX/UFMG

Profa. Dra. Gercina Ângela Borém de Oliveira Lima
ECI/UFMG

Profa. Renata Maria Abrantes Baracho Porto
coordenadora do Programa de Pós-Graduação em Ciência da Informação - ECI / UFMG

Obs: Este documento não terá validade sem a assinatura e carimbo da Coordenadora.

DEDICATÓRIA



Àqueles que foram meus afluentes de alguma forma com suas ideias, opiniões, indagações, conhecimentos, informações, sentimentos, atitudes, trabalhos, obras, ou até mesmo com sua simples presença direta, silenciosa ou longínqua.

Aos que estiveram presentes desde antes do início, como meus pais e irmãos. Aos que foram se somando com o tempo, como amigos, colegas de estudo, cunhados, sobrinhos, e colegas de trabalho.

Dedico principalmente àqueles que eu possa, através dessa dissertação e dos frutos dela, contribuir, mesmo que com um pingo de água, na formação de outros caminhos para uma busca nos oceanos do conhecimento.

AGRADECIMENTOS

Aos meus orientadores:

Renato Rocha Souza
Renata Maria Abrantes Baracho Porto

Aos professores que também colaboraram diretamente no projeto e concepção dessa dissertação:

Beatriz Valadares Cendón
Gercina Ângela Borém de Oliveira Lima
Heliana Ribeiro de Mello
Maria Aparecida Moura
Maria Guiomar da Cunha Frota
Maryualê Malvessi Mittmann
Nair Yumiko Kobashi
Ricardo Hiroshi Caldeira Takahashi
Wagner Meira Júnior

Aos demais professores que tive também o prazer do convívio mais próximo:

Alcenir Soares dos Reis
Alessandro Ferreira Costa
Carlos Alberto Ávila Araújo
Cátia Rodrigues Barbosa
Júlia Gonçalves da Silveira
Lídia Alvarenga

Aos colegas e amigos que tive também a oportunidade de conhecer e com quem pude estreitar laços:

Agnaldo Lopes Martins
Ariane Barbosa Lemos
Christiano Pereira Pessanha
Clotildes Madalena de Avelar Teixeira
Daniela Lucas da Silva
Edson Marchetti da Silva
Fernando Hadad Zaidan
Flávia Virgínia Melo Pinto
Izabel França de Lima
Joel Augusto de Oliveira
Joice Rodrigues Teixeira
José Alimatéia de Aquino Ramos
Juliana Horta de Assis Pinto
Juliana Moreira Pinto
Kátia Cardoso Coelho
Lilian Emanueli Marques
Lívia Ferreira Coutinho
Luciana Emirena dos Santos Carneiro
Luiz Cláudio Gomes Maia
Maria de Fátima Pinto Coelho
Maria Inês Moreira Sepúlveda
Mateus Uerlei Pereira da Costa
Max Cirino de Mattos
Paula Emanuelle Souza
Pedro Alves Barbosa Neto
Rafael Oliveira de Ávila
Raísa Mendes Fernandes de Souza
Rodrigo Moreno Marques
Tatiane Krempser Gandra
Wesley Rodrigo Fernandes

Aos profissionais com cuja ajuda também sempre pude contar:

Gilma Pereira
Gisele Reis
Lucimary Souto de Oliveira Silva
Nely Ferreira
Wanda de Andrade Lara

Às entidades:

Universidade Federal de Minas Gerais (UFMG)
Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq)
Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (Capes)

E, por fim, aos noventa e oito entrevistados que gentilmente também colaboraram com esta pesquisa:

Adriane Maria Arantes de Carvalho
Alexandre Martins Costa Santos
Aline Choucair Vaz
Aline de Marco Viott
Ana Cristina Passos de Paiva Bello
Ana Luiza de Quadros
Ana Paula Ladeira
Andre Belico de Vasconcelos
Andrea Maria Favilla Lobo
Andreia de Assis Ferreira
Andres Manuel Villafuerte Oyola
Andrezza Fernanda Santiago

Breno Rates Azevedo
Bruno Campos de Carvalho
Carlos Alberto Llanes Leyva
Carolina Furtado Torres da Silva
Cecilia Vieira do Nascimento
Celia da Consolacao Dias
Charles Antonio de Paula Bicalho
Daniel Goncalves Chaves
Danielle Gomes Passos Silva
Debora Costa Reis
Debora Fernandes Almeida
Denilson Jose do Carmo
Denis Leandro Francisco
Diana Quintao Lima de Oliveira
Diva Souza Silva
Ednaceli Abreu Damasceno
Edson Jose Carpintero Rezende
Eduardo Henrique Martins Nunes
Eudes Lorencon
Fabio Augusto Rodrigues e Silva
Felipe Masiero Salvarani
Fernando Andrade Souza
Fernando Castro de Oliveira
Fernando Skackauskas Dias
Frederico Cesar Mafra Pereira
Gabrieli Lessa Parrilha
Geide Rosa Coelho
Geraldo Marcio da Costa
Guilherme Rocha Pereira
Helga Gabriela Aleme
Hernan Oliver Daza Gutierrez
Jacques Fux
Janaina Cecilia Oliveira Villanova
Joana Ziller de Araujo Josephson
Jorge Andre Matias Martins
Jose Quintao de Oliveira
Josiley Francisco de Souza
Juarez Fabiano de Alkmim Filho
Juliano Cezar Minardi da Cruz
Julio Cesar Machado de Paula
Karla Emilia de Sa Rodrigues
Karla Moreira Vieira
Leonardo Augusto de Almeida
Lorene dos Santos
Lucia de Fatima Melo
Ludmilla Zago Andrade
Luiz Megale
Magda Francisca Goncalves Rocha
Marcelo Machado Viana
Marcia Ambrosio Rodrigues Rezende
Maria Elisa Rodrigues Moreira
Maria Ines Barreiros Senna
Maria Luciana Brandao Silva
Maria Tereza Gomes de Almeida Lima
Mariana Thiengo
Mario Luis Cabello Russo
Marlécio Maknamara
Marlice de Oliveira e Nogueira
Marta Ribeiro dos Santos
Matheus da Cruz e Zica
Mercia Aleide Ribeiro Leite
Musso Garcia Greco
Nelicio Faria de Sales
Paulo Custodio Furtado Cruzeiro
Paulo Eduardo Ferian
Paulo Henrique Dias Menezes
Priscilla Rochele Barrios
Priscylla Tatiana Chalfun Guimaraes
Renata de Castro Martins
Renata Melo Moreira
Renata Silva Bergo
Renato Pereira de Andrade
Ricardo Bezerra Cavalcante
Rodrigo Drumond
Rogerio Oliveira Rodrigues
Romero Alves Teixeira
Rosane da Silva Gomes
Rosangela Ramos Corgosinho
Sandra Goulart Santos
Shirlei Rezende Sales
Simone Aparecida Fernandes
Tatiane Alves da Paixao
Valeria Barbosa de Resende
Vanessa Ferraz Almeida Neves
Viviane Aguiar Andrade
Viviane Mota Bispo

“Rem tene, verba sequentur.”[1]

(Marcus Porcius Cato)

"As palavras são como um dedo apontando para a Lua;
cuida de saber olhar para a Lua,
não se preocupe com o dedo que a aponta.”

(fragmento de um conto zen budista)

[1] “Retenha o conceito, as palavras vêm em seguida.”, ou ainda "*Abbi chiaro il concetto, e le parole verranno da sole*", “Tenha claro o conceito, e as palavras virão sozinhas”.

RESUMO

O objetivo principal dessa dissertação foi analisar se haveria um comportamento característico de distribuição de termos relevantes ao longo de um texto científico que poderia contribuir como um critério para o processo da sua indexação automática. A distribuição foi analisada de duas formas: uma linear, realizada do início ao fim do texto; e outra que considera algumas de suas partes estruturais (introdução, desenvolvimento e conclusão). Os termos considerados aqui foram somente sintagmas nominais plenos contidos nos próprios textos. Os textos considerados foram um total de 98 teses de doutorado das oito áreas de conhecimento da UFMG. Inicialmente, para cada um dos textos, foram selecionados 20 sintagmas nominais como candidatos a descritores. Os próprios autores das teses, mediante entrevistas, avaliaram a relevância de cada um deles como descritor de suas obras. 77,9% dos candidatos foram considerados relevantes. Os valores de relevância dos descritores foram associados às suas posições no texto. Foram analisados os valores resultantes dessa distribuição considerando dois tipos de posição: uma linear, com valores consolidados em dez partes iguais e consecutivas; outro considerando partes estruturais do texto (como introdução, desenvlvimento e conclusão). Todos os textos apresentaram um comportamento característico único, assim como um comportamento característico quando estavam relacionados às ciências naturais ou às ciências sociais. Todos os comportamentos, inclusive o geral, foram caracterizados em equações polinomiais e podem ser aplicados como critério para indexação automática.

Palavras-chave: linguística computacional; texto científico – estrutura e distribuição de termos relevantes; processamento de linguagem natural; indexação automática; sintagmas nominais.

ABSTRACT

The main goal of this thesis was to analyze whether there was a characteristic behavior regarding the distribution of relevant terms through a scientific text that could contribute as a criterion for its automatic indexing process. The distribution was analyzed in two ways: a linear one, performed from the beginning to the end of the text; and another that considered some of its structural parts (introduction, development and conclusion). The terms considered here were only nominal phrases contained in the texts. The texts considered here are a total of 98 doctoral dissertations from the eight knowledge areas of UFMG. Initially, for each text, 20 nominal phrases were selected as candidates for descriptors. The authors of the theses, through interviews, rated the importance of each nominal phrase as a descriptor of his/her work. 77.9% of candidates were considered relevant. The descriptors' relevance values were associated with their positions in the text. We analyzed the resulting values of this distribution considering two types of position: a linear one, where values were consolidated into ten equal and consecutive portions; and one considering other structural parts of the text (such as introduction, development and conclusion). All texts showed a unique and characteristic behavior, as well as a characteristic behavior when the text was related to the natural sciences or social sciences. All behaviors, including general, were characterized in polynomial equations and can be applied as a criterion for automatic indexing.

Keywords: computational linguistics; scientific text – structure and distribution of relevant terms; natural language processing; automatic indexing; noun phrases.

LISTA DE GRÁFICOS

LISTA DE TABELAS

LISTA DE EQUAÇÕES

LISTA DE ABREVIATURAS

ANNOD - *A Navigator of Natural Language Organized Data*

BDTD - Biblioteca Digital de Teses e Dissertações

CGI.br - Comitê Gestor da Internet no Brasil

CSN - Categoria do Sintagma Nominal

FASIT - *Fully Automatic Syntatically based Indexing Text*

IBM - *International Business Machines Corporation*

ICSI - *International Conference on Scientific Information*

IUPAC – *International Union of Pure and Applied Chemistry*

KWIC – *Key-word-in-context*

PRECIS – *Preserved Context Indexing System*

RI – Recuperação da Informação

SMART - *System for the Mechanical Analysis and Retrieval of Text*

SN - Sintagma Nominal

SPIRIT – *Système Syntaxique et Probabiliste d'Informations Textuelles*

SRI – Sistema de Recuperação da Informação

TF – *Term Frequency*

UFMG – Universidade Federal de Minas Gerais

SUMÁRIO

1 Introdução

A atuação militar durante a II Guerra Mundial intensificou a pesquisa científica como nunca antes na história. Nesse momento, surgiu um cenário no qual cientistas em várias nações passaram a somar esforços numa mesma direção e a inventar artefatos que influenciaram a humanidade (BUSH, 1945). O computador foi um desses inventos que, dessa época em diante, tornou-se uma das principais tecnologias que caracteriza a Revolução da Tecnologia da Informação iniciada no século XX (CASTELLS, 1999). A máquina imaginária de *memória estendida* denominada MEMEX, idealizada por Bush (1945) como uma rede de comunicação interativa no espaço e no tempo, concretizou-se com as redes de computadores. O paradigma da tecnologia da informação consolidou-se mundialmente na sociedade do século XXI com a Internet. Na primeira década deste milênio, o número de internautas cresceu em quase cinco vezes, chegando a cerca de 30% da população mundial (WIUPS, 2011). No Brasil, somente entre os anos de 2008 a 2010, a proporção da população total do país que é usuária da Internet passou de 34% para 41%, de acordo com o Comitê Gestor da Internet no Brasil (CGI.br, 2011).

Segundo Wersig (1993), se a imprensa de Gutenberg do século XV já propiciou para a humanidade um *dilúvio de literatura*, esse crescimento vertiginoso do uso das tecnologias da informação e da comunicação traz benefícios e preocupações. Saracevic (1996) avisa que, no final do século passado, os grandes sistemas de informação, inclusive as bibliotecas, arriscam-se a serem transformadas de uma casa do tesouro em armazém, e deste, em depósito de sucata.

A Ciência da Informação surge nesse momento decorrente das tecnologias novas e mais complexas do pós-guerra, contendo *tanto um componente de ciência pura* quanto *um componente de ciência aplicada* para investigar as propriedades e o comportamento da informação, as forças que governam seu fluxo e os meios para otimizar sua acessibilidade e uso (BORKO, 1968; WERSIG, 1993). A Recuperação da Informação (RI), como uma das áreas da Ciência da Informação, tenta resolver o problema da explosão informacional apontada por Bush (1945). A Ciência da Computação também “desenvolve significativas pesquisas nessa área com o objetivo principal de prover aos usuários de seus sistemas um fácil acesso à informação do seu interesse” (BAEZA-YATES; RIBEIRO-NETO, 2011, p. 1, tradução do autor). Dentre muitas outras áreas que tornam a RI interdisciplinar, a Linguística também contribui significtivamente para o processamento de informações textuais em linguagem natural.

O sucateamento informacional apontado por Goethe (WERSIG, 1993) é pertinente, uma vez que os critérios utilizados pelos sistemas de recuperação da informação (SRIs) podem fazer com que alguns documentos, que seriam de interesse dos usuários,

fiquem mais escondidos nos acervos que outros que possuem características mais favoráveis para tais critérios desses sistemas. Diante de volumes gigantescos de documentos, como aqueles digitais possibilitados pelas redes de computadores, é apresentado ao usuário uma quantidade muito grande de documentos como resultado de uma busca. Na maioria dos casos, o usuário tende a escolher somente os primeiros resultados, deixando de lado os demais que aparecem ao final dessa listagem ordenada.

A simples variação em um critério utilizado para ordenar os documentos como resultado de uma busca pode levar um usuário a utilizar documentos muito diferentes entre si, uma vez que há uma tendência em se utilizar somente uma primeira parte de uma listagem de resultados. O principal critério para a apresentação dos resultados de uma busca é a correspondência entre esta e os termos usados para indexar o documento em um acervo. Logo a indexação é uma das etapas mais importantes em um SRI.

Existem duas principais formas de indexação: a manual, feita por profissionais especialistas, e a automática, realizada por computadores. Esta última forma automatizada mostra-se mais vantajosa, especialmente diante de grandes volumes de informação digital. Salton (1972) apresentou, ao desenvolver um dos primeiros grandes SRIs, que não havia razões técnicas óbvias para a não substituição dos métodos manuais de indexação por métodos automáticos.

Conforme Sayão (1985), a indexação automática começou a ganhar notoriedade com as publicações de Luhn (1957)[2]. Muitos autores contribuíram para a evolução dessa área de pesquisa nas suas primeiras décadas: Baxendale (1958), Swanson (1962, 1963), Borko (1968), Salton (1967, 1968, 1971a, 1971b), Van Rijsbergen (1971), Sparck Jones (1972, 1973, 1978, 1979), Field (1975, 1977), Dillon (1982), Robredo (1980, 1982a, 1982b) e outros. Atualmente existem inúmeros critérios para a indexação automática, sendo que ainda prevalece aqueles apontados no início de sua história, como o uso da frequência de palavras isoladas.

Com o crescimento na área da Ciência da Computação, criou-se algoritmos mais otimizados e processadores mais rápidos[3]; as pesquisas com indexadores automáticos puderam utilizar estruturas lingüísticas mais complexas; sendo uma delas o sintagma nominal (SN). Tal estrutura, de acordo com Perini *et al*. (1996), possui maior valor semântico que a palavra isolada e foi usada para a língua portuguesa por Kuramoto (1999) em sua tese de doutorado. A partir desses estudos, Souza (2005) propôs uma metodologia de escolha automática de SNs como descritores relevantes no processo de indexação

[2] Inicialmente, Luhn (1957) adotava terminologias como *auto-resumo* e *auto-indexação*. Posteriormente esses termos foram substituídos por indexação automática.

[3] Normalmente, menciona-se a "lei de Moore", segundo a qual se acredita que, a cada 18 meses, o número de transistores em um processador deve dobrar e ter seu custo mantido. Tal afirmativa é atribuída a Gordon Moore, então presidente da Intel, fabricante de processadores tais como os usados para desenvolver esta pesquisa.

automática. Esta metodologia foi utilizada por Maia (2008) para o desenvolvimento de uma ferramenta[4] que, dentre outras funcionalidades, extrai tais SNs de forma automática.

O uso do SN apresenta uma significativa evolução para a indexação automática, no entanto, os critérios para a seleção desses sintagmas como descritores utilizados até então ainda são baseados principalmente naqueles das primeiras décadas da indexação automática. Ao final de seu trabalho, o autor prevê que:

> a possibilidade de melhores métodos considerando uma análise de densidade informacional dos sintagmas nominais no documento. As considerações relativas à analise de densidade informacional podem ser incorporadas à metodologia, de maneira que os parsers apresentem algum tipo de ponderação que leve em conta as seções mais importantes do documento (SOUZA, 2005, p. 138).

Para Borges (2009), existem cerca de 16 classes diferentes de critérios para indexação automática. Algumas delas ainda são pouco exploradas, como aquelas referentes a *posição do termo no texto* e o de *tópico frasal (palavras sugestivas)*. Esse segundo critério baseia-se em Baxendale (1958) que aponta, por exemplo, que um termo presente no início ou no final de uma parte textual tem 85% de chances de ser seu descritor. Tais critérios remetem à noção de estrutura no sentido amplo da "relação entre elementos e entre as partes de um todo. [...] que permite distinguir o essencial do acessório" (ORTEGA; LARA, 2010, p. 11) e podem acrescentar à atividade da indexação automática, como linguagem documentária por meio automático, mais qualidade no sentido pragmático (KOBASHI; FERNANDES, 2009).

Baeza-Yates e Ribeiro-Neto (2011) apresentam o uso da informação estrutural de textos em diferentes estágios do processo de recuperação da informação, inclusive no da indexação. São atribuídos a seções, subseções e parágrafos, por exemplo, os elementos estruturais relativos à *posição do termo no texto*. Shah *et al.* (2003) e Galeas, Kretchmer e Freisleben (2009), dentre outros, utilizam em suas pesquisas a posição do termo ao longo de um texto como critério para pontuar a sua relevância como descritor. A posição do termo é considerada de duas formas: a posição linear do termo em relação a todo o texto, desde a primeira palavra até a última (medida em % em relação ao tamanho do texto medido em quantidade de palavras); e a posição em uma estrutura delimitada do texto (como seção de introdução, desenvolvimento ou conclusão, por exemplo).

Alguns sistemas atuais permitem a utilização da posição como critério para a recuperação da informação, no entanto a grande maioria desses sistemas é baseada na

[4] A ferramenta de Maia (2008) se chama Ogma. Existem várias ferramentas de processamento de linguagem natural para a língua portuguesa, dentre elas pode-se destacar o sistema Palavras de Bick (2000), que é fruto de uma tese de doutorado para a análise automática gramatical da língua portuguesa. Ambas permitem a extração dos sintagmas nominais presentes em textos eletrônicos, sendo que a última possui a vantagem de ainda prover informações relativas ao posicionamento estrutural dos sintagmas nominais no texto em relação a frases e parágrafos, inclusive. Outra ferramenta significativa é o MHTX que é decorrente das pesquisas de Lima (2010) em análise facetada e mapas conceituais.

língua inglesa. A língua portuguesa possui substanciais diferenças para com o inglês para que tais ferramentas sejam facilmente adaptadas a ela. Logo, faz-se necessária a criação de conhecimento, não apenas sobre, mas para a língua portuguesa com o uso de tais ferramentas. A partir da análise desses critérios de posição dos SNs em um texto em português, podemos chegar a métodos de escolha automática de descritores que sejam mais relevantes do que simplesmente a sua frequência no texto ou a quantidade total de documentos em que eles ocorrem. As ferramentas e as pesquisas aqui citadas abrem campo para a investigação dessa área na indexação automática dentro da Ciência da Informação.

1.2 Objetivos

O objetivo principal desta pesquisa é investigar a existência de um comportamento de distribuição de termos relevantes ao longo de um texto que possa favorecer à sua indexação automática. A distribuição aqui se refere a duas formas: uma linear, que vai do início ao fim do texto, termo a termo; e outra que considera algumas de suas partes estruturais (introdução, desenvolvimento e conclusão). Os termos considerados aqui são somente SNs contidos nos próprios textos. Os textos considerados aqui são teses de doutorado das oito áreas de conhecimento da Universidade Federal de Minas Gerais (UFMG).

1.2.1 Objetivo geral

O objetivo geral desta pesquisa é analisar se há um comportamento característico de distribuição de termos relevantes ao longo de um texto científico que possa contribuir como um critério para o processo de sua indexação automática.

1.2.2 Objetivos específicos

Esta pesquisa também tem como objetivos específicos:

1. Analisar características linguísticas quantitativas que diferenciam as teses de doutorado das oito áreas de conhecimento da UFMG que podem interferir na extração automática de SNs;
2. Desenvolver um protótipo para a seleção automática de SNs como candidatos a descritores utilizando um processador de linguagem natural;
3. Examinar os fatores que influenciam o processo de seleção automática de SNs como candidatos a descritores;

4. Verificar os principais fatores linguísticos, de forma quantitativa, que influenciam nas diferenças de distribuição de termos relevantes ao longo dos textos e nas partes estruturais (introdução, desenvolvimento e conclusão) das teses de doutorado das oito áreas de conhecimento da UFMG;
5. Determinar funções matemáticas de distribuição de termos relevantes ao longo dos textos das teses de doutorado da UFMG das oito áreas de conhecimento.

2 Conceitos gerais e revisão da literatura

Esta pesquisa envolve três principais áreas: linguística, processamento de linguagem natural e indexação automática. Neste capítulo são apresentados os principais conceitos para a compreensão da pesquisa relativos a tais áreas. O aporte linguístico foi utilizado aqui para fundamentar a utilização dos SNs como descritores e para analisar a sua distribuição nas partes estruturais do texto. O processamento de linguagem natural teve importância para a elaboração da metodologia, assim como para a elaboração e uso de ferramentas para o processamento dos textos. A indexação automática, tema central desta pesquisa, é tratada aqui de forma detalhada nos seus aspectos históricos e que delinearam as principais técnicas de recuperação da informação utilizadas aqui.

1.1 Conceitos linguísticos

Um sistema linguístico é a língua comum entre todos os membros de uma mesma comunidade linguística. Sua atualização ocorre de acordo com o comportamento linguístico dos indivíduos dessas comunidades, sendo que, cada um desses, por sua vez, pode ter, um nível de competência linguística que está relacionado ao grau de conhecimento que esse indivíduo tem do sistema linguístico (LYONS, 1987).

Uma língua pode ser descrita de forma diacrônica, ou seja, considerando-se as mudanças sucessivas que ela sofre ao longo do tempo, a cada etapa histórica constatada. Em um mesmo momento do tempo, uma língua pode ser descrita de forma sincrônica, quando se encontra estável (DUBOIS *et al*., 1973; LYONS, 1987). A diferença entre a diacronia e a sincronia através da descrição de um jogo de xadrez: para a evolução do jogo, considerando onde cada peça estava anteriormente e para onde foram em seguida, teríamos uma descrição diacrônica; para o arranjo em um determinado momento das peças no tabuleiro, teríamos uma descrição sincrônica (LYONS, 1987).

A estrutura esquemática de um texto científico é unânime na literatura: introdução, desenvolvimento e conclusão. A parte de desenvolvimento pode ser composta de formas distintas (materiais e métodos, resultados e/ou discussões), no entanto, a introdução possui a característica comum de ir de assunto geral para específico, assim como a conclusão, em sentido contrário, porém na mesma direção, vai de assunto específico para geral (FELTRIM; ALUISIO; NUNES, 2000).

Embora a introdução tradicionalmente consista principalmente em três estágios (contextualização, revisão bibliográfica e objetivos), é comum que autores usem de *intermináveis discursos*, demorem em descrições e análises históricas demasiado remotas

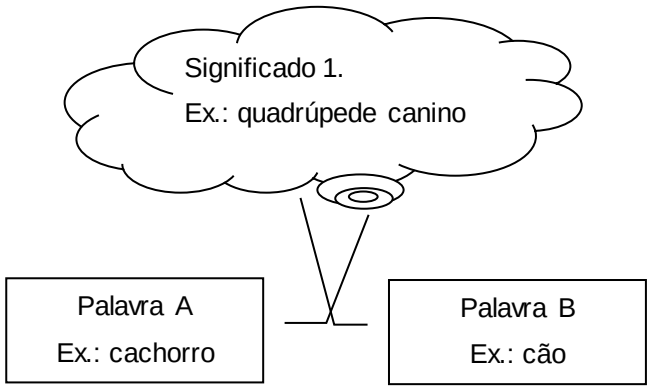
ou ainda descrevam detalhes dos resultados alcançados (FELTRIM; ALUISIO; NUNES, 2000).

A conclusão, além de finalizar o texto científico, conforme Feltrim, Aluisio e Nunes (2000) condensa todas as principais ideias desenvolvidas ao longo do mesmo e faz referências a assuntos abertos durante a introdução. Outra característica da conclusão é abrigar de forma direta as opiniões e visões do autor, assim como indicações de trabalhos futuros, que podem transcender as ideias desenvolvidas na pesquisa. Em todas as partes do texto ocorrem expressões que dependem do contexto para a determinação de seu significado. Segundo Lyons (1987), essas expressões são denominadas referenciais.

Como apresentado adiante, para a indexação automática, a frequência de um termo é usada como peso para determinar a sua relevância como descritor. Um problema que as expressões referencias geram para a indexação automática seria o fato de ocultar a real frequência de um assunto, pelo fato da expressão referencial possibilitar que termos distintos sejam usados para o mesmo assunto.

Para Cintra (2002) outro fator linguístico importante para a linguagem documentária usada na indexação é a *sinonímia* que corresponde ao fato de dois ou mais termos serem equivalentes. Esse fato, assim como para as expressões referenciais, possibilita que a frequência de um significado seja diluída em distintos termos. Os vocabulários controlados permitem minimizar esse fator, uma vez que têm como função normalizar essa distinção de termos para um mesmo significado representado por um só termo.

Figura 1 - Exemplo de sinonímia

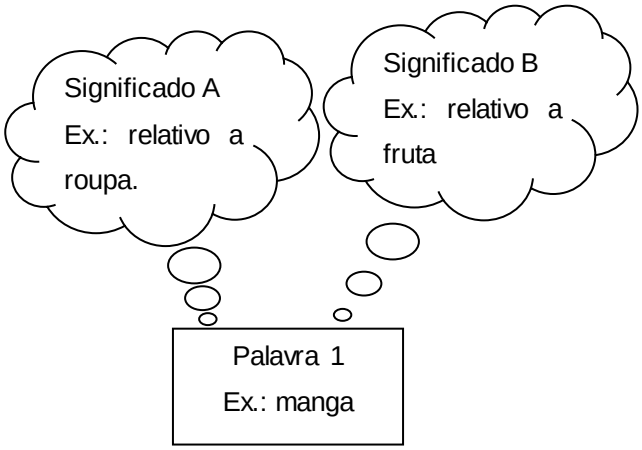


Fonte: Elaborado pelo autor.

Por outro lado, Cintra (2002) afirma que ao invés de dois ou mais termos serem referentes a um mesmo significado, como na sinonímia e nas expressões referenciais, a plurisignificação também pode tornar mais complexo um processo de indexação. A polissemia, a homonímia e a ambiguidade são exemplos de plurisignificação, uma vez que,

em todas elas, um mesmo termo pode possuir mais de um significado. A plurisignificação possibilita que um termo seja compreendido com um significado diferente do que foi a intenção do autor.

Figura 2 - Exemplo de plurisignificação



Fonte: Elaborado pelo autor.

Por fim, de acordo com Lyons (1987) outro conceito linguístico que é importante para a presente pesquisa é a estilística, que é um ramo da macrolinguística. Riffaterre (citado por DUBOIS, 1973, p. 243), afirma que “a língua exprime, o estilo sublinha”. Ou seja, esclarece Dubois (1973, p. 244) que “o estilo é caracterizado como uma marca individual do sujeito, uma gramática particular”. Bally (citado por DUBOIS, 1973, p. 237), define que a estilística é o “estudo dos fatos de expressão da linguagem organizada do ponto de vista de seu conteúdo afetivo, isto é, expressão dos fatos da sensibilidade pela linguagem e ação dos fatos de linguagem sobre a sensibilidade”.

Um texto, mesmo que dentro das normas de uma gramática, possui uma marca do indivíduo que o escreve. Essa marca pode ser percebida, por exemplo, através de desvios de um comportamento lógico esperado. A recorrência desses desvios pode caracterizar um estilo. Então, o estilo seria o lado negativo das estruturas gramaticais. Granger (citado por DUBOIS, 1973), em *Ensaio de uma Filosofia do Estilo*, amplia a noção de estilo para fora da literatura chegando a todas as construções científicas.

> A passagem do amorfo ao estruturado não é jamais o resultado da imposição de uma forma que vem toda constituída do exterior [...]. Toda a estruturação resulta de um trabalho que põe em relação, suscitando-os, a forma e o conteúdo do campo explorado (GRANGER, *apud* DUBOIS, 1973, p. 242).

Podemos considerar para esta pesquisa que qualquer autor imprime em seu texto, mesmo que científico, uma marca que pode ser atribuída à sua personalidade,

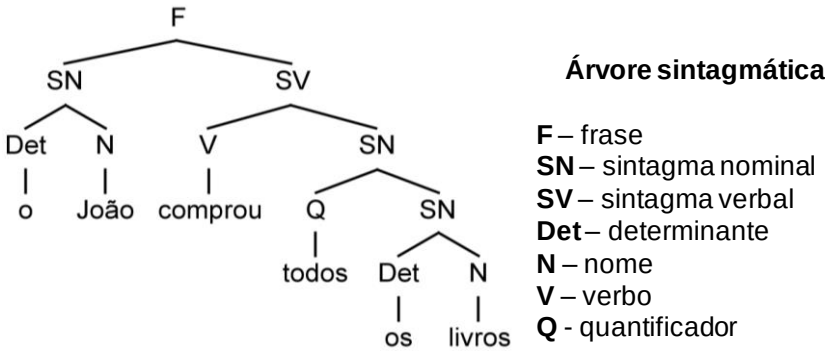
havendo até mesmo o conceito de estilometria que empregaria “a estatística para o estudo dos fatos do estilo” (DUBOIS, 1973, p. 245).

2.1.1 Sintagmas nominais

O SN ou [denominado em inglês *noun phrase*] é definido como a única unidade sintática capaz de funcionar como sujeito ou objeto nas orações da língua portuguesa, sendo normalmente construído com base em um substantivo. Uma forma de verificar se uma expressão é um SN consiste em tentar inseri-lo na seguinte moldura: _____ *sou* / *é* / *somos* / *são* / *bom* / *boa* / *bons* / *boas* (TRASK, 2004).

Abaixo, temos um exemplo de SN. É possível observar que existe a estrutura chamada de sintagma nominal aninhado. Na Figura 3, a seguir a expressão *todos os livros* possui tal estrutura, pois ele é composto por um outro SN (os livros) aninhado dentro dele.

Figura 3 - Exemplo de estrutura de um sintagma nominal



Fonte: Adaptado de OTHERO, 2009.

Outro exemplo é dado por Souza (2005), que apresenta a estrutura sintagmática de *As características do ambiente do mundo dos negócios* que engloba os SNs: *os negócios* (SN1), *o mundo dos negócios* (SN2), *o ambiente do mundo dos negócios* (SN3) e ele mesmo (SN4). Onde SN1 a SN4 correspondem aos níveis apresentados na Tabela 1, que apresenta ainda o nível SN5 para estruturas do seu tamanho ou maiores. O nível SN1 apresenta duas subdivisões, sendo a primeira correspondente à estrutura de um determinante D com um nome N (ex.: os negócios), e a segunda correspondente a qualquer estrutura, exceto essa (ex.: negócios mundanos).

Tabela 1 - Níveis das estruturas dos sintagmas nominais

N	Estrutura e Nível do SN
1ª	Nível 1, estrutura (D+N)
1b	Nível 1, exceto (D+N)
2	Nível 2
3	Nível 3
4	Nível 4
5	Nível 5 ou superior

Fonte: Adaptado de SOUZA, 2005.

Os SNs em um documento apresentam densidade informacional superior à palavras isoladas, mantendo maior proximidade do discurso contido nos documentos por eles descritos (KURAMOTO, 1996; SOUZA, 2005). “Palavras isoladas, como descritores, podem apresentar mais problemas de polissemia ou de plurisignificação” (LYONS, 1987, p. 140). Além de apresentarem menos influência dos problemas acima, “os sintagmas nominais trazem em seu bojo o contexto semântico dos discursos” (SOUZA, 2005, p. 136). Para Baeza-Yates e Ribeiro-Neto (2011) os substantivos, que compõem um SN, possuem maior valor semântico ao serem usados como termos de indexação. Portanto, o uso de SNs como termos de indexação pode apresentar melhores resultados que o uso de palavras isoladas.

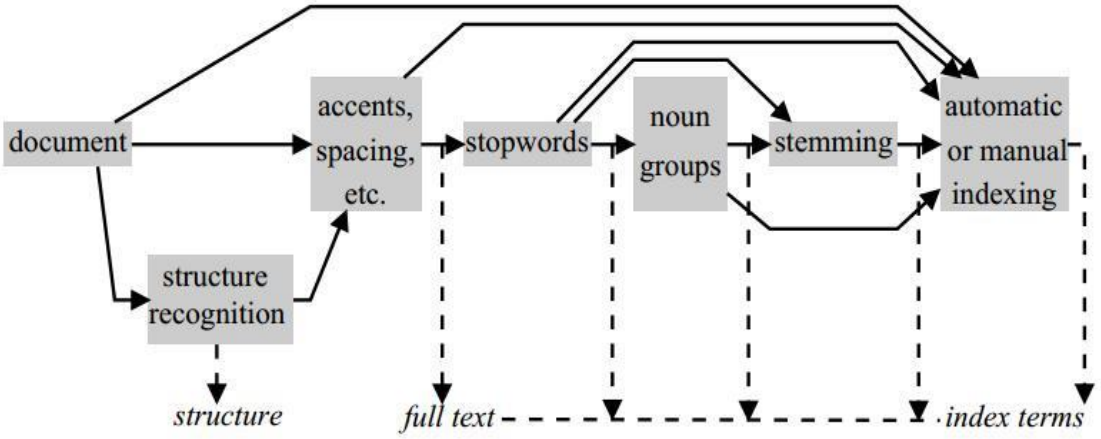
Os SNs podem ser extraídos automaticamente de textos. Os trabalhos de Kuramoto (1996), Souza (2005), Maia (2008), Corrêa *et al.* (2011) e outros apresentam como tema central a utilização de SNs através da sua extração em processadores de linguagem natural de forma semi e automática para a língua portuguesa. A seguir são apresentados alguns conceitos relativos a esses processadores.

1.2 Processamento de linguagem natural

Ladeira (2010), que selecionou e analisou a produção científica brasileira entre 1996 e 2003 relacionada ao processamento de linguagem natural, considera que tal área seja “responsável por manipular automaticamente a linguagem não controlada contida normalmente nos documentos textuais” (LADEIRA, 2010, p. 43).

Baeza-Yates e Ribeiro-Neto (2011) apresentam que um documento pode ser pré-processado seguindo cinco operações que são exemplificadas na Figura 4, a seguir.

Figura 4 - Fases de pré-processamento de um texto



Fonte: BAEZA-YATES; RIBEIRO-NETO, 1999.

Um texto é analisado essencialmente por suas palavras. Logo após um documento ser reconhecido como um texto, a primeira operação consiste na denominada análise léxica, que consiste no tratamento de acentuações (*accents*), espaços *(spacing*), marcas de pontuação, números, hífens etc. Alguns sistemas podem considerar ainda quebras de linha e quebras de parágrafos e realizar uma marcação estrutural.

As palavras que possuem baixa relevância para descrever um assunto ou para serem usadas como termos de indexação são denominadas *stopwords* (o conjunto dessas é denominado *stoplist*). A retirada dessas palavras pode ser feita através de uma *stoplist* ou por métodos estatísticos, como aquelas que ocorrem em todos os documentos e, portanto, não possuem características discriminatórias entre os mesmos.

Os SNs podem ser usados exclusivamente para representar todos os termos de um texto, uma vez que possuem maior valor semântico que qualquer outra estrutura sintagmática (como a verbal, adverbial, etc.).

O *stemming* consiste na transformação de uma palavra para a sua raiz. Uma técnica para isso consiste na retirada de prefixos e sufixos. O objetivo é reduzir as variações sintáticas de um mesmo termo como aquelas provocadas por mudanças de gênero, de grau, ou até mesmo para a redução à forma infinitiva de um verbo.

Finalmente, os termos restantes são eleitos como descritores através de um processo que pode ser automático ou manual. A decisão para que um termo pré-processado seja eleito automaticamente como um descritor de um texto envolve a utilização de critérios que são tratados detalhadamente a partir do subitem a seguir.

Uma coleção de recursos de informação pode aparecer sob a forma de um *corpus*, que consiste em uma coletânea de textos naturais, escolhidos para caracterizar um estado ou variedade de uma língua. Texto natural é aquele que ocorre espontaneamente na língua e que não foi criado com o propósito de figurar no *corpus.* Um *corpus* de artigos

científicos é mais propenso a ser utilizado para estudos de *parsers* (SARDINHA, 2004). Existem grupos de pesquisa que disponibilizam diversos *corpora* de acordo com o tipo de pesquisa a ser feita em processadores de linguagem natural, como os projetos Linguateca (SANTOS, 2009) e AC/DC (SANTOS; SARMENTO, 2002).

1.3 Descritores

O conceito de descritor utilizado aqui é referente ao termo que ressurgiu nos Estados Unidos da América após seu período de industrialização e veio substituir a numerosa terminologia usada até então para seu significado: "índice, cabeçalho de assunto, uni termo, termo coordenado, palavra-chave, frase-chave, indexação coordenada, etc." (SILVA, 1972, p. 28-29). Após a Segunda Guerra Mundial, a adoção do nome Ciência da Informação, que buscava inserir *indivíduos voltados à tecnologia e de fora da Biblioteconomia*, caracteriza o esforço em se utilizar as novas tecnologias para resolver problemas antigos (ORTEGA, 2004). "O termo *palavra-chave* passou a denotar também atividades automáticas, já o termo *descritor* passou a ser empregado mais especificamente no processo de tratamento automático da informação no que era então chamada de tecnologia da documentação"[5] (SILVA, 1972, p. 29).

Mooers (1947, citado por LANCASTER, 1968) passou a utilizar o termo descritor no seu sistema denominado Zato[6] para a classificação de assuntos de documentos a partir de palavras extraídas dos seus próprios textos. Taube (1951, citado por SILVA, 1972) apresentou um processo de indexação coordenada denominado *unitermo*, que também retira palavras ou termos únicos do próprio texto para classificação de assuntos. Posteriormente, Luhn (1957) lança o sistema *Key-word-in-context* (KWIC)[7], com base no princípio de Taube (1951 citado por SILVA, 1972), que retira do próprio título dos textos os seus termos descritores. Em tal sistema computadorizado, já é usado o conceito de *stoplist,* que consiste em uma lista de termos que não seriam descritores, tais como preposições, artigos, pronomes etc. Vários outros sistemas de indexação automática foram desenvolvidos desde então a partir da extração de termos do próprio texto dos documentos.

Lancaster (2004) apresenta as terminologias: indexação por atribuição (também denominada indexação derivada) e indexação por extração. A primeira é condicionada a um vocabulário controlado (como um tesauro) e a segunda é obtida através da extração de

---

[5] Para Ortega (2004) a Documentação tem como principal questão o "registro do conhecimento científico, a memória intelectual da civilização.

[6] O sistema de classificação Zato utilizado por Mooers dependia de cartões perfurados (SILVA, 1972).

[7] 100 anos antes do KWIC já havia na Alemanha, terra natal de Lhun, um sistema manual similar denominado *Schlagwort,* que significa palavra principal ou palavra-chave em alemão (SILVA, 1972).

termos livres do próprio texto. O resultado final de ambos processos são os termos de indexação que podem ser definidos como:

> Um termo de indexação é uma palavra ou grupo de palavras consecutivas em um documento. Em sua forma geral, um termo de indexação é qualquer palavra em uma coleção. Isto é uma interpretação usada por desenvolvedores de sistemas de busca. Em uma interpretação mais restrita, um termo de indexação é um grupo pré-selecionado de palavras que representa um conceito chave ou tópico em um documento. Isto é uma interpretação usada por bibliotecários e cientistas da informação (BAEZA-YATES; RIBEIRO-NETO, 2011, p. 61-62, tradução livre).

Baeza-Yates e Ribeiro-Neto (2011) apresentam uma distinção de definições de termo de indexação para aqueles mais relacionados às tecnologias da informação e aqueles mais relacionados à ciência da informação e biblioteconomia. A primeira definição pode ser considerada mais pragmática, uma vez que visa o desenvolvimento de um sistema, e a segunda, mais conceitual, que se aproxima da prática do indexador ao analisar assuntos.

Nesta pesquisa, a definição de termo de indexação é utilizada como sinônimo de descritor, e está mais relacionada ao processo de indexação automática apresentado a seguir.

1.4 Indexação automática

As origens da indexação têm seus primeiros indícios em processos de manipulação dos *papirus* egípicios e dos registros fiscais da Grécia Antiga. A obra de Aegidius Romanus, *Commentarius inprimum sententiarum*[8], do século XIV, apresenta em suas sete primeiras páginas uma indexação alfabética que já usa as palavras principais como descritores, ao invés de simplesmente as primeiras palavras de cada título (SILVA, 1972, p. 31-32).

A indexação pode ser definida como:

> [...] o processo de analisar o conteúdo informacional dos registros do conhecimento e sua expressão na linguagem do sistema de indexação. Ele implica: a) Selecionar os conceitos indexáveis de um documento; e b) Expressar esses conceitos na linguagem do sistema de indexação. (BORKO; BERNIER, 1978, p. 8)

A leitura que leva o indexador manual a eleger ou atribuir termos descritores a um texto envolve o próprio indexador como "sujeito e toda a sua capacidade subjetiva de interpretar" (DIAS; NAVES, 2007, p. 44). Atualmente, "é ressaltado o papel do leitor como produtor do sentido, numa dinâmica de forças que perpassa a relação do sujeito com o texto" (DIAS; NAVES, 2007, p. 45). O bibliotecário então, como profissional mais indicado

[8] A obra encontra-se na Biblioteca da Universidade Católica dos Estados Unidos, Washington, D.C.

para o exercício da indexação, acaba por refletir sua realidade social em tal eleição/atribuição de descritores.

> [...] a indexação deve ser considerada como um produto que reflete o processo pelo qual foi construído, tendo influências do bibliotecário, do tipo de biblioteca, da comunidade atendida, do vocabulário, da instituição, do próprio processo, do documento, entre outros. (COUTINHO, 2012, p. 77).

A análise de assunto, uma das áreas que estuda os aspectos subjetivos da prática da indexação manual, aponta que, dada a dimensão atual proporcionada pelos inúmeros documentos eletrônicos, sua tendência passa a ser também a de auxiliar a indexação automática.

> [...] a tendência da pesquisa em análise de assunto é no sentido de identificar os padrões de processamento dos indexadores de forma a não apenas auxiliar no aperfeiçoamento desses padrões, como também servir de insumo à cada vez mais necessária automatização do processo. (DIAS; NAVES, 2007, p. 105).

Além da inviabilidade do tratamento de grandes quantidades de documentos, os problemas práticos da atividade de indexação manual encontram-se também na inconsistência praticada pelos indexadores (DIAS; NAVES, 2007), que podem ser interindexadores e intraindexadores (BORKO, 1977). A inconsistência interindexadores ocorre quando dois ou mais indexadores elegem ou atribuem descritores diferentes para um mesmo documento. A inconsistência intraindexadores ocorre quando um mesmo indexador atribui descritores diferentes para um mesmo documento em momentos diferentes.

A indexação automática se justifica então pela sua capacidade de atender o crescente volume de documentos eletrônicos e de forma mais consistente que a manual. As pesquisas em indexação automática ganharam força após a Segunda Guerra Mundial, quando o espírito pragmático e o apoio em pesquisa tecnológica dos Estados Unidos gerou um grande avanço, permitindo várias implementações (ORTEGA, 2004).

Fundada em 1911, a *International Business Machines Corporation* (IBM) destacou-se durante a Segunda Guerra Mundial fornecendo serviços e produtos para o governo americano. Nascido na Alemanha em 1896, Hans Peter Luhn mudou-se para os Estados Unidos logo após a Primeira Guerra Mundial e assumiu a gerência do Departamento de Pesquisa em Recuperação da Informação na IBM. Suas primeiras publicações na área ocorreram no final da década de 1950 em decorrência da *International Conference on Scientific Information* (ICSI) em Washington (SCHULTZ, 1968).

Em 1958, a ICSI promoveu a divulgação de seus *preprints* em um documento[9] juntamente com seus dados preparados em cartões perfurados, de modo a serem

[9] CITRON, J. L.; HART, L.; OHLMAN, H. *A permutation index to the "Preprints of the International Conference on Scientific Information*". Santa Monica, Cal. System Development Corp., 1958. 140p. (SP-44).

processados por máquinas. “Muitas das primeiras experiências em indexação automática foram realizadas com este material” (SAYÃO, 1985, p. 14). A terminologia *indexação automática* é originalmente usada pelo Luhn (1961) que defendia a necessidade de se usar as próprias palavras e termos de um documento para a sua indexação, assim como sua classificação (SCHULTZ, 1968). A terminologia *indexação automática* é, portanto, concebida juntamente com o conceito de indexação por extração ou indexação derivada.

De acordo com Sayão (1985), o índice KWIC foi implementado na IBM por Hans Peter Luhn e foi acompanhado por outras significativas contribuições para a indexação automática nos anos seguintes conforme apresentado no Quadro 1.

Quadro 1 - Algumas contribuições para a indexação automática no período de 1957 a 1984

Período	Autor(es)	Contribuição(ões) para a indexação automática
1957-59	Luhn (1957, 1958a, 1958b 1959)	Introduziu os temas *auto-resumo* e *auto-indexação*.
1958	Baxendale (1958)	Busca de sentenças significativas, processos sintáticos automáticos e seleção automática de expressões.
1961-63	Swanson (1962, 1963)	Localização no texto de *palavras-pista* que identificassem textos de uma mesma área. *Indexação derivativa*. *Indexação atributiva* com textos curtos. Diminuição de ênfase nas palavras-pista através das estratégias de *localização de sinônimos* e associação de pesos à palavra de acordo com sua frequência.
1960	Maron (1960)	Indexação automática baseada em palavras-pista com enfoque probabilístico de associação estatística entre palavras-pista e cabeçalhos de assunto manualmente assinalados.
1963	Trachtenberg (1963)	Métodos teórico probabilístico de indexação e classificação automática com determinação e valor de associação de palavras-pista a diferentes categorias.
1969	Edmundson (1969)	Extração automática de frases relevantes, ao invés de palavras. Estratégia das *palavras pragmáticas*, tais como *significante* e i*mpossível* que indicavam provável relevância da frase. Hipótese de que determinadas posições dentro do texto continham frases mais relevantes. Técnica de atribuir valores de pesos para as frases, somando-os posteriormente.
1965-71	Salton (1967, 1968, 1971a, 1971b)	O *System for the Mechanical Analysis and Retrieval of Text* (SMART) aceita linguagem natural nos documentos e nas consultas. Separa raízes e sufixos de palavras em inglês. Sinônimo de palavras usando a raiz da palavra. Números de conceitos identificam conteúdos e os substituem pela palavra original. Arranjo hierárquico de conceitos que permite cruzamento e identificação de conceitos mais gerais ou específicos. Associação estatística usando coeficientes de co-ocorrência que permite calcular a similaridade de palavras, raízes de palavras ou conceitos. Análise sintática que compara frases e consultas. Uso de dicionários de expressões predefinidas. *Clusterização* de documentos. Retroalimentação da consulta alterada pelo usuário com base no resultado apresentado pelo

Período	Autor(es)	Contribuição(ões) para a indexação automática
		sistema. Análise de citação bibliográfica como indicador de conteúdo.
1969	Moyne (1969)	Afirmação de que era possível o uso da linguagem natural na recuperação da informação.
1970	Graves e Helander (1970)	Constatação de que somente 40% de descritores controlados assinalados manualmente por indexadores estavam presentes em títulos e resumos do *Petroleum Abstracts*.
1973	Sparck Jones (1973, 1978)	Analisou a influência das características de uma coleção sobre o desempenho de um sistema de recuperação da informação e constatou que: a ponderação estatística influencia sensivelmente nos resultados; técnicas de retroalimentação são altamente positivas; a saída ordenada deve ser considerada; o SMART de Salton apresentou praticamente os mesmo resultados que os demais sistemas da época. Avaliou a insuficiência de técnicas linguísticas mais sofisticadas e a necessidade em melhorar mais as técnicas de elaboração de consultas que as de descrição dos documentos.
1974	Bookstein e Swason (1974)	Método puramente probabilístico para agrupamento (*clustering*) por padrões de ocorrência e distribuição de ocupação.
1975	Field (1975)	Projetou um sistema capaz de gerar automaticamente cabeçalhos de assunto controlados, descritores controlados e classes para documentos a partir de uma indexação livre. Usou o *coeficiente de adesão* para medir o grau de associação entre diferentes elementos de indexação (como cabeçalho de assunto e indexação livre). Versão multilíngue de seu sistema que permitia gerar descritores controlados em uma única língua, o inglês, a partir de qualquer outra língua.
1975-82	Salton, Yang e Yu (1975); Salton (1981, 1982)	Teorias para indexação automática usando álgebra veorial e conjuntos nebulosos (*fuzzy sets*). Resumo de todas as técnicas de indexação automática da época de forma didática.
1976	Yu e Salton (1976)	Métodos para aumentar a atribuição de relevância para termos raros e diminuir para termos frequentes. Técnica de associação de termos frequentes a *expressões-termos*. Uso de um tesauro de termos raros que deveriam ser substituídos por identificadores de conteúdo no lugar do termo individual.
1976	Artandi (1976)	Argumentação de que a linguística e a semiótica podem ser aplicadas de forma a contribuir para a criação de algoritmos mais sofisticados para a indexação automática.
1977	Van Rijsbergen (1971)	Apresentação do conceito de dependência de co-ocorrências de termos e de cálculos de funções não-lineares ponderadas entre termos independentes e dependentes.
1977	Barnes, Constantini e Perschke (1978)	Uso de um sistema, o SLC II, que identificava elementos linguísticos e sintáticos e possuía um módulo de “enriquecimento” do tesauro, que incorporava novos termos.
1977	Van der Meulen e Jansen (1977)	Avaliação da indexação automática (usando o sistema DIRECT, semelhante ao SMART) como comparável à manual, sendo que as suas diferenças residiam principalmente na formulação da consulta.

Período	Autor(es)	Contribuição(ões) para a indexação automática
1978	Dunhan, Pacak, Pratt (1978)	Aplicação de uma linguagem de indexação estruturada interativa o sistema *Systematized Nomenclature of Pathology* com análise morfológica e sintática.
1981-84	Dillon et al. (1981); Dillon (1982); Dillon e Gray (1983); Dillon e McDonald (1983); Dillon e Federhart (1984)	Desenvolvimento de um sistema experimental baseado em um *software* de indexação automática (com uso de tesauro) que analisava textos completos e identificava e substituía termos por suas formas controladas, tais como nomes de autores. Apresentação de um sistema de indexação automática de livros através de seus textos completos (pré-formatados pelo sistema SCRIPT) com a conclusão de que seu sistema funcionaria melhor em áreas de vocabulário altamente específico com filosofia de indexação exaustiva. Descrição do sistema *Fully Automatic Syntatically based Indexing Text* (FASIT) de indexação totalmente automática por sintaxe de qualquer texto com experimentos em textos técnicos (manuais de processamento de dados de bibliotecas). Estratégia estatística para SRIs que identificava *termos tópicos* como segmentos do texto que indicavam sobre o seu assunto.
1981	Borko (1982)	Revisão dos procedimentos tradicionais e apresentação dos avanços da época em indexação automática. Defesa do abandono definitivo da indexação manual e adoção da automática. Apresentação do conceito de medida de qualidade em substituição aos parâmetros de revocação e precisão, como forma de métrica mais apropriada para medir os benefícios que chegam ao usuário.
1982	Aitchinson e Harding (1982)	Comparação de custos entre a indexação manual e a automática com a conclusão de que, para a geração de termos livres, estes apresentavam custos baixos, e que para termos controlados, assim como para classificações automáticas, os custos eram menores que o sistema manual, com altos índices de revocação.
1982	Stokolov (1982)	Descrição de uma técnica baseada em linguagem especialmente formalizada para a representação semântica de textos biológicos (BIOSIS) com a finalidade de diminuir o seu volume de vocabulário. Denominou tal linguagem de “linguagem dos conceitos primitivos”.
1983	Nishida, Takamatsu e Fujita (1984)	Descrição de um método sintático (de forma precisa) e semântico (de forma superficial) para extração semi-automática de informações em textos completos (de língua inglesa e japonesa) com sua posterior normalização para “expressões internas”.
1983	Brozozowski e Masquarade (1983)	Desenvolvimento do sistema MASQUERADE com interface amigável para recuperação de informação em relatórios de geologia e exploração. Combinação de técnicas já dominadas na época como: indexação automática, consulta livre e por lógica booleana, sistema de ponderação para ordenação de saída, retroalimentação de de consultas e possibilidade de aplicação em outras bases de dados.
1984	Bernstein e Willianson (1984)	Apresentação do sistema de recuperação da informação denominado *A Navigator of Natural Language Organized Data* (ANNOD) combinando elementos probabilísticos, linguísticos e empíricos. Ordena parágrafos em textos completos em função de similaridade com consulta formulada em linguagem natural.
1984	Edmundson (1984)	Levantamento sobre modelos matemáticos de textos e revisão de conceitos de análise linguística (grafema, morfologia, sintaxe e semântica) relacionando-os com a Ciência da Informação.

Fonte: Adaptado de Sayão (1985, p. 14-32).

Os primeiros anos da indexação automática apresentaram questões recorrentes. Primeiramente, ela passou por um período de autoafirmação com Luhn (1957, 1958a, 1958b 1959). Moyne (1969) amplia essa autoafirmação assumindo a linguagem natural como viável para a recuperação da informação. Sparck Jones (1973) avaliou a insuficiência das técnicas linguísticas para os SRIs. Artandi (1976) somou a linguística e a semiótica como insumos para a criação de algoritmos mais sofisticados para a indexação automática. Salton, Yang e Yu (1975), Yu e Salton (1976), Salton (1981, 1982) e Borko (1982) finalmente consolidaram os conhecimentos adquiridos até então, tornando-os mais acessíveis com suas publicações e apresentações.

Outra questão recorrente foi sua comparação com a indexação manual. Graves e Helander (1970) relataram ineficiências do uso de vocabulários controlados. Van der Meulen (1977) considerou que as duas seriam comparáveis e apontou a formulação de consulta como o fator principal de diferenças. Borko (1982) defendeu o abandono em definitivo da indexação manual. Aitchinson e Harding (1982) apresentaram que os custos da indexação automática eram equiparáveis aos da indexação manual, e ainda menores com o uso de termos controlados.

A questão mais recorrente foi aquela que pode ser considerada como essencial para a indexação automática: o uso de estratégias e técnicas baseadas em cálculos, estatísticas e probabilidades. Os critérios mais recorrentes, inclusive os contidos nessas inúmeras técnicas e estratégias observadas nessas primeiras décadas, são apresentados na próxima subseção, juntamente com outros mais atuais e encontrados também em português.

No Brasil, os primeiros registros de utilização do pioneiro sistema de indexação automática KWIC ocorreram em 1968, 11 anos após a sua invenção e foram destinados a um processo de automação da informação em Física, sendo este ano considerado o início brasileiro da indexação automática (ZAHER, 1969; SAYÃO, 1985).

Belluzzo *et al.* (1990 citado por Borges, 2009) apresenta que, em 1968, Dr. Derek Austin criou o sistema de indexação *Preserved Context Indexing System* (PRECIS). O PRECIS foi estudado na tese de doutorado de Assumpção (1978). Nos anos seguintes, a indexação automática é abordada como tema principal em dissertações de mestrado e artigos em Biblioteconomia e Ciência da Informação, em congressos brasileiros de computação e informática, assim como no desenvolvimento de sistemas que eram baseados principalmente naqueles já desenvolvidos para outras línguas, como, por exemplo, a

francesa[10](BRAGA, 1982; HALLER, 1982; HALLER, 1983; MARTINS, 1983; VON STAA, 1983; BASTOS, 1984).

A principal necessidade dos sistemas de indexação automática, como apresentado no Quadro 1, na maioria das contribuições iniciais, é o tratamento linguístico de textos em linguagem natural. Os primeiros sistemas foram desenvolvidos para a língua inglesa que, por ser de origem não latina, tende a dificultar sua adaptação para a língua portuguesa. A partir de 1980, a língua portuguesa recebe significativas contribuições para o desenvolvimento de sistemas de indexação automática, principalmente com os trabalhos de Robredo (1980, 1982a, 1982b) e Robredo e Ferreira (1980), que utilizava a lei de Zipf como principal critério para eleição de descritores.

Desde a década de 1950, a diversidade de critérios para eleição de descritores cresceu juntamente com a quantidade de sistemas de indexação automática desenvolvidos em diversas línguas. A seguir, são analisados os critérios mais recorrentes encontrados nos sistemas de indexação automática.

2.4.1 Modelos de recuperação da informação

"Um modelo de recuperação da informação pode ser caracterizado como uma função que determina uma pontuação para a relação de um determinado documento com uma determinada consulta" (BAEZA-YATES; RIBEIRO-NETO, 2011, p. 57). No sentido inverso, é possível considerar que um critério para indexação automática seja uma função que determina a pontuação da relação de um determinado documento com um determinado descritor.

Outra inversão foi apresentada de modo similiar por Hjϕrland (2001), citado e criticado como confuso por Lancaster (2004), ao tratar da decisão de um indexador (por atribuição) ao estabelecer a relação entre documento e descritor (de um vocabulário controlado). Hjϕrland (2001) apresentou que, se um indexador deve decidir qual descritor dentre vários de um vocabulário controlado deve ser atribuído a um documento; no sentido contrário, o indexador pode (e deve) se perguntar: Sob quais descritores pareceria relevante para o usuário encontrar esse documento?. No entanto, Lancaster (2004) considera que Hjϕrland (2001) pode haver confundido duas etapas distintas na indexação: uma anterior, que seria a *análise conceitual*, onde o indexador decide quais assuntos seriam relevantes para o usuário criando assim o vocabulário controlado; e uma etapa posterior, a de

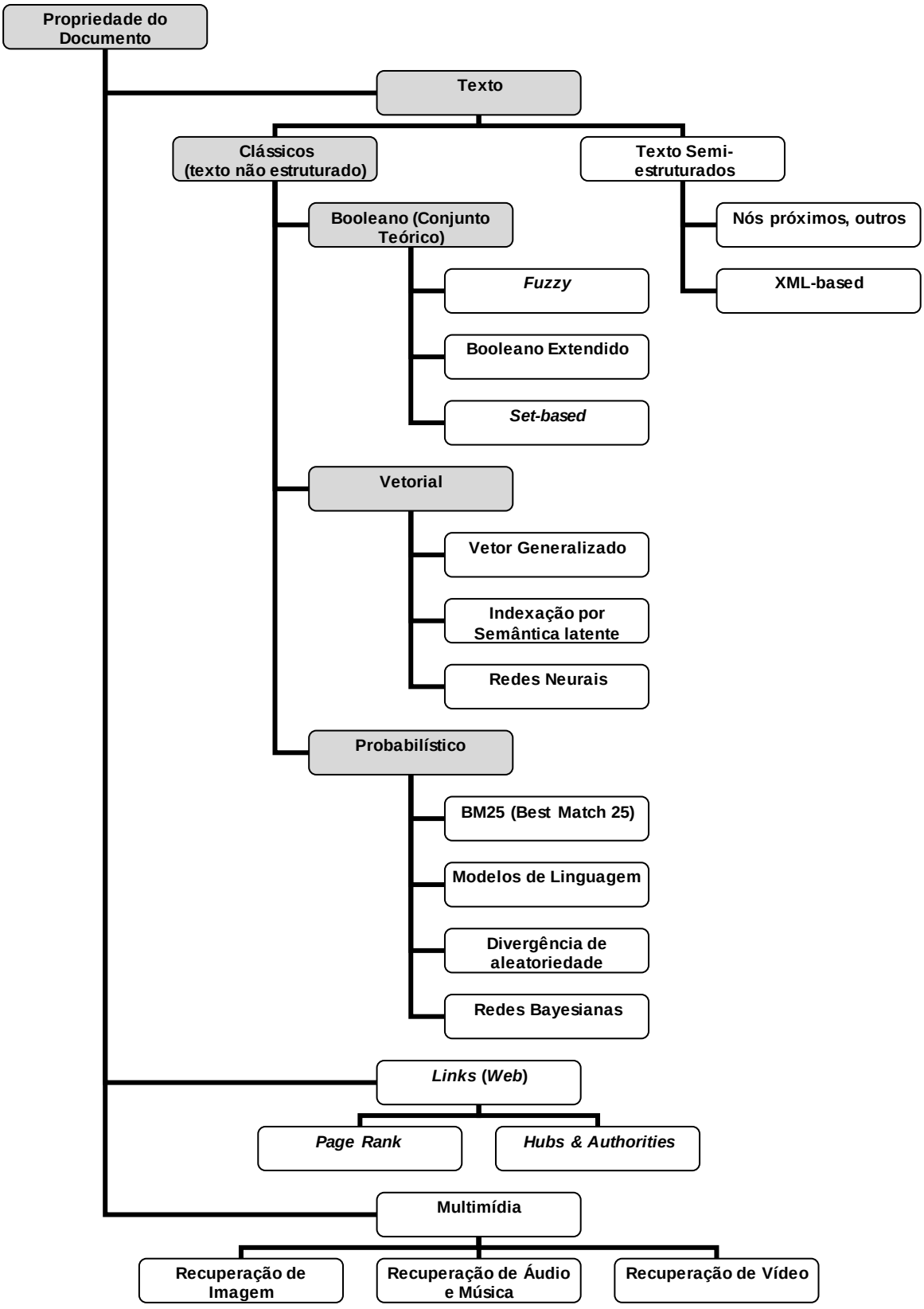
[10] O SPIRIT – *Système Syntaxique et Probabiliste d'Informations Textuelles* foi desenvolvido pelo Centre National de La Recherche Scientifique e foi adaptado para a língua portuguesa por Andreewsky e Ruas em 1982 na Universidade Católica do Rio de Janeiro (SAYÃO, 1985).

tradução, na qual o indexador seleciona os melhores descritores para aquele documento em um vocabulário controlado.

Embora a busca e a indexação possam ocorrer em momentos distintos em um SRI, os modelos de critérios procuram estabelecer um grau de relação entre um documento e um termo (seja este definido pelo usuário na busca, ou definido/atribuído por um indexador durante a inserção do documento no acervo).

Os modelos de recuperação de informação, já exemplificados aqui anteriormente no Quadro 1, foram atualizados e compilados em sua maioria no capítulo 3 do livro de Baeza-Yates e Ribeiro-Neto (2011). A seguir, a Figura 5 resume a taxonomia de alguns modelos de recuperação da informação, desde os clássicos às tendências contemporâneas.

Figura 5 - Taxonomia de modelos de recuperação da informação



Fonte: Adaptado de BAEZA-YATES; RIBEIRO-NETO, 2011, p. 60.

Os modelos de recuperação da informação clássicos destacados na Figura 5 fazem referências a muitos autores que fizeram contribuições iniciais para a indexação automática apresentadas no Quadro 1. Alguns destes autores e suas contribuições podem ser classificados de acordo com a taxonomia proposta por Baeza-Yates e Ribeiro-Neto (2011). No Quadro 2, a seguir, são apresentados todos os autores que apresentaram contribuições iniciais para a indexação automática e também foram citados para cada uma das classificações clássicas iniciais dos modelos de RI.

Quadro 2 - Classificação das principais contribuições iniciais para indexação automática de acordo com a taxonomia de modelos clássicos iniciais de RI.

Autor	Descrição da contribuição	Classificação no modelo de RI
Luhn (1957)	O peso de um descritor em um documento é simplesmente proporcional à frequência do termo neste documento.	Clássico. Conceitos básicos.
Salton, Yang e Wong (1975)	Uso da combinação da frequência do termo com a frequência inversa do documento (TF-IDF).	Clássico. Conceitos básicos e vetorial.
Salton e McGill (1983)	Livro que compila vários modelos.	Clássicos: booleano, vetorial e probabilístico.
Van Rijsbergen (1979)	Livro que compila vários modelos.	Clássicos: booleano, vetorial e probabilístico.
Bookstein (1978, 1985)	Discussão sobre os problemas do uso conjunto do nodelo booleano com atribuição de pesos. As implicações da estrutura booleana para o modelo probabilístico.	Clássicos: booleano e probabilístico.
Salton e Lesk (1968)	Popularização do modelo de vetores através da publicação dos resultados obtidos com seu sistema de recuperação da informação SMART. Verificou a eficiência do uso do valor inverso da frequência do documento. Uso de pesos por termos simples**Erro! Indicador não definido.**. Aprofundou nos estudos dos pesos dos termos no *ranking* final.	Clássicos: vetorial.
Sparck Jones (1972, 1973)	Introdução do uso da frequência inversa do documento.	Clássicos: vetorial.
Maron e Kuhns (1960)	Discussão do uso da relevância e indexação probabilística.	Clássico: probabilístico.
Robertson e Sparck Jones (1976)	Conceito de resposta ideal para uma determinada busca realizada por um processo interativo com o usuário, para o qual são apresentadas possíveis respostas com base em probabilidades.	Clássico: probabilístico.
Sparck Jones (1979)	Estudos experimentais com o modelo probabilístico com o uso do *feedback* dos usuários para estimar as probabilidades inciais.	Clássico: probabilístico.

Fonte: Adaptado de SAYÃO, 1985 e BAEZA-YATES; RIBEIRO-NETO, 2011.

Os conceitos básicos para os modelos de recuperação da informação surgem com Luhn (1957) assumindo a frequência do termo (*term frequency - TF*) como critério para atribuição de pesos em um documento.

> **Definição**: *Frequência do Termo.* O valor, ou peso, de um termo k_i que ocorre em um documento d_j é simplesmente proporcional à frequência do termo $f_{i,j}$. Isto é, quanto mais o termo k_i ocorre em um texto do documento d_j, mais alto é seu peso por frequência de termo $TF_{i,j}$ (LUHN, 1957, tradução do autor).

Sparck Jones (1972) apresentou o conceito de *especificidade do termo* que foi denominado como frequência inversa do documento e se baseou nas noções de exaustividade e especificidade dos termos.

> **Definição**: *Exaustividade e Especificidade*. Exaustividade é uma propriedade de descrição do documento, especificidade é uma propriedade dos termos de indexação. A exaustividade da descrição do documento é interpretada como a sua cobertura para os principais tópicos do documento. A especificidade de um termo de indexação é interpretada como o quão bem o termo descreve um tópico do documento (BAEZA-YATES; RIBEIRO-NETO, 2011, p. 70, tradução livre).

O nível de exaustividade adotado é considerado como a principal decisão da política de indexação e vai determinar estatisticamente a quantidade de termos de indexação usada em média para cada documento. Uma indexação exaustiva elege/atribui termos de indexação para todos os assuntos de um documento, por outro lado, a indexação seletiva elege/atribui uma quantidade limitada de termos de modo a representar somente os assuntos principais de um documento (LANCASTER, 2004). A *exaustividade ótima* considera que o número de termos de indexação deva ser otimizado de modo que a probabilidade de relevância do documento recuperado seja maximizada (BAEZA-YATES; RIBEIRO-NETO, 2011). Ou seja, para uma provável consulta, a quantidade de termos de indexação deve possibilitar uma máxima recuperação de documentos considerados relevantes por um usuário.

A especificidade é a propriedade semântica do termo que depende do seu significado. Por exemplo, *moradia* é menos específico que *casa* ou *apartamento*. A especificidade pode ser ainda definida através da estatística em substituição da propriedade semântica do termo de indexação. Ou seja, o valor de especificidade de um termo pode ser calculado através do inverso da quantidade de documentos nos quais ele ocorre. Se um termo ocorre em todos os documentos, sua especificidade é baixa.

Antes de serem apresentadas as equações para o cálculo dos pesos da frequência e da especificidade dos termos, é importante frisar o comportamento da frequência de termos encontrado por Zipf (1932), que caracterizou a ordenação descrescente das frequências dos termos de um documento como uma função exponencial, exemplificada no Gráfico 9. Logo, para obter um peso com variação linear em função da

frequência, pode ser usada uma escala logarítmica da frequência de cada termo. Esse mesmo recurso matemático pode ser usado para o cálculo dos pesos relacionados à especificidade.

Baeza-Yates e Ribeiro-Neto (2011) apresentam três recomendações[11] de equações para o cálculo de pesos para termos em um documento. No Quadro 3, as três equações utilizam as seguintes expressões:

- $f_{i,j}$ → frequência do termo *i* no documento *j* (TF);
- N/n_i → número total de documentos dividido pelo número de documentos nos quais ocorre o termo *i* ao menos uma vez (especificidade ou IDF).

Quadro 3 - Recomendações de equações para o cálculo de pesos de termos

Peso do termo em um documento
$f_{i,j} \cdot \log N/n_i$
$1 + \log f_{i,j}$
$(1 + \log f_{i,j}) \cdot \log N/n_i$

Fonte: Adaptado de BAEZA-YATES; RIBEIRO-NETO, 2011, p. 74.

Os modelos clássicos em recuperação da informação foram delineados inicialmente para textos não estruturados, como apresentado na Figura 5, sendo que os modelos que os compõem são os booleanos, vetoriais e probabilísticos.

O modelo booleano, que considera a teoria de conjuntos e a álgebra booleana, possui como principal vantagem sua simplicidade ao usar pesos para termos de indexação de forma binária. Sua principal desvantagem é a ausência de uma pontuação que permita uma ordenação (*ranking*) de acordo com a relevância do termo. Em sua essência binária, o modelo booleano considera um termo somente como relevante e não relevante.

O modelo vetorial que foi delineado principalmente por Sparck Jones (1972) reconhece as limitações do modelo booleano e apresenta a possibilidade de *ranking* dos termos. Os pesos utilizados no modelo vetorial são basicamente calculados a partir da frequência do termo e do inverso da frequência dos documentos, como apresentado no Quadro 3. Para Baeza-Yates e Ribeiro-Neto (2011), o modelo vetorial é o modelo mais popular e é aplicado em diversas coleções de documentos, sendo o modelo mais utilizado para a avaliação de novos modelos de recuperação de informação.

O modelo probabilístico assume que para uma determinada busca existe um conjunto ideal de documentos que são relevantes. A questão central do modelo

[11] A primeira recomendação utilizada nesta pesquisa e é apresentada no capítulo sobre a metodologia.

probabilístico é o desconhecimento das propriedades que caracterizam esse conjunto ideal, sendo necessário atribuir uma probabilidade do que seria relevante. Essa probabilidade pode ser refinada com interações com o usuário em direção ao que ele considera como resposta ideal. Tal necessidade caracteriza uma desvantagem desse modelo, que precisaria de informações que estariam fora do próprio sistema.

Outros tipos de documentos ganharam importância a partir da década de 1990: *links* web e documentos multimídia. Os *links* são propriedades que caracterizam os denominados hipertextos, cujo principal conceito reside na noção não linear de sequência. Os documentos, denominados *Web Pages*, não se encontram em um único repositório e necessitam ser mapeadas em um processo denominado *crawling*, no qual a cada *link* encontrado em um documento mapeia-se outro documento que contém outros *links*. Esse processo recursivo pode fazer referência a uma quantidade de documentos muito grande ou até mesmo incalculável.

Os documentos multimídia, com elementos como imagem, áudio e vídeo, necessitam de modelos muito distintos dos empregados para os textuais, principalmente na formulação das consultas. A principal questão nesses modelos está na delimitação de unidades semânticas que devem considerar elementos tanto espaciais como temporais. Alguns experimentos, como em funções encontradas no *Youtube*, possibilitam a busca de texto na fala de interlocutores de um vídeo, por exemplo.

Dentre os demais modelos de recuperação da informação apresentados na Figura 5 é possível ainda destacar, em relação a documentos textuais, aqueles que tratam de textos semi-estruturados. As estruturas nesses modelos consideram, por exemplo, partes do texto como seções, subseções, parágrafos, figuras, títulos, nome de autores, rótulos de figuras etc. Pode ser considerada também como parte da estrutura de um texto a posição linear de um termo entre o início e fim do texto completo.

Como apresentado no início desse subitem, podemos considerar que as estratégias e técnicas apresentadas nos modelos de recuperação de informação nos diferentes modelos podem ser associadas com as técnicas de indexação automática. Para esta pesquisa, o recorte dentre os modelos foi para aqueles relacionados a documentos textuais. O modelo vetorial foi considerado o mais indicado dentre os modelos clássicos Os critérios para indexação automática que fazem referência a estruturas e as posições do texto fazem parte do objetivo principal desta pesquisa e são tratados no subitem a seguir.

2.4.2 Critérios de posição e estrutura para indexação automática

Borges (2010), em sua dissertação de mestrado, faz uma listagem e caracteriza os critérios mais recorrentes nos artigos técnico-científicos, dissertações, teses, livros e

outros, publicados eletronicamente até 2008. Sua pesquisa, para a língua inglesa e portuguesa, considerou 28 fontes de informação, dentre base de dados, periódicos e anuários; elegendo um total de 103 documentos. Ele elaborou uma listagem que enumera 16 tipos de critérios de indexação, dentre eles estão o de frequência de termo e o do inverso da frequência dos documentos. Outros dois são relacionados a posições e as estruturas do texto: um considera certas partes (como títulos e resumos) como mais relevantes, outro considera *tópicos frasais* que considera posições iniciais e finais como mais relevantes.

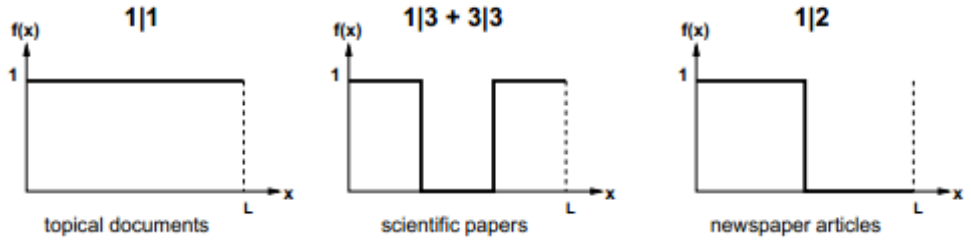
Os critérios que fazem referência a posições e as estruturas do texto têm o propósito de aumentar a eficiência do processo de indexação. Esse aumento de eficiência é possibilitado uma vez que algumas partes do texto podem ser analisadas prioritariamente com maior probabilidade de oferecer informações para a indexação. Por exemplo, um termo que ocorre no título ou no resumo tem maiores chances de ser um descritor relevante, enquanto que as demais poderiam ser ignoradas pelo indexador de modo a economizar tempo.

Feltrim, Aluísio e Nunes (2000) apresentam que, para textos científicos, a introdução e a conclusão concentram os assuntos gerais do texto, sendo que há uma evolução *geral-específico-geral* na qual o leitor é conduzido de assuntos gerais no início, em seguida são tratados os assuntos específicos e, no final, há uma volta para os assuntos gerais. Kobashi (1994) caracterizou que um texto científico possui o quê ela denomina de superestruturas textuais: no início é apresentada e delimitada a questão a ser discutida no texto, em seguida são apresentados os dados que procuram fundamentar um ponto de vista e, por fim, na conclusão ocorrem os comentários finais e soluções encontradas.

Baxandele (1958), de um modo similar, caracterizou que a primeira frase de um parágrafo possui 85% de seus termos relevantes, enquanto que a última frase possui 7% desses termos. Shah *et al.* (2003) analisaram textos completos de artigos científicos segmentando-os em cinco partes, por ordem de posição no texto: resumo, introdução, métodos, resultados e discussão, e concluíram que o resumo possui a maior densidade de termos relevantes. A introdução e a discussão seriam as partes textuais mais relevantes, enquanto que a metodologia teria a menor densidade de termos relevantes para a área biomédica.

Galeas, Kretschmer e Freisleben (2009) analisaram a distribuição linear de termos ao longo de um texto através de redução para séries matemáticas de Fourier. Os coeficientes de tais séries foram usados por esses autores para determinar o grau de similaridade entre a distribuição esperada dos termos de busca e a apresentada nos documentos, além de realizarem experimentos com três tipos de distribuição linear apresentados a seguir.

Figura 6 - Caracterização de distribuição de termos relevantes (f) por posição no texto (x)



Fonte: GALEAS; KRETSCHMER; FREISLEBEN, 2009, p. 4.

Os textos científicos teriam uma tendência a concentrar termos relevantes em suas extremidades. Artigos jornalísticos possuem como característica a apresentação inicial de todos os dados relevantes para a notícia. Outros textos podem ser generalizados com uma distribuição homogênea.

Esta pesquisa tem como objetivo principal a análise da distribuição de termos relevantes ao longo de textos científicos, seja pela posição relativa ao tamanho total do documento, como apresentado por Galeas, Kretschmer e Freisleben (2009), seja por partes estruturais, como apresentado por Shah *et al.* (2003). No entanto, os termos utilizados aqui são SNs extraídos do próprio texto assim como realizado por Souza (2005). A seguir descreve-se a metodologia empregada nessa pesquisa.

3 Metodologia

A fundamentação teórica apresentada anteriormente teve o objetivo de tornar possível a compreensão dos principais conceitos usados para o desenvolvimento desta pesquisa. Outros conceitos são especificados ainda neste capítulo e foram decorrentes do tratamento dos dados analisados.

Este capítulo descreve inicialmente um pré-teste realizado que foi fundamental para delinear a metodologia final aplicada. Em seguida, são apresentados em detalhes: o método empírico utilizado desde a seleção, obtenção e tratamento do *corpus*, assim como o processo para a extração dos SNs, a metodologia empregada para a determinação dos descritores candidatos, a aplicação dos questionários aos entrevistados e, por fim, o processo de distribuição dos valores de relevância dos descritores por suas respectivas posições nas teses. Tal distribuição é analisada em detalhes no capítulo seguinte.

1.5 Pré-teste

Souza (2005) utilizou em sua pesquisa 60 artigos na sua metodologia consolidada, sendo que aqui foram escolhidos 10 desses artigos[12] para a realização do pré-teste. O motivo da escolha dos mesmos artigos foi pelo fato de Souza (2005) já ter apresentado a avaliação da relevância dos SNs como descritores, assim como outros dados que foram usados para avaliar a metodologia aqui desejada.

Foram escolhidos somente 10 artigos aqui, pois todo o pré-teste foi realizado manualmente, ou seja, sem o uso de *scripts* de computador especialmente desenvolvidos para automatizar os processos. Tais *scripts* foram desenvolvidos após o pré-teste, como será detalhado posteriormente.

Os 10 artigos foram facilmente transformados em texto puro. Foram retirados os campos não considerados textuais e o restante foi exportado como documentos eletrônicos identificados como *an.txt* (onde n era o número do documento e variou de 1 a 10).

A ferramenta Ogma foi utilizada nas etapas de: etiquetagem, extração dos sintagmas nominais e classificação da sua estrutura.

Para cada SN, calculou-se o total de suas ocorrências em cada documento, o total de documentos no *corpus* que o SN ocorria e seu valor associado. Com esses três fatores foi possível verificar a aplicação da metodologia empregada por Souza (2005) e seus respectivos valores de relevância atribuídos a cada SN como descritores.

[12] Os artigos foram retirados da revista eletrônica DataGramaZero, disponível em: http://www.dgz.org.br/.

Foram encontrados 9.874 SNs nos 10 artigos. A ferramenta Ogma levou cerca de 10 minutos[13] para extrair todos os SNs. Conforme é detalhado na Tabela 2, foi extraída uma quantidade de 97% de SNs utilizando-se a ferramenta Ogma em relação à extração realizada por Souza (2005).

Tabela 2 - Avaliação da extração de sintagmas nominais pelo Ogma

Artigos	Souza (2005)	Ogma	%
1	1.673	1.404	**84%**
2	842	886	**105%**
3	783	713	**91%**
4	801	999	**125%**
5	1.478	1.092	**74%**
6	984	809	**82%**
7	638	643	**101%**
8	779	924	**119%**
9	1.104	982	**89%**
10	1.146	1.422	**124%**
Total	10.228	9.874	**97%**

Fonte: Elaborado pelo autor.

Para cada ocorrência de um SN eleito como descritor em cada artigo, foi atribuída à sua posição um valor correspondente à sua relevância. Para realizar essa atribuição, foi considerado que: cada SN ficaria na mesma sequência original do texto (retirando-se tudo que não seria um SN); cada posição seria numerada proporcionalmente (em %) à quantidade total de SNs do texto; o valor da relevância atribuída em cada ocorrência do SN seria dividido pelo total de ocorrências naquele texto.

Na

Tabela 3 é dado um exemplo dessa distribuição dos valores de relevância em um trecho de um dos artigos.

[13] Foi utilizado um computador com processador Core 2 Duo, 2,0GHz.

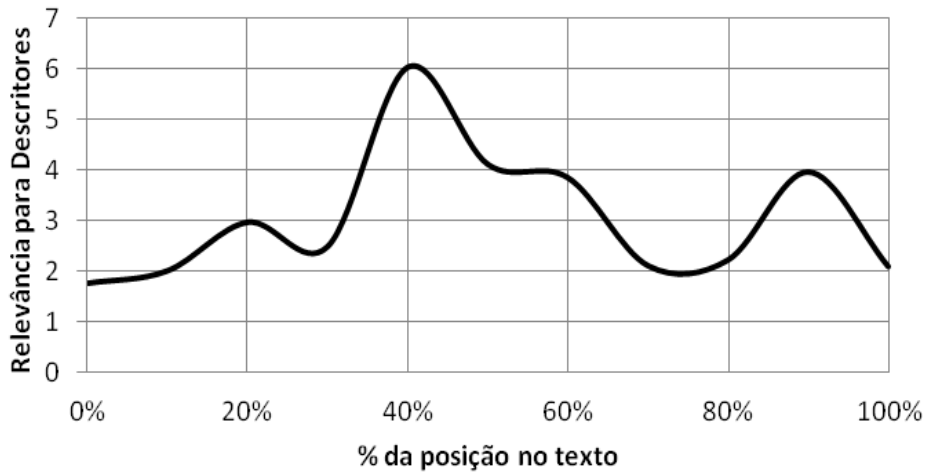
Tabela 3 - Exemplo de distribuição de valores de relevância em um artigo

% Posição no texto	Sintagma Nominal	Valor
...	...	...
48%	2000	0
48%	o auto-arquivamento	0,25
49%	as soluções	0
49%	justamente este conceito	0
49%	auto-arquivamento	0,5
49%	conceituação segundo descrito no site	0
...	...	...

Fonte: Elaborado pelo autor.

Em seguida, todos os valores nos 10 artigos foram consolidados em uma única distribuição, representando assim a distribuição total dos valores de relevância de descritores no *corpus*. Para possibilitar uma análise, as posições foram divididas em dez partes e os valores encontrados podem ser vistos no Gráfico 1 a seguir.

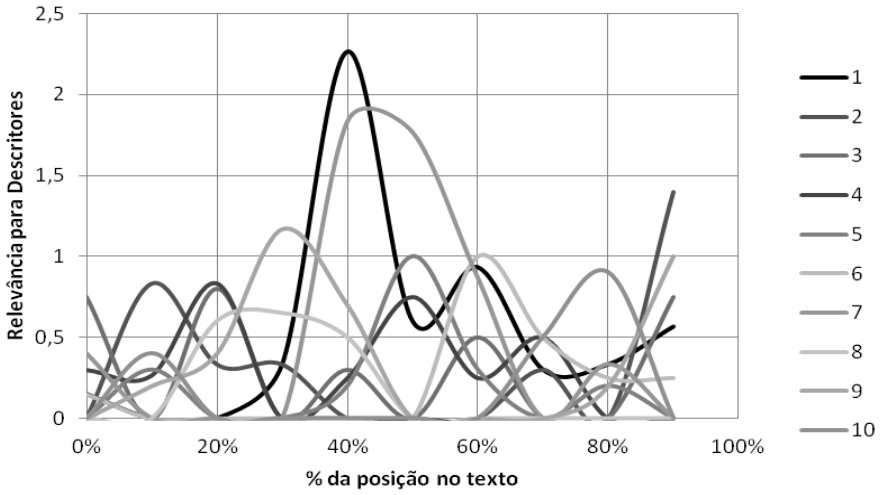
Gráfico 1 - Relevância para descritores por posição em um *corpus* de pré-teste



Fonte: Elaborado pelo autor.

Podem-se observar, nos dados obtidos, que a dimensão textual dos artigos poderia dificultar a análise da distribuição de valores de relevância, uma vez que os artigos apresentaram distribuições bem diferentes entre si, como pode ser visto no Gráfico 2.

Gráfico 2 - Relevância para descritores por posição por artigo no pré-teste



Fonte: Elaborado pelo autor.

Em virtude da discrepância tão acentuada da distribuição nos diferentes artigos, foi postulada a hipótese de que, se os textos fossem maiores e com mais descritores, haveria mais homogeneidade de distribuição. De fato, com a adoção de teses de doutorado isso ocorreu, como é apresentado adiante.

Com o pré-teste, foi possível concluir positivamente que:

1. A ferramenta Ogma apresentou uma quantidade de SNs extraídos bem próxima à obtida por Souza (2005);
2. Seria viável realizar o cálculo de valores de relevância dos SNs como descritores tal como em Souza (2005);
3. Seria viável a análise da distribuição dos valores de relevância em um *corpus*.

De fato, todas estas conclusões se confirmaram no restante da pesquisa, como é apresentado a seguir.

3.2 Seleção, obtenção e tratamento do c*orpus*

Em virtude da necessidade de um *corpus* com textos mais longos, buscou-se por teses de doutorado, como textos mais longos e acessíveis digitalmente. O Portal de Periódicos da CAPES possui 64 bases de teses e dissertações, sendo que 58 delas são brasileiras. Dessas bases, foi escolhida a Biblioteca Digital de Teses e Dissertações (BDTD) da UFMG, uma vez que seria necessário entrevistar seus autores, e a proximidade poderia aumentar a adesão dos entrevistados.

Para uma tese, que “possui aproximadamente entre cem e quatrocentas páginas relacionadas a uma área de estudos” (ECO, 2007, p. 27), acredita-se, como dito anteriormente, que essa ordem de grandeza textual pode favorecer ao estudo da distribuição dos SNs como descritores. Essa hipótese é baseada nos seguintes aspectos: as repetições de um mesmo SN tendem a aumentar conforme o crescimento da quantidade de palavras em um texto que trata de uma mesma área; com uma quantidade maior de repetições de um mesmo sintagma, pode-se avaliar com mais detalhes suas variações da distribuição ao longo do texto.

A escolha aqui de teses como elementos de pesquisa implica em maior *custo computacional* de processamento da extração dos SNs, em comparação a artigos, uma vez que estes últimos, geralmente, possuem um tamanho da ordem de dez vezes menor. No entanto, com o desempenho dos recursos computacionais atuais em relação aos mais antigos[14] usados em outras pesquisas, que se basearam em artigos, o processamento de teses mostrou-se viável (cerca de 16 horas para 98 teses), como pode ser visto adiante na análise de resultados.

Para Babbie (1999, p. 113), “o principal motivo em se fazer amostragem em uma pesquisa é custo e tempo”. A BDTD/UFMG possui, atualmente, aproximadamente 2.000 teses. Avaliar toda essa população levaria um tempo contraproducente para o escopo desta dissertação. Como é de conhecimento geral, o emprego da estatística[15] está sujeito a níveis de confiança e precisão, os quais, com uma seleção cuidadosa de amostras, podem tornar a pesquisa viável sem a necessidade de se avaliar todos seus itens. A inferência estatística usada aqui seria então sobre o processo de generalização dos resultados dessas amostras para toda a população.

Inicialmente foram levantadas todas as quantidades de teses na BDTD/UFMG, encontrando-se 1.921 referências pertencentes a 54 programas de pós-graduação (os outros 13 programas só apresentaram dissertações de mestrado). A listagem completa dos programas e suas respectivas quantidades de teses encontram-se no APÊNDICE A.

Para atingir um maior grau de representatividade e um menor erro amostral, foi utilizada uma amostragem estratificada, ou seja, os elementos de pesquisa (as teses) foram agrupados de modo a representar sua heterogeneidade (BABBIE, 1999), sendo separados por programas de pós-graduação. Objetivou-se também representar as oito áreas de conhecimento nas quais esses programas estão inseridos: Ciências Agrárias, Ciências

---

[14] Souza (2005) utilizou um computador com processador AMD Athlon XP 2600+ com 256MB de memória RAM. O utilizado aqui possui processador Intel Core i5-2430M 2,4GHz com 4GB de RAM.

[15] Define-se aqui estatística como: “Ciencia que se ocupa del estudio de fenómenos de tipo genérico, normalmente complejos y enmarcados en un universo variable, mediante el empleo de modelos de reducción de la información y de análisis de validación de los resultados en términos de representatividad." (BENITEZ, G. S.; ARRONDO, V. M. Sobre la definición de estadística. *DataGramaZero - Revista de Ciência da Informação*. V. 6, N.4. 2005. Disponível em: <http://www.dgz.org.br/ago05/Art_02.htm>. Acesso em: 07/04/2012).

Biológicas, Ciências da Saúde, Ciências Exatas e da Terra, Ciências Humanas, Ciências Sociais Aplicadas, Engenharias e, por fim, Linguística, Letras e Artes. O método de eleição dos programas consistiu em ordenar decrescentemente por quantidade de teses dos 54 distintos programas e eleger aqueles que possuíssem mais teses dentro da sua área de conhecimento. Os oito grupos eleitos foram denominados pelas letras de A até H conforme aTabela 4:

Tabela 4 - Eleição dos programas de pós-graduação para amostragem

Grupo Eleito	Ordem	Programa de Pós-Graduação	Qtd. Teses	Área de Conhecimento
A	1º	Pós-Graduação em Educação: Conhecimento e Inclusão Social	214	Ciências Humanas
B	2º	Pós-Graduação em Ciência Animal	128	Ciências Agrárias
C	3º	Pós-Graduação em Letras: Estudos Literários	105	Linguística, Letras e Artes
D	4º	Pós-Graduação em Engenharia Metalúrgica e de Minas	91	Engenharias
	5º	Pós-Graduação em Estudos Linguísticos	90	Linguística, Letras e Artes
	6º	Pós-Graduação em Engenharia Elétrica	88	Engenharias
E	7º	Pós-Graduação em Química	76	Ciências Exatas e da Terra
	8º	Pós-Graduação em Física	75	Ciências Exatas e da Terra
	9º	Pós-Graduação em Ciência da Computação	72	Ciências Exatas e da Terra
F	10º	Pós-Graduação em Bioquímica e Imunologia	61	Ciências Biológicas
G	11º	Pós-Graduação em Ciência da Informação	58	Ciências Sociais Aplicadas
H	12º	Pós-Graduação em Medicina (Pediatria)	56	Ciências da Saúde

Fonte: Adaptado de BDTD/UFMG, 2012.

Uma vez eleitos os grupos, a determinação para o *tamanho da amostra* levou em conta o objetivo da entrevista, que foi avaliar a relevância de SNs candidatos como descritores de uma tese. Foi adaptada a equação usada por Levine, Berenson, Stephan (2000) e usada para determinar o tamanho da amostra para uma proporção:

Equação 1 - Tamanho da amostra para uma proporção

$$n = Z^2 p(1-p)/e^2$$

n -> ta
tamanho da amostra;
Z -> relativo ao nível de confiança desejado;
p -> relativa à verdadeira proporção de aceite das respostas das entrevistas;
e -> relativo ao volume de erro de amostragem que se está disposto a aceitar.

Fonte: Adaptado de LEVINE; BERENSON; STEPHAN, 2000, p. 301.

A seguir, temos o resultado da determinação do tamanho da amostra de cada grupo em função dos três parâmetros definidos por Levine, Berenson, Stephan (2000):

- nível de confiança (relativo a Z) = 90%;
- verdadeira proporção (relativo a *p*) = proporção para todas as teses;
- nível de erro de amostragem (relativo a *e*) = 10%.

Tabela 5 - Determinação do tamanho da amostra de cada grupo

Grupo	Qtd. da subpopulação	Proporção na População	Tamanho da Amostra	% Amostral do Grupo	% Amostral da População
A	214	11,1%	**24**	11,2%	1,2%
B	128	6,7%	**16**	12,5%	0,8%
C	105	5,5%	**13**	12,4%	0,7%
D	91	4,7%	**12**	13,2%	0,6%
E	76	4,0%	**10**	13,2%	0,5%
F	61	3,2%	**8**	13,1%	0,4%
G	58	3,0%	**8**	13,8%	0,4%
H	56	2,9%	**7**	12,5%	0,4%
Total	**789**	**41,1%**	**98**	**12,4%**	**5,1%**

Fonte: Elaborado pelo autor.

Para cada programa de pós-graduação, foram selecionadas teses que foram disponibilizadas na BDTD/UFMG mais recentemente. O *recorte temporal* aqui faz parte de qualquer processo de entrevistas (BABBIE, 1999). Foi utilizada uma amostragem sistemática iniciando-se da publicação mais recente em direção à mais antiga. Seguiram-se duas hipóteses para que esse recorte favorecesse à pesquisa: os autores convidados poderiam responder ao questionário com base em uma memória mais recente de quando da elaboração das suas respectivas teses; e os convidados poderiam ser localizados mais facilmente, uma vez que haveria mais chances de seus dados na Plataforma Lattes estarem atualizados[16].

Uma vez então definido cada grupo de amostragem com um tamanho finito, representativo estatisticamente, e ainda de forma sistemática na sua homogeneidade possibilitada pelo recorte temporal, foi considerado aqui que esses grupos comporiam um *corpus* limitado ao seu tempo. Para DUBOIS (1973, p. 47):

[16] Por outro lado, esse recorte sistemático temporal em cada programa de pós-graduação pode ser influenciado por vieses de temas de pesquisas, considerando-se aqui que, na comunidade científica, há períodos em que as publicações costumam versar mais sobre um ou mais determinados temas. Nas técnicas estatísticas, trabalha-se comumente com amostragens probabilísticas, geralmente atribuídas à eleição de elementos de forma aleatória, ao contrário daqui, minimizando que os resultados tenham um viés (consciente ou inconsciente). A reflexão de tal viés, sendo aqui de forma consciente, para os métodos de representação da informação aqui aplicados poderia causar uma menor representatividade desses temas modais, uma vez que um dos critérios de eleição dos descritores consiste na raridade que os mesmos ocorrem em cada grupo do *corpus*. Ou seja, se a amostragem for feita, por exemplo, em um período no qual muitas teses falam de um mesmo assunto, refletindo aí o uso de um sintagma nominal relativo a ele, tal termo terá uma menor chance de ser eleito como descritor. Isso implica que o recorte temporal sistemático aqui adotado pode acabar por mascarar descritores que caracterizariam uma identidade de uma época analisada que fosse diferente das demais.

> Infere-se a língua de *corpus* por generalização. A determinação de um *corpus* é feita segundo determinado número de critérios que devem garantir seu caráter representativo e a homogeneidade dos enunciados, afastando *a priori* as variações de situação. [...] Trata-se então de descrever os elementos de uma língua pela sua aptidão (possibilidade ou impossibilidade) para se associar entre si a fim de chegar à descrição total de um estado de língua em sincronia (DUBOIS, 1973, p. 47).

Através do nome do principal autor, foi realizada uma busca do seu currículo na Plataforma Lattes[17]. Nesta plataforma, é permitido o envio de um *e-mail* para o autor[18]. Para esta primeira mensagem[19], com a solicitação de sua participação na pesquisa, foi tomado o cuidado para que o conteúdo possuísse um formato que se distanciasse ao máximo possível de uma mensagem automática[20].

Nesta primeira mensagem enviada aos autores das teses, foi solicitada a participação na pesquisa através da resposta a um questionário. A confirmação da participação do autor previamente ao processamento dos textos possibilita uma maior eficiência na pesquisa, uma vez que tal processamento demandou significativos recursos computacionais e humanos.

Uma vez recebida a confirmação de que o autor da tese concordara em participar da pesquisa, sua tese foi obtida a partir da BDTD/UFMG no formato PDF[21].

Para cada programa de pós-graduação foram gerados vários arquivos, assim como planilhas eletrônicas. Todos estão disponibilizados em mídia digital anexa a esta pesquisa. Nos procedimentos a seguir são detalhadas as sintaxes dos nomes de tais arquivos digitais de modo a possibilitar suas referências.

Os textos foram convertidos do seu formato PDF para TXT (texto simples) adotando-se os seguintes procedimentos:

1. Foram descartadas as partes pré-textuais, tais como capa, dedicatórias, agradecimentos, resumos, listas de ilustrações, lista de tabelas, listas de abreviaturas, sumários, e ainda as partes pós-textuais, como referências bibliográficas, apêndices e anexos;
2. Foram descartadas todas as informações cujo formato digital não fosse o textual, tais como gráficos, imagens e figuras[22];
3. Foram eliminados espaços em branco consecutivos;

---

[17] Disponível em http://buscatextual.cnpq.br/buscatextual/busca.do?metodo=apresentar

[18] Para evitar a prática de *spam*, a Plataforma Lattes permite o envio de mensagem para o autor mediante uma confirmação enviada para o email do solicitante do envio da mensagem. Para evitar o uso automatizado desse recurso por *softwares*, durante o processo é solicitada a digitação de caracteres presentes em uma imagem.

[19] Um exemplo da mensagem enviada encontra-se no APÊNDICE B.

[20] Evitou-se usar o termo introdutório “Caro(a)”, que revela uma pré-concepção do texto desvinculada do gênero do destinatário. Foi utilizado na introdução do texto somente o primeiro nome do autor, em um tom menos formal, o que seria mais propício para uma comunicação entre colegas de pesquisa de uma mesma instituição. Esse procedimento visou dar mais credibilidade à mensagem eletrônica, diminuindo a possibilidade de ser classificada como *spam*, por exemplo, e aumentando a adesão dos autores em participar da pesquisa.

[21] O PDF é um padrão aberto de arquivo (*Portable Document Format*) desenvolvido pela *Adobe Systems*.

[22] Os textos contidos em formatos digitais não textuais, tais como em imagens ou figuras, também foram descartados.

4. Uma vez que na conversão do formato PDF para o TXT não houve distinção entre a mudança de linha e mudança de parágrafo, sendo convertidos todos como mudanças de parágrafo, optou-se por eliminar todos esses, tornando o texto uma sequência de frases sem parágrafos[23];
5. Foram inseridos demarcadores[24] logo após a introdução e antes da parte final, como conclusão e/ou considerações finais.

Todos os procedimentos descritos neste item foram realizados manualmente. Ao final deles, cada texto pré-processado foi nomeado usando-se a seguinte sintaxe *ann.tx"*: [letra do grupo] + [número sequencial com dois dígitos] + *.txt* (extensão de arquivo do tipo texto). Exemplos: a01.txt até a24.txt; b01.txt até b16.txt; c01.txt até c13.txt; etc.

3.3 Extração dos sintagmas nominais

Para cada texto, foram obtidos seus SNs e apresentados, um em cada linha, em um novo texto. Considerou-se aqui cada SN máximo, desconsiderando-se os SNs aninhados. Essa escolha deve-se ao fato da ferramenta Ogma fornecer a listagem sequencial de sintagmas somente nesse formato.

A posição de cada SN foi definida somente em relação aos outros SNs. Ou seja, embora possa haver termos que não sejam SNs entre dois destes, a posição de um em relação ao outro foi considerada aqui como sendo consecutiva.

A ferramenta Ogma 0.10[25] e o *software* Microsoft Office Word 2007 foram utilizados para a extração dos SNs através dos seguintes procedimentos:

1. Etiquetagem: a partir de cada texto pré-processado com o nome no formato *ann.txt* foi gerado um novo arquivo. Esse arquivo é utilizado como uma etapa intermediária para a extração dos SNs. Nela é realizada a etiquetagem do texto no modelo ED-CER (MAIA, 2008). Usou-se a seguinte sintaxe de comando para este procedimento:
    - ogma e *ann*.txt *ann*-e.txt (pode-se observar que o nome do arquivo etiquetado gerado é o mesmo do original acrescido de "-e". Exemplo: ogma e a01.txt a01-e.txt).

---

[23] Na conversão do formato PDF para o TXT também não há distinção entre texto e cabeçalhos e rodapés. Logo, elementos tais como numeração de páginas foram misturados ao texto. Esse problema foi contornado eliminando-se tais números posteriormente.

[24] Utilizou-se aqui uma sequência de caracteres improvável de ser uma palavra de nossa língua e que fosse considerada pelo extrator de sintagmas nominais como uma palavra: "lamboriscadela". Tais demarcadores foram retirados logo após tal extração e anotadas as suas respectivas posições em relação à sequência de sintagmas nominais extraídos.

[25] O criador da ferramenta Ogma disponibilizou gentilmente uma nova versão, a 0.10 (sendo a anterior a 0.9), para que a mesma atendesse às necessidades dos recursos usados nesta pesquisa.

2. Extração dos SNs: a partir de cada texto etiquetado com o nome no formato *ann*-e.txt foi gerado um novo arquivo. Esse arquivo é o resultado da extração dos SNs do texto com base nas regras definidas por Maia (2008). Usou-se a seguinte sintaxe de comando para este procedimento:
    - ogma s *ann*-e.txt *ann*-s.txt (pode-se observar que o nome do arquivo gerado com a sequência de SNs extraídos é o mesmo do original acrescido de "-s". Exemplo: ogma s a01-e.txt a01-s.txt).
3. Limpeza dos SNs: a partir de cada listagem de SNs foi realizado um procedimento para a melhoria dos resultados baseado na elaboração pelo autor de macros de aplicação[26] dentro do Microsoft Office Word 2007 (o nome do arquivo gerado com a sequência de SNs extraídos já limpos é o mesmo do original acrescido de "-sl". Exemplo: a01-sl.txt). A limpeza dos SNs considerou os seguintes resultados encontrados[27] a partir do Ogma:
    - Alguns SNs extraídos apresentaram no seu início palavras como preposições, pronomes definidos, pronomes indefinidos, pronomes possessivos, pronomes demonstrativos, conjunções, verbos no gerúndio, artigos e advérbios, assim como suas respectivas contrações; e ainda *stopwords* da língua inglesa.
    - Alguns SNS extraídos pelo Ogma foram números puros (como aqueles decorrentes das numerações de páginas) ou até mesmo compostos somente por *stopwords*.
4. Classificação da estrutura: após a limpeza da sequência de SNs extraídos foi gerado um novo arquivo que contém a classificação da estrutura de cada SN. Usou-se a seguinte sintaxe de comando da ferramenta Ogma para este procedimento:
    - ogma tra *ann*-sl.txt *ann*-tral.txt (pode-se observar que o nome do arquivo gerado com a sequência de sintagmas nominais extraídos é o mesmo do original acrescido de "-tral". Exemplo: ogma tra a01-sl.txt a01-tral.txt).

As *macros* do Microsoft Office Word 2007 aqui elaboradas pelo autor para a limpeza dos SNS extraídos pelo Ogma encontram-se no APÊNDICE D. Ao final desses procedimentos descritos, para cada tese obteve-se a listagem final de todos os SNs já com

[26] As *macros* de aplicação consistem na automatização da execução de funções.
[27] A listagem completa dos termos retirados no processo de limpeza encontra-se em APÊNDICE C.

os procedimentos de limpeza aplicados (arquivos com a seguinte sintaxe "*ann*-sl.txt"), assim como a sua respectiva lista de classificação estrutural dos seus SNs (arquivos "*ann*-tral.txt").

1.6 Determinação dos sintagmas nominais como candidatos a descritores

Após a extração de todos os SNS, em um total de 995.688, de todas as teses de cada uma das seções do *corpus*, foi possível determinar um conjunto de SNs como candidatos a descritores de cada tese. Para isso, os SNs de cada texto foram processados através do *software* Microsoft Office Excel 2007, com o objetivo de atribuir a cada um deles uma pontuação referente à sua relevância como possível descritor de tal texto.

Foi aplicada nesta etapa a metodologia proposta por Souza e Raghavan (2006) para a atribuição de uma pontuação de SNs como descritores e que pode ser resumida na sua seguinte fórmula:

Equação 2 - Pontuação de um sintagma nominal como descritor

$$\text{Score}(NP) = f_{ij} * \log\left(\frac{N}{n_i}\right) * CNP$$

Fonte: SOUZA; RAGHAVAN, 2006.

Onde:

- NP→ *noun frase* = sintagma nominal;
- f_{ij}→ frequência do SN *i* no documento *j*;
- N→ número de documentos no corpus;
- n_i→ número de documentos que contém o SN *i*;
- CNP→ categoria do SN.

Para cada categoria do sintagma nominal (CSN) acima foram atribuídos os seus correspondentes valores também propostos por Souza e Raghavan (2006) e descritos na Tabela 6.

Tabela 6- Valores das categorias de sintagmas nominais (CSN)

Categoria	Nível e Estrutura do SN	Valor
1a	Nível 1, nome + determinante (N + D)	0,2
1b	Nível 1, exceto N + D	0,8
2	Nível 2	1,1
3	Nível 3	1,4
4	Nível 4	1,2
5	Nível 5 ou maior	0,8

Fonte: SOUZA E RAGHAVAN, 2006.

Para a obtenção da pontuação acima foram realizados os seguintes procedimentos para cada *corpus*:

1. Foi criado um arquivo no Excel 2007, para cujo nome usou-se a seguinte sintaxe "*A*.xlsx": [letra do grupo] + ".xlsx" (extensão de arquivo do Excel 2007). Exemplos: A.xlsx, B.xlsx, etc.;
2. Cada arquivo gerado com a sequência de SNs extraídos já limpos (do tipo *ann*-sl.txt, como por exemplo, a01-sl.txt) foi importado para o arquivo do Excel do grupo (*A.*xlsx) em uma planilha com nome seguindo a sintaxe *ann* (Exemplo: A01, A02, etc.);
3. Em cada uma das planilhas *ann* foi realizada uma nova limpeza para a retirada dos SNs que não tiveram sua estrutura classificada pelo Ogma. Para isso, foi importado para a mesma planilha *ann* o arquivo com a classificação de estruturas dos SNs *ann*-tral.txt. Os SNs que não estavam presentes em ambas as listagens foram movidos de *ann* para uma planilha de nome *a.*erros. Para a comparação entre as duas listagens foi utilizada a função do Excel PROCV.
4. Em cada uma das planilhas *ann* foi identificada a posição dos demarcadores logo após a introdução e antes da parte final, como conclusão ou considerações finais. Tais demarcadores foram retiradas e as suas respectivas posições anotadas na planilha denominada "*a.*Corpus";
5. A partir de cada planilha *ann* foi criada uma nova planilha, com o nome cuja sintaxe foi definida como *anns* (Ex.: A01s, A02s, etc.). Nessa planilha foi contabilizada inicialmente a quantidade de ocorrências de cada SN no texto. Para realizar essa contagem foi utilizado o recurso de Tabela Dinâmica do Excel 2007;

6. Na mesma planilha anterior, *anns,* foi contabilizada também a quantidade de textos do *corpus* nos quais há ocorrência de cada SN. Para realizar isso foi utilizada uma fórmula do Excel CONT.SE.
7. Ainda na mesma planilha, foi levantada, em uma nova coluna, a classificação da estrutura do SN (entre 1a, 1b, 2, 3, 4 e 5). A partir desta coluna, foi levantada uma segunda, com o respectivo valor definido na Tabela 6. Para preencher ambas as colunas foi usada novamente a função "PROCV" do Excel;
8. Por fim, na planilha *anns*, foi definida a pontuação de cada SN usando-se a Equação 2 em função dos resultados intermediários dos procedimentos anteriores e os mesmos foram ordenados de forma decrescente.

Na Tabela 7, é apresentado um exemplo do resultado obtido em uma das teses do corpus[28].

Tabela 7 - Exemplo de sintagmas nominais eleitos como candidatos a descritores

Sintagma Nominal	Frequência	Documentos	CSN Valor	Pontuação
escolar dos filhos	62	1	1,4	0,348262604
meses de abril	44	1	1,1	0,194192512
professores do município	46	2	1,1	0,158740032
questionário aplicado	46	1	0,8	0,147650505
escolarização dos filhos	17	1	1,4	0,095491359
escolha do estabelecimento	18	1	1,1	0,079442391
maio e junho de 2009	18	1	1,1	0,079442391
escolares dos filhos	13	1	1,4	0,073022804
professora de ciências	20	2	1,1	0,069017405
professor de ciências	19	2	1,1	0,065566535
caso dos pais	11	1	1,4	0,061788527
professores do grupo	21	3	1,1	0,060643543
professora de geografia	17	2	1,1	0,058664794
dois filhos	108	3	0,2	0,05670565
professora de matemática	22	4	1,1	0,054742036
famílias do grupo	12	1	1,1	0,052961594
famílias fortemente orientadas	8	1	1,1	0,035307729

[28] A tese usada como exemplo possui o título "Pais professores e a escolarização dos filhos" do Programa de Pós-Graduação em Educação: Conhecimento e Inclusão Social.

Sintagma Nominal	Frequência	Documentos	CSN Valor	Pontuação
para o sucesso				
escola dos filhos	6	1	1,4	0,033702833
bom aluno	10	1	0,8	0,032097936
total 114	10	1	0,8	0,032097936

Fonte: Elaborado pelo autor.

As macros do Microsoft Office Excel 2007 aqui elaboradas pelo autor para a determinação dos SNs como candidatos a descritores encontram-se no APÊNDICE E. Ao final dos procedimentos descritos, para cada tese, obteve-se a listagem final de 20 SNs candidatos a descritores, tal como o exemplo na Tabela 7 (acessíveis nas planilhas “ann” das pastas de trabalhos a.xlsx, sendo que a representa o grupo e nn a tese).

1.7 Aplicação dos questionários aos entrevistados

Os SNs de cada tese, uma vez pontuados de acordo com Equação 2, foram submetidos aos respectivos autores das teses de modo que eles avaliassem a relevância de tais sintagmas como descritores. Optou-se pelos próprios autores pois assumiu-se que eles seriam os especialistas mais viáveis para realizar o julgamento de relevância dos descritores de suas próprias teses. Outra técnica, empregada por Souza (2005) em um *corpus* de artigos de Ciência da Informação, poderia ter sido a do próprio pesquisador realizar tal julgamento com base nos títulos e resumos de cada tese. Porém, dada à diversidade de áreas de conhecimento no *corpus*, optou-se por recorrer aos especialistas.

A quantidade de SNs submetida aos autores foi vinte. Esse recorte foi devido a dois fatores: o primeiro seria relativo à quantidade, que deveria ser a maior possível, para que mesmo SNs com pontuação baixa pudessem ter chance de serem avaliados como bons descritores; e a segunda questão seria que quantidade não poderia ser muito extensa, de modo a impactar no tempo necessário para o preenchimento dos questionários, o que poderia levar a uma menor adesão dos respondentes.

Os vinte SNs foram alfabeticamente ordenados, retirando-se a ordem gerada pela pontuação. Evitou-se aqui criar uma tendência de respostas de acordo com a pontuação, pois “a maneira como os dados são procurados determina a natureza dos dados recebidos” (BABBIE, 1999, p. 193). Ou seja, “a ordem na qual são feitas as perguntas pode afetar a resposta, bem como toda a coleta de dados” (BABBIE, 1999, p. 205).

Para cada SN foi adotada a escala Likert, que permite uma ponderação uniforme em uma direção e “são usados para se fazer uma *análise de itens* levando à escolha dos melhores itens” (BABBIE, 1999, p. 233). Os limitadores da escala adotada foram: *Não*

Relevante e *Extremamente Relevante*. Foram usadas sete categorias entre os limitadores: os números de 1 (um) a 7 (sete)[29]. Optou-se por aumentar o número de níveis, uma vez que o necessário seriam somente quatro, para haver maior probabilidade de precisão nas respostas.

Cada questionário foi elaborado através de um formulário eletrônico[30] acessível pela Internet através de uma URL que foi informada ao entrevistado por email. Cada formulário continha: dados da tese, uma breve orientação para o preenchimento e a listagem dos SNs seguidos dos respectivos campos da escala. Para evitar dados faltantes nas respostas dos autores, usou-se ainda como recurso eletrônico a obrigatoriedade de resposta de todos os itens do questionário antes do envio do mesmo. Um exemplo do modelo de formulário eletrônico enviado está no APÊNDICE F.

As respostas dos formulários de cada entrevistado foram automaticamente salvas em planilhas eletrônicas. Além dos dados das respostas, foi disponibilizado em tais planilhas o horário e a data de envio realizado pelo entrevistado. O tempo total decorrido entre o envio da primeira resposta e o da última foi de exatamente 4 meses, sendo que, para alguns entrevistados, foi necessário enviar até quatro emails até se obter uma resposta.

1.8 Distribuição dos valores de relevância dos descritores por suas respectivas posições nas teses

Foram obtidas as respostas de 100% dos entrevistados (um total de 98), avaliando a relevância de cada um dos descritores candidatos (um total geral de 1.960). Aproximadamente 1/5 dos descritores candidatos (22%) foram considerados não relevantes.

Embora os entrevistados tenham optado, para cada descritor, por um dentre sete níveis de resposta (com limitadores *Não Relevante* e *Extremamente Relevante*), suas respostas foram tratadas em somente quatro níveis, adotando-se o mesmo procedimento empregado por Souza (2005, p. 93) como demonstrado na Tabela 8.

[29] Na análise de dados foram consideradas somente as respostas de 2 a 7 em três níveis proporcionais de relevância: moderada (níveis 2 e 3), razoável (4 e 5) e extrema (6 e 7).
[30] Foi utilizado o recurso de edição de formulários presente no pacote de aplicativos Google Drive da empresa multinacional Google Inc. disponível em http://drive.google.com.

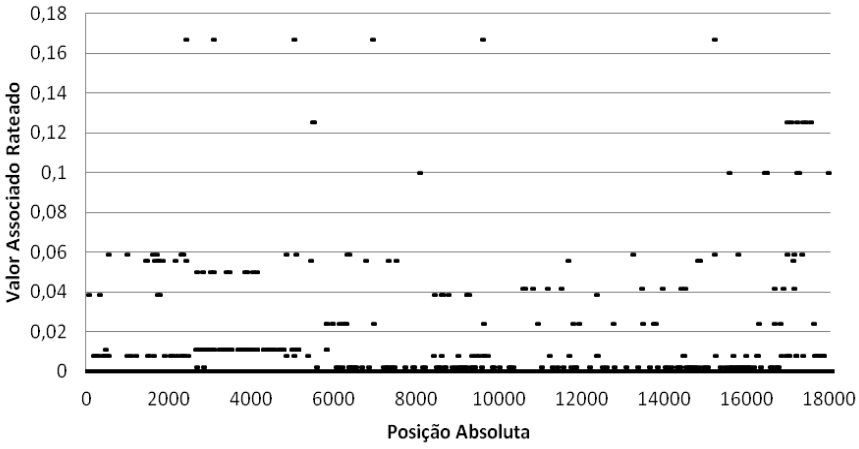
Tabela 8 - Valor associado aos níveis de respostas dos questionários

Questionário		Tratamento das Respostas	
Escala (nível)	**Limitadores**	**Nível de Relevância**	**Valor Associado**
7	Extremamente Relevante	Extremamente	1,00
6		Extremamente	1,00
5		Razoavelmente	0,50
4		Razoavelmente	0,50
3		Moderadamente	0,25
2		Moderadamente	0,25
1	Não Relevante	Não Relevante	0,00

Fonte: Elaborado pelo autor.

O valor associado para cada resposta referente aos descritores candidatos foi distribuído em posições nas suas respectivas teses. O valor associado de cada descritor foi rateado entre todas as suas ocorrências ao longo da tese. O Gráfico 3 exemplifica a distribuição desses valores rateados.

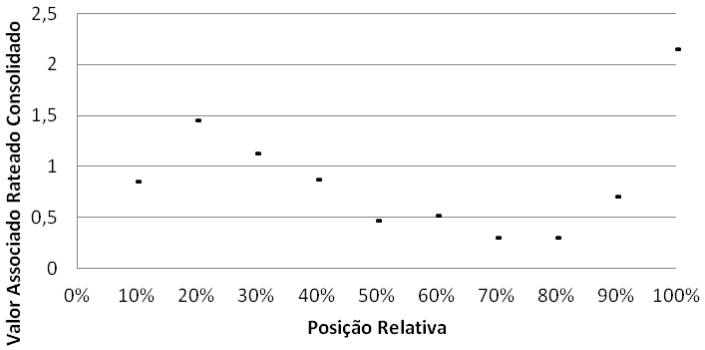
Gráfico 3 - Exemplo de valor associado rateado por posição absoluta



Fonte: Elaborado pelo autor.

Foi utilizada, assim como no pré-teste, a posição relativa (em %) ao tamanho total da tese (medido em número de SNs extraídos). Os valores associados rateados foram consolidados a cada 10% da posição relativa.

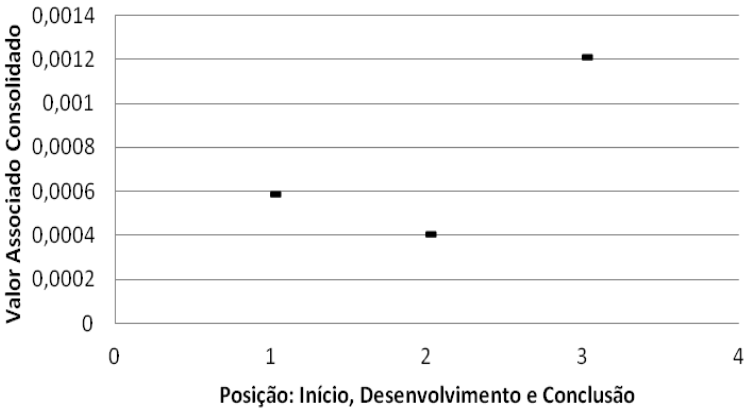
Gráfico 4 - Exemplo de valor associado rateado consolidado por posição relativa



Fonte: Elaborado pelo autor.

Outra estrutura de posição foi considerada: a relativa ao início, desenvolvimento e conclusão da tese. Para esta forma de distribuição, os valores associados foram consolidados de acordo com os delimitadores de início/desenvolvimento e desenvolvimento/conclusão levantados durante o tratamento dos *corpora*.

Gráfico 5 - Exemplo de valor associado consolidado por posição de início, desenvolvimento e conclusão

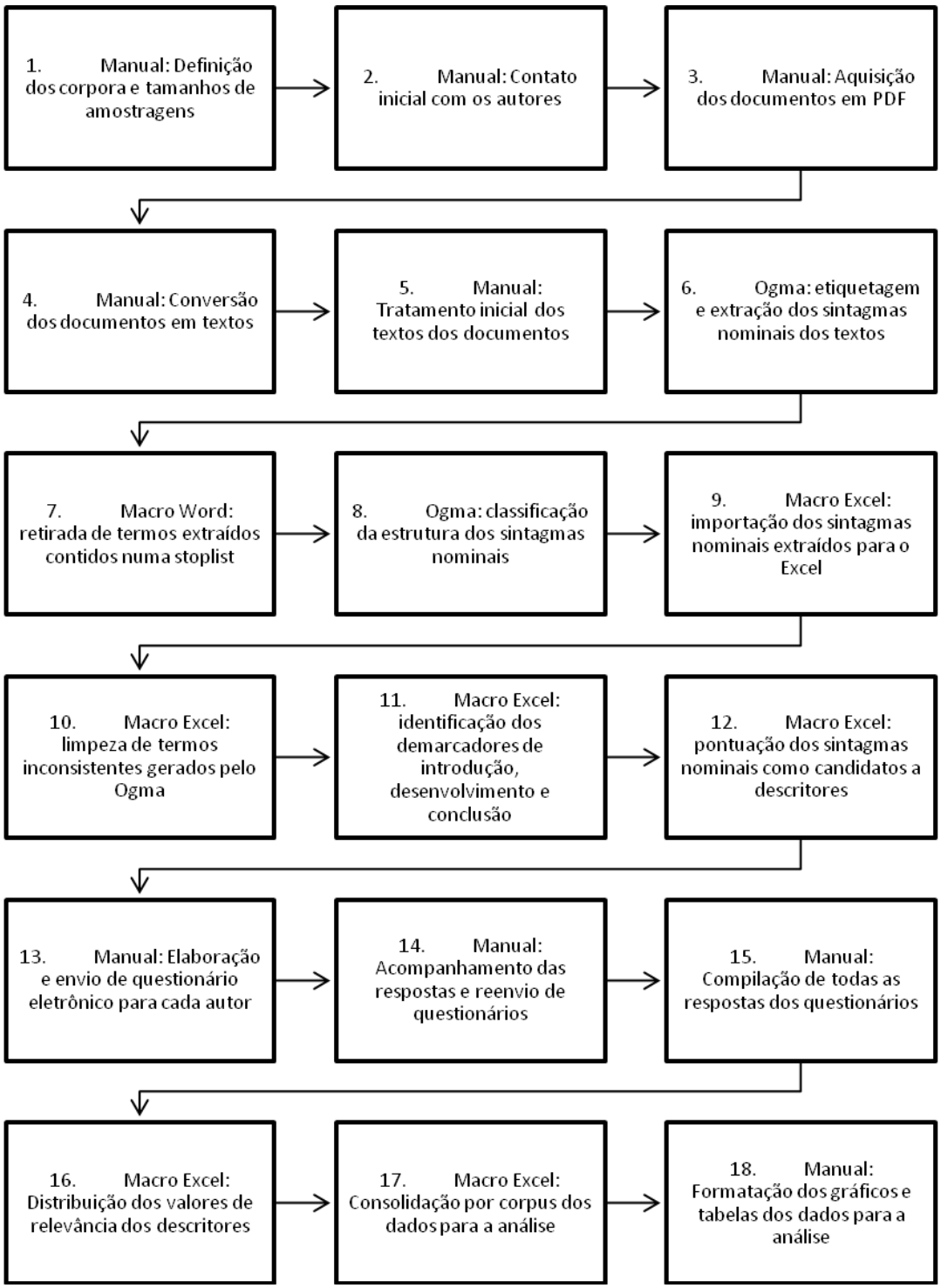


Fonte: Elaborado pelo autor.

Para a realização da consolidação dos valores associados conforme descrito anteriormente, foram elaboradas pelo autor *macros* do Microsoft Office Excel 2007 que se encontram no APÊNDICE G. Ao final desses procedimentos descritos, para cada tese obteve-se a listagem final de todos os valores consolidados por posição relativa e por posição de início/desenvolvimento/conclusão. Todos esses valores são analisados no próximo capítulo.

A seguir, a metodologia aplicada foi resumida na Figura 7.

Figura 7 - Fluxograma da metodologia aplicada



Fonte: Elaborado pelo autor.

4 Apresentação e análise dos resultados

A metodologia descrita no capítulo anterior e aplicada nesta pesquisa teve como principal pressuposto analisar a distribuição de descritores relevantes ao longo de um texto. O intuito foi verificar a existência de um comportamento padrão que pudesse ser usado como critério de indexação. Tal objetivo foi confirmado e está detalhado ao final deste capítulo.

Outro pressuposto foi avaliar a metodologia de escolha automática de descritores utilizando SNs elaborado por Souza (2005) e adaptado por Souza e Raghavan (2006). Tal pressuposto também obteve êxito e é analisado neste capítulo.

Outro pressuposto secundário foi o de avaliar a diferença de comportamento linguístico entre os oito programas de pós-graduação, tais como: proporção entre início/desenvolvimento/conclusão, quantidade média de SNs por tese e seu consequente tamanho numérico médio de palavras; e, por fim, sua variabilidade de distribuição de descritores relevantes ao longo do texto.

Durante a experimentação empírica e sua análise, outros resultados foram obtidos, tais como a verificação da existência de um comportamento padrão na distribuição de descritores relevantes ao longo de um texto. Foi possível determinar equações matemáticas que podem prever uma variação de, aproximadamente, 12% na relevância de um SN de acordo com sua posição no texto.

Também foi possível estimar que um aumento na quantidade de descritores candidatos para cerca de 40 (ao invés dos 20 usados em cada tese) pode levar a uma eleição aproximada de 100% dos descritores relevantes. A partir desse recorte na quantidade de descritores candidatos, também foi possível chegar a outro resultado: a determinação de uma pontuação de corte na metodologia proposta por Souza e Raghavan (2006).

4.1 Análise do corpus e teses selecionadas

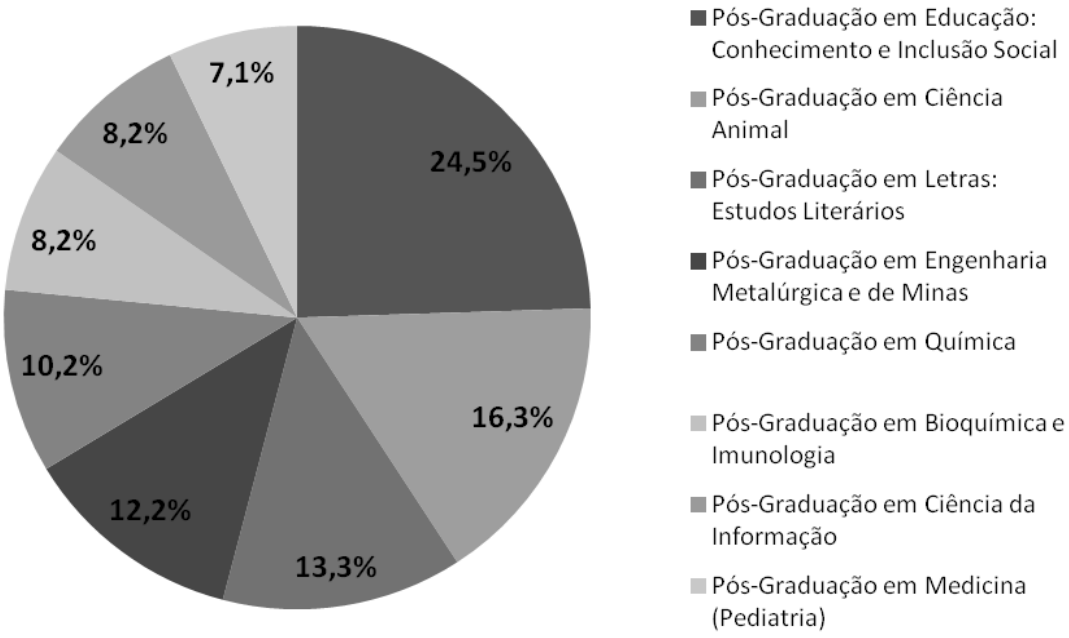
O *corpus* foi constituído de oito seções, sendo que cada uma delas representou umas das oito áreas de conhecimento da UFMG. O total de teses analisadas foram noventa e oito, distribuídas para cada programa de pós-graduação conforme o Gráfico 6 e a Tabela 9.

Tabela 9 - Distribuição da quantidade de teses analisadas nos programas de pós-graduação

Seção do *corpus*	Área de Conhecimento	Programa de pós-graduação com maior nº de teses na mesma área de conhecimento	Qtd. Teses Analisadas	%
A	Ciências Humanas	Pós-Graduação em Educação: Conhecimento e Inclusão Social	24	24,5%
B	Ciências Agrárias	Pós-Graduação em Ciência Animal	16	16,3%
C	Linguística, Letras e Artes	Pós-Graduação em Letras: Estudos Literários	13	13,3%
D	Engenharias	Pós-Graduação em Engenharia Metalúrgica e de Minas	12	12,2%
E	Ciências Exatas e da Terra	Pós-Graduação em Química	10	10,2%
F	Ciências Biológicas	Pós-Graduação em Bioquímica e Imunologia	8	8,2%
G	Ciências Sociais Aplicadas	Pós-Graduação em Ciência da Informação	8	8,2%
H	Ciências da Saúde	Pós-Graduação em Medicina (Pediatria)	7	7,1%
Total			**98**	100%

Fonte: Elaborado pelo autor.

Gráfico 6 - Quantidade de teses analisadas por programa de pós-graduação



Fonte: Elaborado pelo autor.

O período de publicação de todas as teses analisadas corresponde a aproximadamente 4,5 anos (de fev./2008 a ago./2012), sendo que, para cada programa de pós-graduação analisado, o período médio foi de 2,3 anos entre a tese mais antiga e a mais recente. O intervalo médio[31] entre as publicações na BDTD/UFMG para cada programa foi de 2,5 meses, conforme a Tabela 10.

Tabela 10 - Datas de publicação das teses analisadas na BDTD/UFMG

Seção do *corpus*	Publicação da Tese no BDTD/UFMG		Período analisado (anos)	Média de intervalo entre publicações (meses)
	Data mais antiga	Data mais recente		
A	26/02/2010	28/02/2012	**2,0**	**1,0**
B	26/02/2008	25/11/2011	**3,7**	**2,9**
C	08/07/2010	27/02/2012	**1,6**	**1,5**
D	26/02/2008	09/11/2011	**3,7**	**3,8**
E	24/02/2011	17/08/2012	**1,5**	**1,8**
F	19/02/2009	12/09/2011	**2,6**	**3,9**
G	30/11/2009	14/12/2011	**2,0**	**3,1**
H	26/02/2010	07/04/2011	**1,1**	**1,9**
Todos	26/02/2008	17/08/2012	**4,5**	**0,6**
Média do *corpus*			**2,3**	**2,5**

Fonte: Adaptado de BDTD/UFMG, 2012.

Pelo período médio de todas as teses de uma mesma seção do *corpus* ser de 2,3 anos, considera-se que as descrições linguísticas feitas aqui são *sincrônicas*, ou seja, foi considerado que todas as teses fizeram parte de um mesmo momento histórico social dos respectivos programas de pós-graduação.

A listagem completa de todas as teses analisadas no *corpus,* com suas respectivas datas de publicação na BDTD/UFMG, assim como título e autor, estão no APÊNDICE HH.

4.2 Análise da extração dos sintagmas nominais no *corpus*

[31] Para alguns programas, algumas teses dentro do período não foram analisadas: umas por não estarem disponíveis integralmente na BDTD/UFMG, outras por seus autores não poderem ser contactados.

Para a extração dos SNs, foram realizados, como descritos anteriormente, os processos de: escolha das teses, solicitação de confirmação de participação do autor da tese na pesquisa, obtenção da tese em PDF, conversão para o formato texto, retirada das partes pré e pós-textuais, demarcação entre início, desenvolvimento e conclusão. Todos esses processos foram realizados manualmente e duraram cerca de quatro meses, contando com a participação de terceiros.

Para a extração dos SNs, foram utilizadas as ferramentas Ogma, *macros* no Microsoft Word e *macros* no Microsoft Excel, como também descrito anteriormente. Durante o uso destas ferramentas, pôde-se calcular com precisão os tempos gastos em horas e minutos. Na Tabela 11, a seguir, é possível verificar que a média de tempo para a extração foi de aproximadamente 81,8% somente para a ferramenta Ogma.

Tabela 11 - Tempo de processamento para extração dos sintagmas nominais

Tempo (hora:min.)	A	B	C	D	E	F	G	H	Total	Total (%)
1º Processamento do Ogma	03:32	00:53	02:14	00:36	00:58	00:25	00:50	00:24	**09:52**	**60,1%**
Processamento de Macro do Word	00:30	00:13	00:25	00:14	00:09	00:11	00:13	00:05	**02:00**	**12,2%**
2º Processamento do Ogma	01:02	00:21	00:55	00:18	00:17	00:08	00:22	00:11	**03:34**	**21,7%**
Processamento de Macro do Excel	00:31	00:03	00:14	00:02	00:02	00:01	00:04	00:02	**00:59**	**6,0%**
Total Tempo	**05:35**	**01:30**	**03:48**	**01:10**	**01:26**	**00:45**	**01:29**	**00:42**	**16:25**	**100,0%**
Quantidade de Teses (unid.)	24	16	13	12	10	8	8	7	**98**	
Média de tempo portese (hora:min.)	**00:13**	**00:05**	**00:17**	**00:05**	**00:08**	**00:05**	**00:11**	**00:06**	**00:10**	

Fonte: Elaborado pelo autor.

A média de tempo de processamento para a extração dos SNs foi de dez minutos por tese. Podemos objetivar que o tempo de processamento é proporcional à quantidade de sintagmas nominais extraídos, sendo que a média aproximada foi de 1 (um) minuto para cada 1.000 (mil) extrações, conforme pode ser visto na Tabela 12, a seguir:

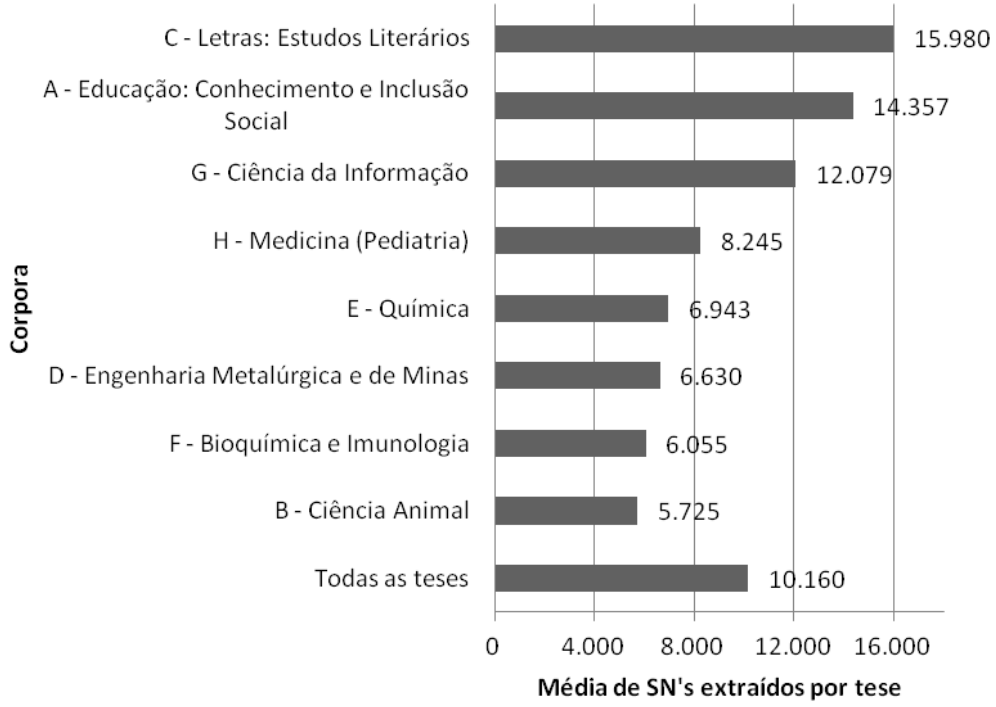
Tabela 12 - Média de tempo de processamento por 1.000 sintagmas nominais extraídos

Corpus	Tempo total de processamento (hora:min.)	Quantidade total de sintagmas nominais extraídos	Média de tempo por 1.000 sintagmas nominais extraídos (min.:seg.)
A	05:35	344.576	00:58,3
C	03:48	207.746	01:05,8
G	01:29	96.631	00:55,3
B	01:30	91.599	00:59,0
D	01:10	79.560	00:52,8
E	01:26	69.429	01:14,3
H	00:42	57.714	00:43,7
F	00:45	48.436	00:55,7
Todos	**16:25**	**995.691**	**00:59,4**

Fonte: Elaborado pelo autor.

As seções do *corpus* que apresentaram maiores médias de tempo por tese, apresentadas na Tabela 11, também foram aquelas que apresentaram as maiores médias de SNs extraídos por tese, conforme pode ser visto a seguir no Gráfico 7.

Gráfico 7 - Média de sintagmas nominais extraídos por tese em cada seção do corpus



Fonte: Elaborado pelo autor.

Podemos considerar tradicionalmente a existência das ciências naturais e das ciências sociais em um nível mais generalista. Embora haja uma tendência de superação dessa dicotomia[32] (SANTOS, 1996), pôde-se perceber, no Gráfico 7, que nas seções do *corpus* de programas de pós-graduação mais relacionados às ciências sociais houve uma quantidade acima da média de SNs extraídos, assim como, em todas as seções do *corpus* relacionadas às ciências naturais, essa quantidade foi abaixo da média. Para Dubois *et al.* (1973), há uma concepção distinta de estruturas[33] para as ciências humanas e para as ciências mais relacionadas aos sistemas lógicos e matemáticos, existindo para estas uma maior *autorregulação*, na medida em que permanecem mais estáveis temporalmente[34]. Tal estabilidade é considerada aqui como fator primordial para a constatação da maior objetividade[35] das teses relacionadas às ciências naturais considerando-se o seu menor uso em quantidade de SNs.

Em relação à quantidade de SNs, dentre as principais pesquisas referenciadas aqui e que realizaram extração de SNs na língua portuguesa, assim como a presente pesquisa, podemos citar Kuramoto (1999) e Souza (2005), que utilizaram artigos científicos da Ciência da Informação nos seus *corpora*; Maia (2008) que utilizou artigos científicos também da Ciência da Informação e textos jornalísticos de outras áreas; e ainda Corrêa *et al*. (2011) que utilizaram resumos de teses e dissertações nas áreas de Direito, Computação e Nutrição. Neste momento, podemos comparar inicialmente a quantidade de SNs extraídos entre todas essas pesquisas conforme Tabela 13, a seguir:

---

[32] Para Santos (1996), todo conhecimento científico-natural é científico-social, sendo que esta última preferiu "a compreensão do mundo à manipulação do mundo" (ibidem, p. 71).

[33] "Uma estrutura é um sistema caracterizado por noções de totalidade, de transformação, de autorregulação" e "se definem por uma série de relações entre os elementos; não é nem o elemento nem o todo, mas suas relações que constituem a estrutura, e o todo não é senão o seu resultado" (DUBOIS, 1973, p. 247).

[34] Ainda para Dubois *et al.*(1973) um sistema linguístico está em constante transformação e ocorre de acordo com o comportamento linguístico dos integrantes de uma comunidade linguística.

[35] A objetividade de um texto, assim como as características determinadas por influências culturais, como a disparidade entre as tradições anglo-americanas e francesas; podem ser melhor analisadas com a Teoria dos Gêneros Textuais. Essa análise está fora do escopo dessa pesquisa em Ciência da Informação e é indicada para trabalhos futuros pela Linguística.

Tabela 13 - Comparação de extração de sintagmas nominais entre pesquisas

Pesquisas	Quant. de Documentos	Tipo de Documentos	Modo de Extração	SNs Extraídos	Média de SNs por Documento
Kuramoto (1999)	15	artigos científicos	manual	**8.818**	**588**
Souza (2005)	60	artigos científicos	automática	**76.739**	**1.279**
Maia (2008)	210	artigos científicos (50) e textos jornalísticos (160)	automática	**153.386**	**730**
Corrêa *et al.* (2011)	30	resumos de teses e dissertações	automática	**951**	**32**
Esta Pesquisa	98	teses	automática	**995.691**	**10.160**

Fonte: Elaborado pelo autor.

A quantidade de SNs extraídos nesta pesquisa corresponde a aproximadamente 6,5 vezes mais que a maior quantidade de SNs extraídos em pesquisas anteriores. Esse fato deve-se ao tipo de documento escolhido (tese), com o principal propósito da análise da distribuição de relevância (apresentada ainda neste mesmo capítulo), e a quantidade amostral utilizada para representar todas as áreas de conhecimento da UFMG.

Na Ciência da Informação, podemos comparar com precisão a diferença de tamanho médio, em quantidade de SNs, de um artigo científico, 1.279 (SOUZA, 2005, p. 127), e uma tese, 12.079 (valor apresentado aqui anteriormente), sendo este 9,4 vezes maior que o primeiro. É irresistível salientar aqui a curiosa coincidência numérica entre os dois valores, que são diferentes entre si apenas por um zero no meio de um deles.

Assim como em outras pesquisas, durante a extração de SNs, ocorreram extrações automáticas que não resultaram propriamente em SNs devido a falhas nos processos de extração. Corrêa *et al*. (2011) explicitaram uma taxa de erros de extração através do Ogma de 42%. Devido à pequena quantidade de SNs extraídos em tal pesquisa, os autores puderam constatar manualmente a efetividade de cada resultado da extração.

Para esta pesquisa, os erros puderam ser contatados em dois momentos de forma automática: através da retirada de *stopwords* residuais (APÊNDICE C) com o uso de *macros* do Microsoft Word (APÊNDICE D) e através da comparação de saídas

inconsistentes do próprio Ogma[36], usando-se para isso *macros* do Microsoft Excel (especificamente a sub-rotina LimpaSintagmaErroSlxTral no APÊNDICE F).

A taxa de erros encontrada aqui foi bem inferior (3,5 vezes menor) que a encontrada por Corrêa *et al*. (2011), conforme pode ser visto na % total de extrações excluídas na Tabela 14, a seguir:

Tabela 14 - Quantidade de exclusões de extrações de sintagmas nominais do Ogma

Seção do *corpus*	Sintagmas Nominais				
	Extraídos pelo Ogma	**Excluídos por *Stopwords* residuais**	**Excluídos por inconsistência no próprio Ogma**	**Considerados nesta pesquisa**	**% total de extrações excluídas**
A - Educação: Conhecimento e Inclusão Social	387.825	34.477	8.772	344.576	**11,2%**
B - Ciência Animal	105.499	12.269	1.631	91.599	**13,2%**
C - Letras: Estudos Literários	232.788	18.267	6.775	207.746	**10,8%**
D - Engenharia Metalúrgica e de Minas	92.151	11.330	1.261	79.560	**13,7%**
E - Química	83.635	13.020	1.186	69.429	**17,0%**
F - Bioquímica e Imunologia	54.532	5.140	956	48.436	**11,2%**
G - Ciência da Informação	109.712	10.884	2.197	96.631	**11,9%**
H - Medicina (Pediatria)	64.815	5.671	1.430	57.714	**11,0%**
Total	**1.130.957**	**111.058**	**24.208**	**995.691**	**12,0%**

Fonte: Elaborado pelo autor.

Uma análise manual em cada um dos SNs extraídos, como realizada por Corrêa *et al*. (2011), provavelmente chegaria a uma taxa de erros de extração superior aos 12,0% encontrados aqui. No entanto, dada a dimensão dessa análise para a quantidade aproximada de 1,1 milhões de SNs extraídos, mesmo que feita de forma estatisticamente amostral, e à baixa relevância para os objetivos fins desta pesquisa, tal taxa ficou limitada aos dados obtidos de forma automática.

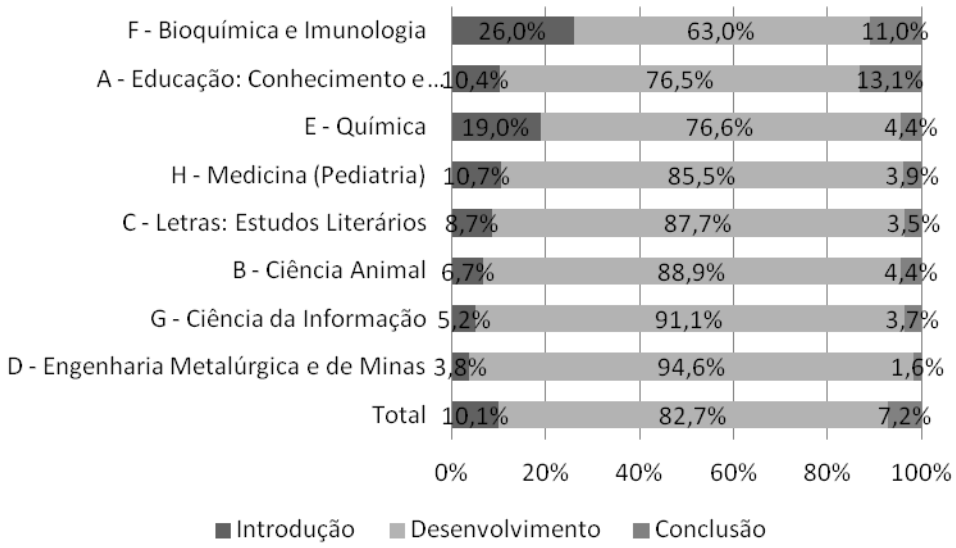
[36] O Ogma pode gerar uma lista dos sintagmas nominais em um texto através da opção "-s" assim como pode gerar uma análise da estrutura de cada sintagma nominal em um texto através da opção "-tral". Para as duas saídas, pôde-se constatar que alguns sintagmas nominais presentes em decorrência da saída "-s" não constavam na saída de "-tral", sendo verificados que eram erros de extração. Tais erros foram movidos para uma planilha com o nome padrão para cada seção do *corpus* denominada *A*.ERROS.

A seção do *corpus* que apresentou maior taxa de erros foi a correspondente ao programa de pós-graduação em Química, que possui como característica de seu sistema linguístico o uso de fórmulas químicas. No entanto, os fatores que influenciaram na sua elevada taxa de erros aqui foram: a elevada presença de números que foram descartados como *stopwords* residuais e o recorrente uso de expressões em inglês. Tais fatores foram constatados por uma exploração de leitura pelo autor nos resultados das extrações feitas pelo Ogma.

A seção do *corpus* que apresentou menor taxa de erros foi a correspondente ao programa de pós-graduação em Letras – Estudos Literários, que podemos considerar o mais metalinguístico dentre os outros programas. Ou seja, como afirma Dubois *et al*. (1973) aquele que usa a própria língua como objeto de seu discurso fazendo assim um distanciamento maior de outros sistemas linguísticos mais especialistas, como o lógico-matemático, que são mais passíveis de incorrerem em erros de extração em processadores de linguagem natural, que usam como base um dicionário geral da língua, como o Ogma.

Para o objetivo principal desta pesquisa, foi considerada para cada SN extraído a sua posição estrutural correspondente às partes de introdução, desenvolvimento e conclusão. Dentre essas, a de desenvolvimento conteve 82,7% dos SNs, enquanto as outras duas dividiram o restante em 10,1% para a introdução e 7,2% para a conclusão, como pode ser visto no Gráfico 8, a seguir:

Gráfico 8 - Distribuição de sintagmas nominais por partes da tese



Fonte: Elaborado pelo autor.

A maior distribuição de SNs nas partes de introdução e conclusão ocorreu no programa de pós-graduação em Bioquímica e Imunologia, enquanto o programa que

concentrou mais SNs na parte de desenvolvimento foi o de Engenharia Metalúrgica e de Minas. O comportamento linguístico que levou a essas diferenças de distribuição pode merecer uma análise estilística da linguística, na qual é considerada a marca individual do sujeito (DUBOIS, 1973, p. 243) deva ser considerada. Tal análise foge ao escopo dessa pesquisa, por ser necessária uma leitura integral de todas as obras sob um olhar crítico, sendo que o objetivo aqui está relacionado a procedimentos automatizados.

4.3 Análise da seleção dos sintagmas nominais candidatos a descritores

Nesta pesquisa, como foi apresentado no capítulo da metodologia, para um SN ser eleito como descritor considerou-se em sua candidatura: sua frequência na própria tese, a quantidade de teses da seção do *corpus* nas quais ele ocorre, seu nível de estrutura como SN e, para sua eleição como descritor, a avaliação da relevância do SN dada pelo próprio autor da tese.

Como a Equação 2, apresentada na página 57, indica a frequência de um SN no mesmo documento é um dos fatores mais importantes para a escolha de um SN como descritor. Foi possível concluir aqui que um mesmo SN ocorre, em média, aproximadamente duas vezes em uma mesma tese.

O total de SNs identificados em cada tese correspondeu a 53,5% do total dos que foram extraídos. Ou seja, esse valor corresponde à quantidade de SNs que são distintos entre si frente ao total extraído. A Tabela 15 a seguir apresenta um detalhamento desses dados por seção do *corpus*.

Tabela 15 - Sintagmas nominais identificados em relação aos extraídos

Seção do *corpus*	Sintagmas Extraídos	Sintagmas Identificados	% Sintagmas Identificados
A - Educação: Conhecimento e Inclusão Social	344.576	180.737	52,5%
B - Ciência Animal	91.599	49.793	54,4%
C - Letras: Estudos Literários	207.746	116.324	56,0%
D - Engenharia Metalúrgica e de Minas	79.560	42.977	54,0%
E – Química	69.429	34.691	50,0%
F - Bioquímica e Imunologia	48.436	25.892	53,5%
G - Ciência da Informação	96.631	52.612	54,4%
H - Medicina (Pediatria)	57.714	30.138	52,2%
Total	**995.691**	**533.164**	**53,5%**

Fonte: Elaborado pelo autor.

A respeito da relação entre a quantidade de SNs identificados e o total de extraídos, Kuramoto (1999) obteve manualmente 8.818 destes e identificou 75,2% deles como sem repetições. Souza (2005) utilizou artigos da Ciência da Informação e extraiu automaticamente 76.739 SNs, sendo que 78,9% destes eram únicos. Já nesta pesquisa, esse mesmo valor caiu consideravelmente para 53,5%. Presume-se aqui que o principal motivo para essa queda seja a dimensão das teses (apresentadas aqui, para a Ciência da Informação, por exemplo, como em média 9,4 vezes maior que um artigo).

A probabilidade de um mesmo autor repetir termos em um discurso aumenta com o tamanho do texto, uma vez que a quantidade de possíveis SNs deriva da quantidade de palavras de uma língua, que é limitada sincronicamente[37]. Essa probabilidade é acentuada uma vez que o discurso de cada tese, como já indica o seu próprio pertencimento a um único programa de pós-graduação, deve centrar-se em uma *área específica de atuação*[38]. E, por fim, como todo texto científico, ao manter uma estrutura coerente, uma tese tende a fazer referências de conceitos já mencionados em seu próprio texto, aumentando assim as chances de repetição de termos.

Novamente, pôde ser observada uma maior singularidade na seção do *corpus* correspondente ao programa de pós-graduação em Letras – Estudos Literários, cuja porcentagem de SNs identificados é a maior dentre os demais programas. Embora a diferença entre as demais seções seja relativamente pequena, podemos ainda perceber que, em tais teses, há uma possibilidade de maior densidade de conceitos, associados aqui aos SNs identificados. Outra hipótese pode estar relacionada ao estilo caracterizado pelo emprego de referências diversificadas, ou seja, quando o autor, para falar de um mesmo conceito, evita usar os mesmos termos. Para confirmar tais hipóteses, novamente, faz-se necessária uma análise diretamente nas teses usadas sob esse viés[39].

Já o programa de pós-graduação em Química apresenta, além da maior incidência de exclusões de extração já demonstrada, o maior índice de repetições de um mesmo SN. Foi considerada a seguinte hipótese para a causa deste fato: em tal comunidade ocorreria um uso do sistema linguístico mais especializado e mais controlado

---

[37] Embora aqui haja a possibilidade de um sintagma nominal ter tamanho arbitrário, é considerado aqui que em um sistema linguístico haja um máximo empregado dentre a totalidade de comportamentos linguísticos de seus indivíduos.

[38] "O Doutorado tem por objetivo desenvolver a capacidade de propor e conduzir pesquisas originais, de forma autônoma, em área específica de atuação" (SODS, Secretaria dos Órgãos de Deliberação Superior -. Normas Gerais de Pós-Graduação da UFMG - Resolução Complementar 01/2009, de 27 de outubro de 2009. Disponível em: https://www2.ufmg.br/sods/Sods/CEPE/Documentos/Resolucoes-Complementares. Acessado em: 06 de abril de 2012).

[39] Outra hipótese ainda seria a melhor proeficiência de autores que realizam pesquisas em Literatura. Eles utilizariam estruturas mais diversificadas e de formas mais criativas. Essa hipótese pode ser melhor analisada com aportes na Linguística de *Corpus*.

que os outros. Ou seja, foi considerada como hipótese um maior grau de autorregulação, proporcionado pelo próprio sistema linguístico ou pela comunidade (como normatizações, por exemplo). Tal hipótese foi justificada com a constatação da existência de um compêndio de terminologia química, denominado por *Gold Book*, adotado internacionalmente e disponibilizado livremente pela *International Union of Pure and Applied Chemistry (*IUPAC). Tal compêndio, que está em língua inglesa, justifica a maior incidência de erros constatada na extração (que aqui foi feita para a língua portuguesa), e, por assemelhar-se a um vocabulário controlado, justifica sua maior homogeneidade de SNs dentre os demais programas de pós-graduação.

Dentre esses SNs identificados, aqueles que ocorreram ao longo da tese uma única vez corresponderam a 80,6%. Dentre aqueles que tiveram mais de uma ocorrência, a média da máxima repetição em cada seção do *corpus* correspondeu a 1,6% do total extraído. A Tabela 16 a seguir apresenta um detalhamento desses dados por seção do *corpus*.

Tabela 16 - Frequência única e máxima dos sintagmas nominais

Seção do *corpus*	Sintagmas Nominais Identificados que ocorreram apenas uma vez	Máxima repetição de um sintagma nominal dentre os sintagmas nominais extraídos
A - Educação: Conhecimento e Inclusão Social	82,1%	1,6%
B - Ciência Animal	76,8%	1,7%
C - Letras: Estudos Literários	83,2%	1,9%
D - Engenharia Metalúrgica e de Minas	79,5%	1,7%
E – Química	76,3%	2,0%
F - Bioquímica e Imunologia	78,2%	1,7%
G - Ciência da Informação	79,8%	1,2%
H - Medicina (Pediatria)	77,6%	1,4%
Total	**80,6%**	**1,6%**

Fonte: Elaborado pelo autor.

Embora a média de repetição de um mesmo SN tenha sido apresentada aqui como aproximadamente duas, foi possível perceber que somente um quinto dos SNs identificados ocorre mais de uma vez ao longo de uma tese. Além de esse valor possuir uma

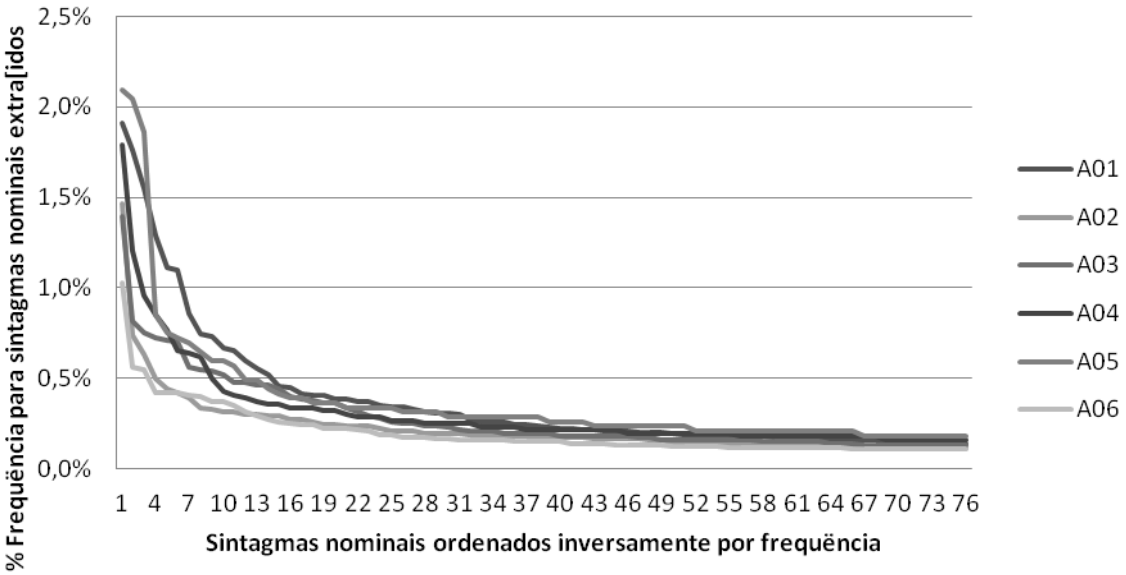
relação com o princípio de Pareto[40], uma vez que todos os SNs selecionados como candidatos a descritores apresentaram mais de uma ocorrência[41] (19,4% do total identificado), foi possível comprovar o comportamento da distribuição de frequências de acordo com a Lei de Zipf[42], como exemplificado em seis teses no Gráfico 9.

[40] O princípio de Pareto é conhecido por relacionar 80% de consequências a 20% de causas.

[41] A listagem completa dos sintagmas nominais eleitos como candidatos a descritores, assim como suas respectivas frequências e outros valores, está no APÊNDICE I.

[42] A lei do linguísta Zipf nasceu em conjunto com o princípio do menor esforço, postulando que o caminho mais natural é por onde haja menos resistência, e foi publicado em ZIPF, G.K. *Human Behavior and the Principle of Least Effort.* Cambridge, Massachusetts: Addison-Wesley. 1949.

Gráfico 9 - Exemplo de maiores frequências ordenadas de acordo com a Lei de Zipf



Fonte: Elaborado pelo autor.

A seção do *corpus* do programa de pós-graduação em Letras – Estudos Literários apresentou a maior média de SNs únicos (83,2%). Uma vez que seus textos são os relativamente mais longos (como já apresentado aqui) há mais probabilidade de haver ocorrências de termos diferentes[43], seja por tratar de assuntos mais distintos, seja por usar termos mais distintos para os mesmos assuntos. O programa de pós-graduação em Química apresentou a maior quantidade de SNs com mais de uma ocorrência, assim como o maior índice de repetições de um mesmo SN (2,0%). Esse fato pode estar, mais uma vez, relacionado ao uso do que se assemelha a um vocabulário controlado internacional, como o *Gold Book*.

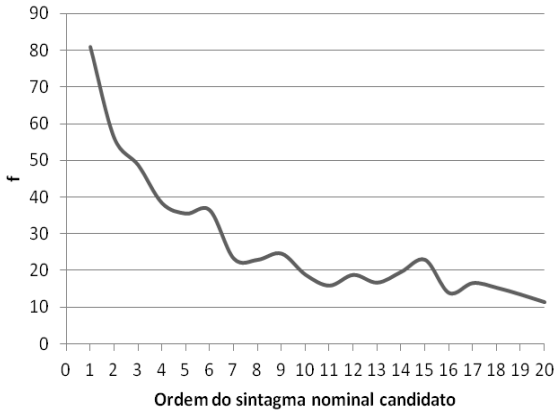
Conforme já descrito na metodologia apresentada aqui, foi utilizada a Equação 2 para a pontuação de todos os SNs identificados em cada tese e, em seguida, realizado o corte dos 20 primeiros de cada uma como seus respectivos candidatos a descritores. Logo, para as 98 teses, foram selecionados um total de 1.960 candidatos. Todos os valores encontrados e usados para a pontuação de cada sintagma nominal selecionado como candidato estão no APÊNDICE I.

A frequência (f) de cada SN foi determinante para a ordenação dos SNs candidatos a descritores. Os outros dois fatores, o número de documentos na seção do *corpus* que

[43] Graciliano Ramos, em "A Terra dos Meninos Pelados" de 1939, descreve uma cena na qual, em um mundo imaginário (Tatipirun), macacos jogam dados de letras até formarem palavras. Um personagem, que observa tal cena, acredita então que se esse jogo continuar infinitamente, todas as palavras serão formadas, assim como todos os poemas já escritos e até mesmo todos os livros. Uma bela forma de se ver probabilidades linguísticas.

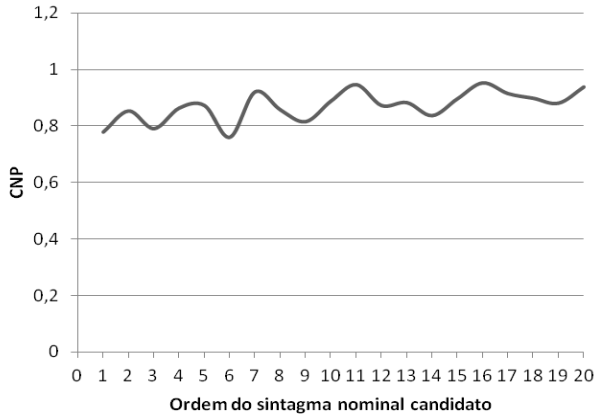
continham o SN (Log N/n) e o valor da CNP, mantiveram-se praticamente estáveis do primeiro ao vigésimo candidato selecionado, conforme pode ser visto nas médias de todos os candidatos por ordem inversa de pontuação nos Gráfico 10 a Gráfico 13 a seguir:

Gráfico 10 - Média da frequência por ordem de sintagma nominal candidato



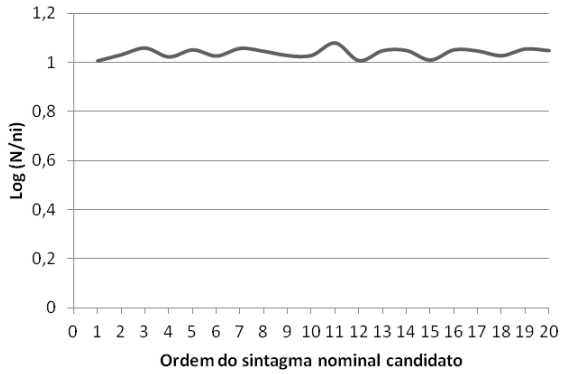
Fonte: Elaborado pelo autor.

Gráfico 12 - Média do valor da categoria do sintagma nominal



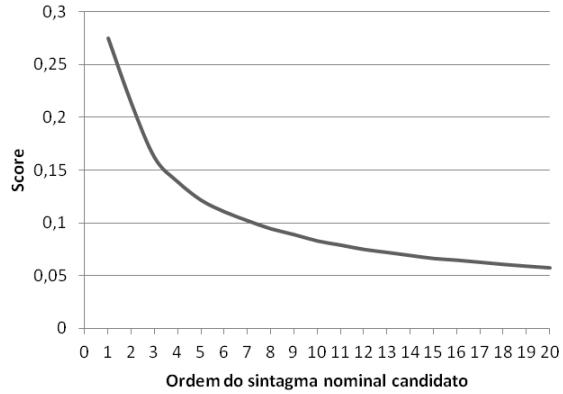
Fonte: Elaborado pelo autor.

Gráfico 11 - Média do log da razão do tamanho da seção do *corpus* pelo número de documentos na seção que contém o sintagma nominal



Fonte: Elaborado pelo autor.

Gráfico 13 - Média da pontuação (*score*) do sintagma nominal



Fonte: Elaborado pelo autor.

O comportamento da pontuação dos SNs candidatos deriva do comportamento da frequência, que, por sua vez, está relacionado à Lei de Zipf. O fator Log (N/n_i) foi importante como corte para SNs considerados como *stopwords*, ou seja, quando ocorrem em mais de 80% dos documentos (BAEZA-YATES; RIBEIRO-NETO, 2011, p. 226). O fator CNP apresentou comportamento mais aleatório e indica uma possível necessidade de revisão dos valores atribuídos na Tabela 6, localizado na página 58. Essa proposição é

detalhada a seguir com a análise dos mesmos fatores acima, porém associados às relevâncias atribuídas pelos autores das teses a cada candidato a descritor.

4.4 Análise da relevância como descritores dos sintagmas nominais candidatos

Todos os 1.960 SNs selecionados como candidatos a descritores foram submetidos aos próprios autores das teses para avaliação de suas respectivas relevâncias, como foi descrito no capítulo da metodologia. Esse procedimento durou cerca de quatro meses e obteve 100% de adesão dos entrevistados.

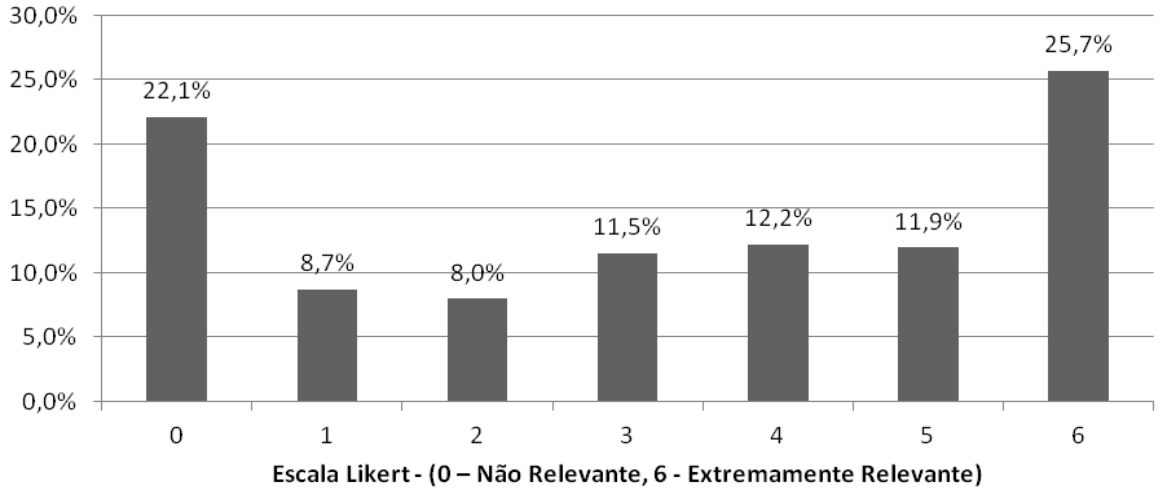
Os autores avaliaram que 77,9% dos SNs candidatos são relevantes como descritores. As respostas dadas na escala Likert (aqui apresentadas como que de 0 a 6) de todos os autores estão detalhadas por seção do *corpus* na Tabela 17 e apresentadas no seu total no Gráfico 14 a seguir:

Tabela 17 - Avaliação de relevância na escala Likert dos sintagmas nominais candidatos

Seção do *corpus* - Programa de Pós-graduação	Respostas de relevância na escala Likert (0 – Não Relevante, 6 - Extremamente Relevante)							% Relevante
	0	1	2	3	4	5	6	
A - Educação: Conhecimento e Inclusão Social	**20,0%**	9,0%	5,4%	12,9%	12,7%	12,3%	27,7%	**80,0%**
B - Ciência Animal	**26,6%**	8,4%	7,2%	11,3%	18,4%	11,6%	16,6%	**73,4%**
C - Letras: Estudos Literários	**16,2%**	10,4%	10,8%	12,7%	10,8%	13,8%	25,4%	**83,8%**
D - Engenharia Metalúrgica e de Minas	**25,0%**	9,6%	9,6%	13,8%	9,6%	11,3%	21,3%	**75,0%**
E - Química	**26,5%**	9,0%	5,0%	11,0%	11,0%	9,5%	28,0%	**73,5%**
F - Bioquímica e Imunologia	**28,1%**	6,9%	12,5%	8,8%	10,0%	11,9%	21,9%	**71,9%**
G - Ciência da Informação	**16,9%**	10,0%	11,3%	11,3%	13,8%	14,4%	22,5%	**83,1%**
H - Medicina (Pediatria)	**17,9%**	3,6%	5,7%	5,0%	5,7%	10,0%	52,1%	**82,1%**
Todos os *corpora*	**22,1%**	**8,7%**	**8,0%**	**11,5%**	**12,2%**	**11,9%**	**25,7%**	**77,9%**

Fonte: Elaborado pelo autor.

Gráfico 14 - Avaliação de relevância na escala Likert dos sintagmas nominais candidatos



Fonte: Elaborado pelo autor.

A avaliação dos autores apresenta uma concentração de distribuição nas extremidades. Esse fato pode estar relacionado ao fato do questionário enviado aos autores, conforme exemplo no APÊNDICE F, possuir descrição somente para os valores de extremidade, sendo que para os demais intermediários é apresentado somente o valor numérico. Somado a esse fato, há uma tendência do entrevistado minimizar seu esforço para responder o questionário e assumir como respostas uma dimensão binária (se é ou não relevante), desconsiderando os níveis intermediários.

Para a extremidade não relevante, seu valor estaria também associado a necessidades de melhoria no processo de seleção de SNs candidatos a descritores. Uma dessas melhorias pode ser direcionada de modo a evitar, por exemplo, que os seguintes candidatos enviados aos autores fossem selecionados de forma automática:

- Referências numéricas: 1996a, 1996b, 200°c, 240***, 300°c, 5ºc até 750ºc, inúmeros outros, total 240;
- Referências temporais: dezembro de 1948, dezembro de 2006, fins dos anos, janeiro de 2010, julho de 2009, maio e junho de 2009, ano de 2006, meados dos anos, meses de abril, meses de idade, período de janeiro de 2004, período de março de 2004;
- Nomes próprios: cecília, érica, heliane, herbert, janaina, jederson, júlia, lúcio, mariano, mateo, patrick, paula, sílvia, tiago, vanessa, vicente, wanda;
- Expressões sobre o próprio discurso: fig, foto alterada, foto da autora, graf, grifos do autor, grifos meus, idem, nome em citações bibliográficas, nota do organizador, página, páginas do livro, participantes da pesquisa, resultados e

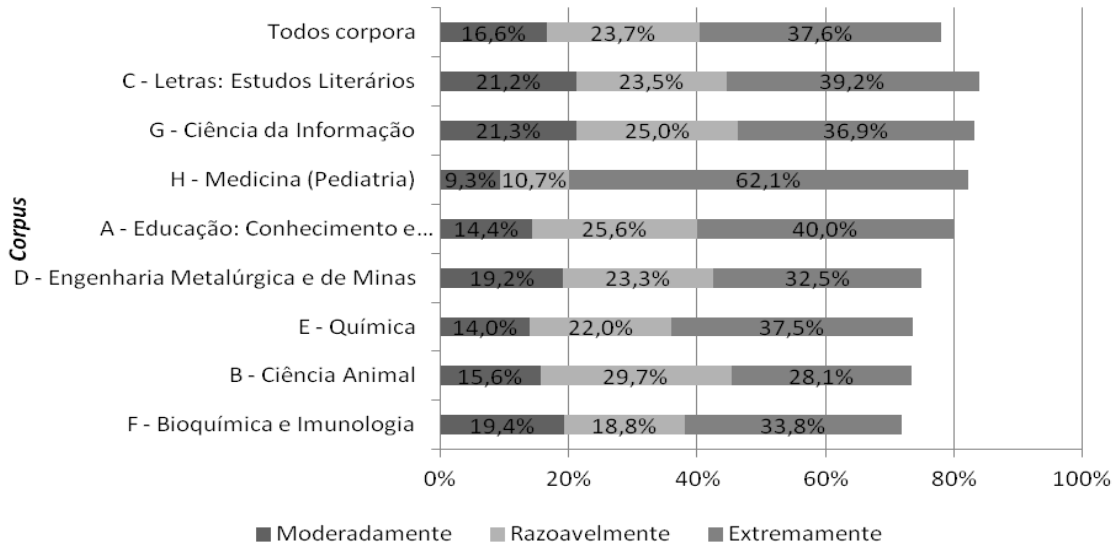
discussão, segundo os autores, tradução da autora, tradução nossa, valores em negrito, valores mostrados, valores teóricos em parêntesis;

- E outros: associadas, gravada, maioria das vezes, maioria dos docentes, maioria dos valores, média de idade, média de três experimentos independentes, média de um experimento realizado em triplicata, média dos valores obtidos, média percentual de triplicatas das células tratadas em relação, médio com comprimento de quadro igual, médio do título de antitoxina, seguintes, etc.

Para o levantamento dos exemplos agora citados, assim como sua classificação, foram listados todos os 422 (22,1% do total) candidatos a descritores que receberam a avaliação como não relevante, em seguida, para cada um deles foi atribuída uma propriedade linguística. Consideraram-se aqui somente os maiores grupos dessas propriedades linguísticas.

Através de comparação com resumos e palavras-chaves dos artigos em Ciência da Informação que utilizou, Souza (2005, p. 132) avaliou que 88,9% dos SNs candidatos eram relevantes como descritores. Já nesta pesquisa, somente para o programa de pós-graduação em Ciência da Informação, o valor encontrado foi relativamente próximo, 83,1%, conforme pôde ser visto na Tabela 17 anterior. Dentre os demais programas, esse foi o segundo melhor valor encontrado, sendo que a seção do *corpus* do programa de pós-graduação em Letras – Estudos Literários apresentou o melhor resultado: 83,8%. O Gráfico 15 apresenta essa ordem de melhores resultados, assim como a distribuição entre os níveis de relevância considerados aqui.

Gráfico 15 - Avaliação de níveis de relevância por seção do corpus



Fonte: Elaborado pelo autor.

A distinção entre as ciências naturais e as sociais, assim como na média da quantidade de SNs de cada seção do *corpus* apresentada no Gráfico 7 da página 69, pode ser percebida com uma tendência para piores resultados nas áreas das ciências naturais, exceto na seção H referente à área de Medicina (Pediatria).

A seção do *corpus* referente ao programa de pós-graduação em Medicina (Pediatria) apresentou valores como extremamente relevante (62,1%) muito acima dos demais. Conforme pode ser obervada nas respostas dos autores desse grupo no APÊNDICE I, um deles avaliou todos os candidatos como extremamente relevantes. Uma vez que esta seção é a menor (somente 7 teses), somente um questionário impactou de forma considerável no resultado dos demais em conjunto. Foi possível ainda perceber, na mesma seção, a avaliação como extremamente relevante de alguns candidatos como: pessoas doentes, área da saúde, humano, alunos, realização de procedimentos, revisão da literatura, corpo, espelho, imagem, observada diferença, saúde perfeita, total de pacientes e através do fio.

O enunciado do questionário enviado aos autores, que se encontra no APÊNDICE F, continha o seguinte texto: “Para cada SN abaixo determine o grau de relevância do mesmo como descritor de sua tese”. Em virtude das respostas encontradas aqui, foi possível perceber que alguns autores consideraram a relevância do descritor para o seu discurso empregado na sua própria tese. Ou seja, consideraram se o descritor era coerente com seu próprio texto, se ele emergia do mesmo. De fato, isso ocorre para todos os descritores candidatos, uma vez que foram selecionados principalmente em virtude da sua maior frequência de ocorrência no mesmo. Nota-se então que, para alguns autores, o conceito de descritor como um identificador que o diferencie de outras obras, sobretudo no mesmo programa de pós-graduação, foi pouco considerado.

Na Equação 2, há dois fatores que são coerentes com esses dois tipos de visão de um descritor: a frequência f_{ij}, como a visão do descritor que emerge do texto; e n_i como a visão do descritor de fora do texto, ou seja, em relação a todos os textos da mesma seção do *corpus*. Este último, como visto no Gráfico 11, foi pouco relevante para a ordem dos candidatos selecionados, uma vez que preponderou a frequência do SN no mesmo texto. Somado a esse viés de visão a partir do próprio texto, para uma amostra relativamente pequena, como a do programa de pós-graduação em Medicina (Pediatria), termos comuns da área, como os exemplificados (área da saúde, corpo, humano, etc.), acabaram sendo eleitos como candidatos.

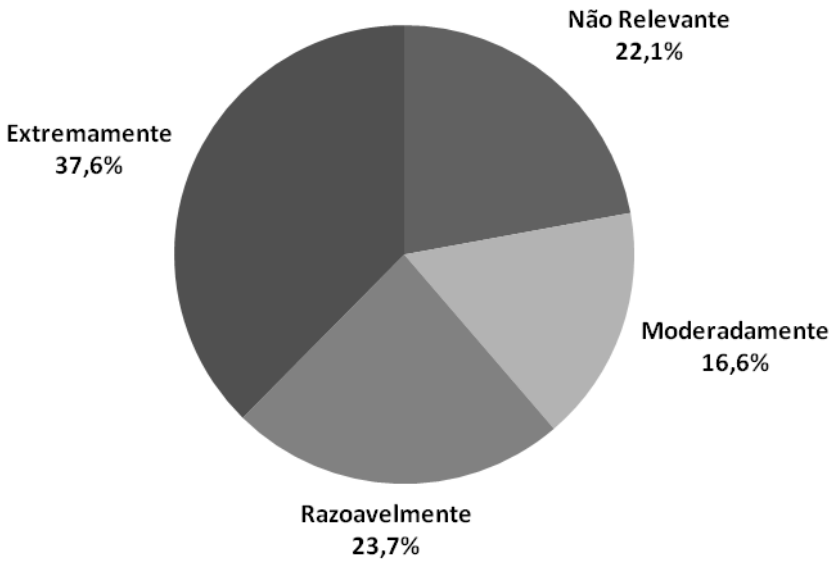
Pode-se resumir aqui então que:

a. na Equação 2 o fator n_i, que serve para minimizar a seleção de *stopwords* de uma área do conhecimento no *corpus*, necessita de maior ponderação diante do fator f_{ij};

b. que amostras pequenas favorecem à seleção de descritores comuns à área, e;
c. que o enunciado apresentado aos autores no questionário deu margem à interpretação do termo descritor como aquele que é mais referente a aspectos internos de uma tese e menos a aspectos que envolvem distintos assuntos da área como um todo.

De um modo geral, a avaliação da relevância dos candidatos selecionados foi positiva, não somente pela quantidade total de 77,9%, como também por apresentar uma ordem crescente do menor nível (moderadamente com 16,6%) para o maior nível de relevância (extremamente com 37,6%), como é apresentado no Gráfico 16 a seguir.

Gráfico 16 - Avaliação total de níveis de relevância



Fonte: Elaborado pelo autor.

Os SNs candidatos de cada tese foram ordenados de acordo com a pontuação obtida pela Equação 2 e então selecionados somente os seus vinte primeiros. Como já descrito também no capítulo de metodologia, os autores receberam os questionários com os candidatos ordenados alfabeticamente, e não pela ordem de pontuação obtida para a seleção. As respostas dos autores foram agrupadas de duas em duas (1 e 2 para moderadamente, 3 e 4 para razoavelmente, e 5 e 6 para extremamente) de modo a formar três níveis com valores associados respectivamente a 0,25, 0,50 e 1,00, como já apresentado na Tabela 8, na página 62. A seguir, na Tabela 18 são apresentadas as médias desses valores associados de relevância para cada ordem do SN candidato (do primeiro ao vigésimo) distribuídos por cada seção do *corpus*.

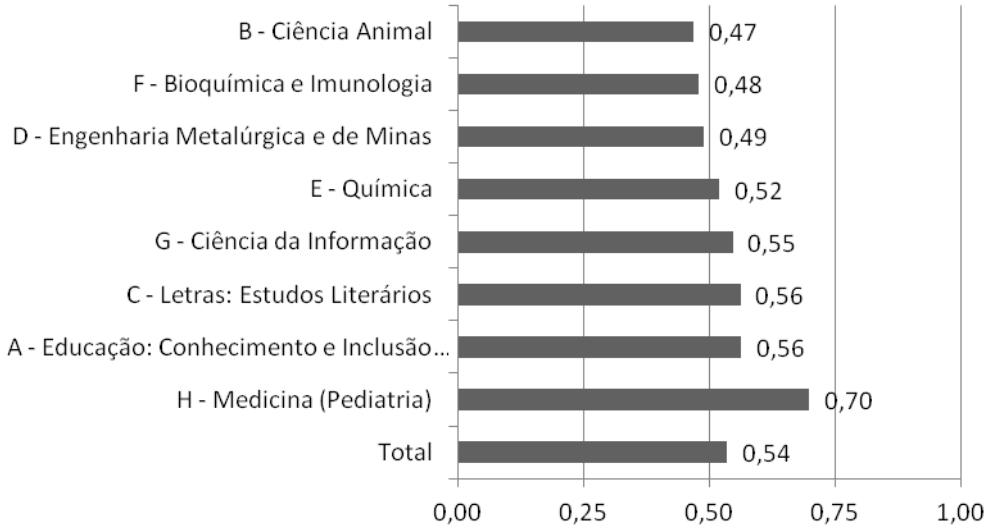
Tabela 18 - Valor associado médio de relevância por ordem dos candidatos a descritor

Ordem	Valor de relevância por ordem dos candidatos a descritor em cada seção do *Corpus*								Total
	A	B	C	D	E	F	G	H	
1	0,44	0,59	0,85	0,54	0,48	0,47	0,63	0,86	**0,58**
2	0,69	0,59	0,75	0,50	0,65	0,69	0,75	0,75	**0,66**
3	0,54	0,53	0,58	0,44	0,58	0,75	0,72	0,75	**0,58**
4	0,61	0,56	0,73	0,63	0,65	0,53	0,81	0,71	**0,64**
5	0,75	0,50	0,56	0,48	0,58	0,66	0,59	0,54	**0,60**
6	0,60	0,41	0,56	0,35	0,63	0,63	0,66	0,54	**0,54**
7	0,47	0,39	0,58	0,48	0,45	0,41	0,47	0,54	**0,47**
8	0,57	0,64	0,56	0,65	0,60	0,75	0,59	0,86	**0,63**
9	0,71	0,47	0,50	0,65	0,40	0,44	0,56	0,71	**0,57**
10	0,45	0,39	0,52	0,63	0,45	0,41	0,53	1,00	**0,51**
11	0,54	0,41	0,60	0,44	0,53	0,63	0,50	0,86	**0,54**
12	0,50	0,34	0,50	0,31	0,65	0,38	0,53	0,71	**0,47**
13	0,43	0,70	0,52	0,42	0,68	0,44	0,63	0,43	**0,53**
14	0,47	0,42	0,63	0,42	0,40	0,38	0,50	0,68	**0,48**
15	0,51	0,39	0,54	0,40	0,35	0,31	0,47	0,79	**0,46**
16	0,66	0,39	0,62	0,50	0,38	0,38	0,53	0,75	**0,53**
17	0,48	0,53	0,29	0,65	0,45	0,28	0,38	0,61	**0,46**
18	0,65	0,34	0,46	0,42	0,45	0,34	0,31	0,61	**0,47**
19	0,61	0,42	0,46	0,31	0,48	0,56	0,59	0,57	**0,50**
20	0,60	0,34	0,46	0,60	0,60	0,19	0,19	0,71	**0,48**
Total	**0,56**	**0,47**	**0,56**	**0,49**	**0,52**	**0,48**	**0,55**	**0,70**	0,54

Fonte: Elaborado pelo autor.

Como já foi analisado mais profundamente, o programa de pós-graduação em Medicina (Pediatria) apresentou a melhor média de avaliação dos candidatos a descritores. Embora os demais programas tenham apresentado um média bem próxima da total de 0,54, novamente é possível perceber um agrupamento entre os programas relacionados às ciências sociais, com uma melhor avaliação dos candidatos, e os das ciências naturais, com uma pior avaliação. O Gráfico 17 apresenta a posição de cada seção do *corpus* em relação à sua média total de avaliação dos candidatos a descritores.

Gráfico 17 - Média de valor associado à relevância dos candidatos a descritores por seção do *corpus*



Fonte: Elaborado pelo autor.

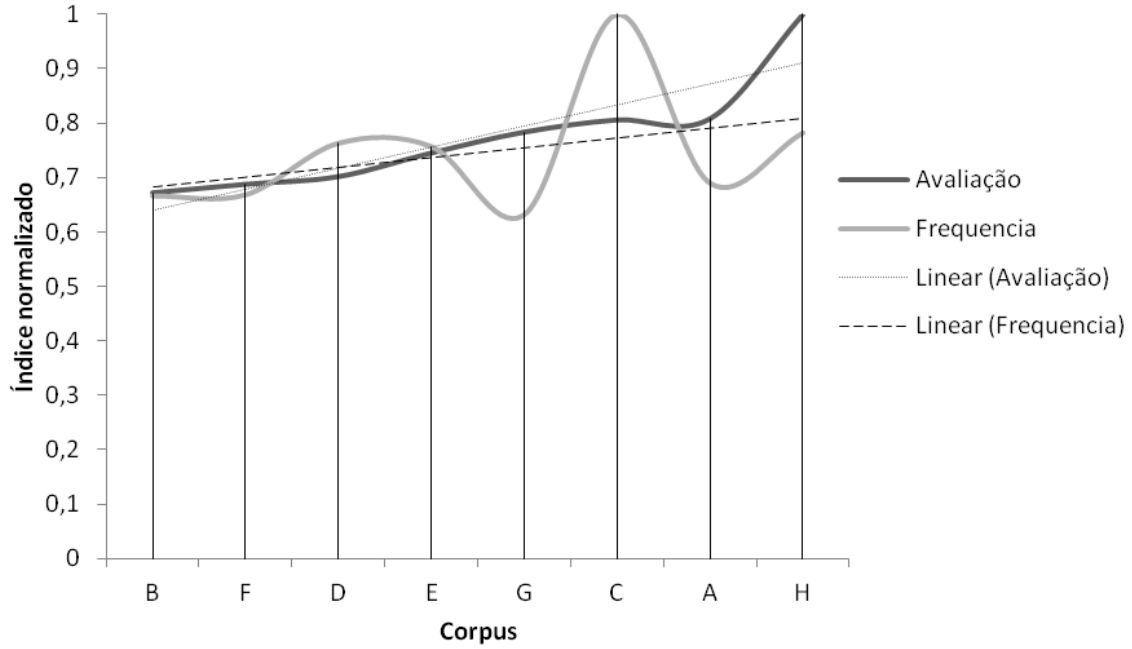
O fato das ciências naturais apresentarem uma pior avaliação dos candidatos a descritores pode ter como causa duas hipóteses levantadas aqui:

a. A linguagem das ciências naturais, por assemelhar-se mais a uma linguagem lógica matemática, dificulta o processamento empregado aqui por distanciar-se mais da linguagem natural, como já foi analisado aqui neste capítulo em relação aos erros de extração;
b. O comportamento linguístico nas ciências naturais teria uma tendência a empregar menos repetições[44] de um mesmo sintagma nominal em seu discurso, sendo que isso dificultaria a eleição de melhores candidatos.

Para esta última hipótese, foram confrontados os dados obtidos para os candidatos considerados relevantes. Analisou-se a média de frequência desses descritores nas próprias teses e a média de valor associado à avaliação do autor. Para permitir uma análise visual, os dados foram normalizados para o máximo encontrado em cada um de acordo com o Gráfico 18 a seguir.

[44] Na Linguística de *Corpus* existe uma medida semelhante denominada *Token/Type Ratio*.

Gráfico 18 - Análise da relação frequência versus relevância entre as seções do *corpus*



Fonte: Elaborado pelo autor.

Considerando-se a mesma ordem das seções do *corpus* por média de valor associado à relevância (iniciando em B e terminando em H), como no Gráfico 17, foram normalizados tais valores de avaliação de relevância assim como os de frequência média dos mesmos descritores eleitos[45]. Com a normalização foi possível confrontar o comportamento de ambos os dados e comprovar a última hipótese. Para isso, foi utilizada a regressão linear de ambos os dados, encontrando-se que o valor associado à relevância e a frequência do descritor tendem a crescer juntos.

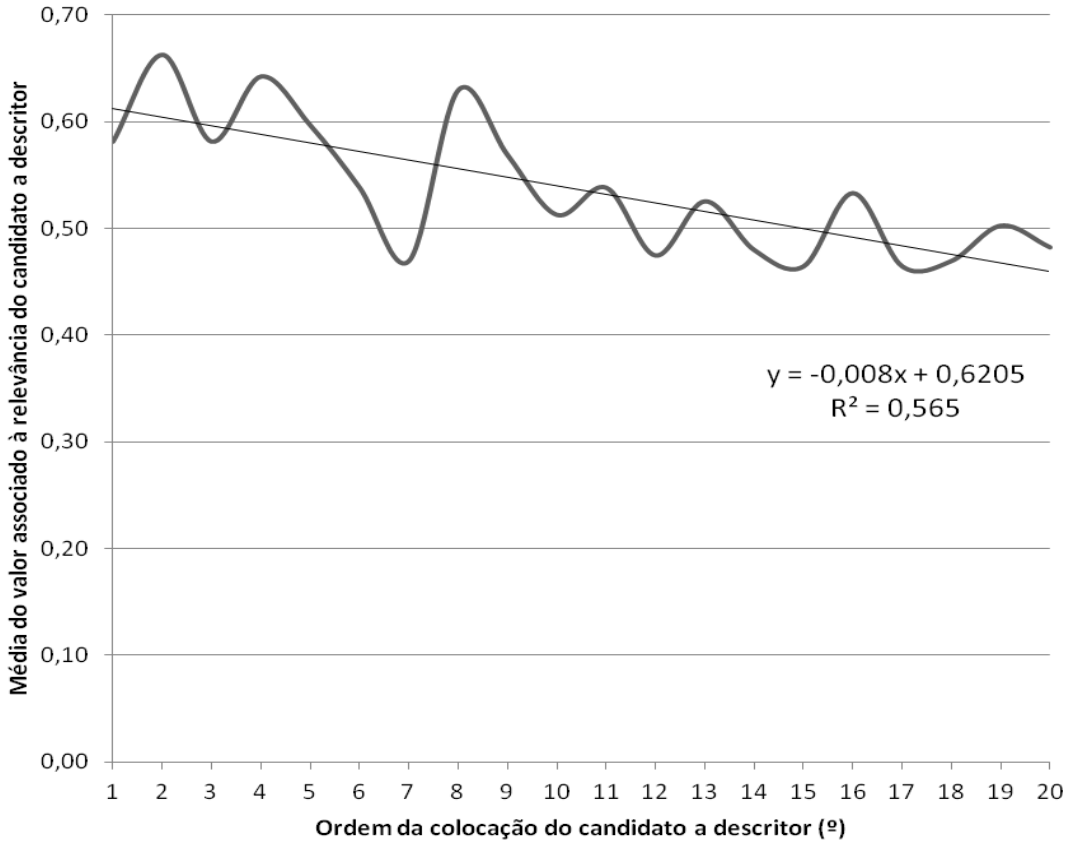
Podemos inferir que a metodologia empregada aqui para a eleição de candidatos a descritores a partir de SNs tende a ser mais eficiente quanto maior for a possibilidade de repetições desses sintagmas ao longo do texto.

Voltando à Tabela 18, com sua associação entre a ordenação obtida pela pontuação para a seleção do candidato a descritor (de 1 a 20) e a média do valor associado à relevância atribuída pelos autores (podendo variar de 0 a 1), esperou-se encontrar um comportamento exponencial de decaimento no valor dessa relevância para cada colocação assim como pôde ser visto no Gráfico 13. No entanto para todas as seções do *corpus* esse

[45] Para uma maior precisão seria necessário o cálculo de tal média em todos os sintagmas nominais identificados. Porém, considerou-se suficiente aqui, como amostra para fins de comparação, os próprios descritores eleitos.

decaimento ficou mais próximo de uma função linear decrescente, como pode ser observado no Gráfico 19 a seguir.

Gráfico 19 - Média de valor de relevância por colocação do candidato a descritor



Fonte: Elaborado pelo autor.

O decaimento do valor associado à relevância do descritor em relação à sua colocação obtida a partir da Equação 2 demonstra que a metodologia empregada aqui foi coerente com a visão dos autores quando compararam um candidato a descritor mais relevante que o outro. Considerando esse decaimento de forma linear, tal qual apresentado no Gráfico 19 e reduzido à Equação 3 a seguir:

Equação 3 - Relação entre avaliação de relevância e colocação do candidato a descritor

$$Avaliação = -0{,}008 \bullet Colocação + 0{,}6205$$

Fonte: Elaborado pelo autor

O corte dos SNs candidatos a descritores melhores pontuados, ao invés do valor de vinte, que foi adotado nesta pesquisa, deveria ser de 78 para valores associados à

relevância maiores que zero. Ou seja, os questionários enviados aos autores deveriam ter aproximadamente 4 vezes mais candidatos a descritores. Caso fosse adotada essa estimativa, haveria o risco de uma menor adesão à pesquisa. No entanto, devido à facilidade para se responder ao questionário relatada por alguns autores, é possível considerar essa dimensão sem prejuízo de adesão.

De acordo com a Equação 3, é possível fazer uma estimativa do valor de corte médio dos candidatos a descritores por objetivo mínimo de relevância, sendo que este foi calculado em função de seu valor associado de relevância imediatamente inferior. A seguir, na Tabela 19, apresenta-se para cada seção do *corpus,* os coeficientes (*a* e *b*) de modo a determinar a sua equação específica, tal como a Equação 3. O coeficiente R^2 também é apresentado e determina a % de variabilidade que pode ser previsível pela equação (LEVINE; BERENSON; STEPHAN, 2000). As quantidades estimadas de candidatos para poder determinar de forma mínima cada um dos três níveis de relevância são apresentadas a seguir.

Tabela 19 - Quantidade estimada de candidatos por objetivo mínimo de relevância

Seção do *corpus*	Relevância = a.Candidatos + b			Qtd. Candidatos por objetivo mínimo de relevância		
	a	b	R^2	Extremamente	Razoavelmente	Moderadamente
A - Educação: Conhec. Inc. Soc.	-0,0011	0,5752	0%	68	296	**523**
B - Ciência Animal	-0,0091	0,5640	25%	7	35	**62**
C - Letras: Estudos Literários	-0,0142	0,7120	50%	15	33	**50**
D - Engenharia Metal. e Minas	-0,0032	0,5228	3%	7	85	**163**
E - Química	-0,0067	0,5903	14%	13	51	**88**
F - Bioquímica e Imunologia	-0,0184	0,6729	46%	9	23	**37**
G - Ciência da Informação	-0,0187	0,7433	58%	13	26	**40**
H - Medicina (Pediatria)	-0,0045	0,7453	4%	55	110	**166**
Total	**-0,0080**	**0,6205**	**57%**	**15**	**46**	**78**

Fonte: Elaborado pelo autor

Dentre todos as seções, somente o correspondente ao programa de pós-graduação em Letras – Estudos Literários e ao de Ciência da Informação obtiveram uma regressão linear com um fator de variabilidade que considera ao menos metade dos valores

encontrados como possíveis de serem previstos pela equação. Isso demonstra que, nesses dois grupos, houve uma maior coerência linear no decaimento da relevância com a ordem do candidato selecionado, embora em todos os demais tenha-se encontrado, de todos os modos, um grau desse decaimento apontado pelo índice negativo do coeficiente *a*.

Analisando ainda o coeficiente *a*, que denota também o quão rápido ocorre o decaimento da relevância em função da colocação do candidato descritor, o programa de Bioquímica e Imunologia também apresentou, assim como os dois programas citados anteriormente, um forte decaimento. Podemos apontar que, nestes três, o conjunto de descritores tende a ser mais reduzido. A causa dessa menor necessidade de descritores nessas seções merece uma atenção mais linguística que a quantitativa empregada aqui, uma vez que, para tais seções, não foi encontrada semelhança de pertencimento às ciências sociais/naturais, no volume de SNs extraídos/identificados, na frequência média dos candidatos ou outros fatores.

A seguir são analisadas as distribuições das relevâncias ao longo dos textos, de modo que haverá mais possibilidades, dentre outros objetivos, de associarmos uma possível justificativa para uma necessidade menor de descritores nesses três programas citados.

4.5 Análise da distribuição da relevância dos descritores em posições do texto

Nos itens anteriores, foram analisadas todas as etapas para obtenção da relevância dos descritores extraídos de forma automática das teses. Neste item, são analisadas as distribuições das relevâncias desses descritores em posições ao longo do texto.

Em relação às possíveis posições que um SN pode ocorrer em um texto, foram apresentadas, no item 4.2 deste capítulo, a dimensão média da quantidade de SNs extraídos, apresentada no Gráfico 7 na página 69, e a distribuição média da quantidade de SNS encontrada por partes estruturais introdução/desenvolvimento/conclusão, apresentado no Gráfico 8 na página 73, ambas com detalhes para cada seção do *corpus*.

Como objetivo principal dessa dissertação, buscou-se encontrar um comportamento na distribuição da relevância dos descritores ao longo do texto. Isso permitiria, por exemplo, um olhar mais direcionado, seja por um indexador manual ou automático, para partes específicas do texto, além das comumente empregadas estruturas, tais como títulos, resumos, palavras-chaves e outras.

A análise da posição no texto empregada aqui é puramente referente à sequência dos seus SNs, desconsiderando-se outras estruturas linguísticas que não estas, e às partes já ditas: introdução/desenvolvimento/conclusão. Foram considerados os SNs

máximos encontrados, ou seja, os de maior categoria cujos níveis são detalhados na Tabela 6 na página 58. Foram desconsiderados os SNs aninhados dentro desses e de menor categoria. E, como também já descrito aqui no capítulo da metodologia, foram descartadas todas as partes pré-textuais (capa, contracapa, resumos etc.) e pós-textuais (referências, anexos, apêndices e etc.).

Para a análise da distribuição da relevância dos SNs como descritores, idealmente poderíamos considerar que um único indexador avaliasse a relevância de cada um dos 533.164 SNs identificados nas 98 teses. Tal prática seria inviável, por três motivos: a dimensão, a diversidade de áreas de conhecimento e a variabilidade de critérios empregados no decorrer desse longo processo, mesmo que feito por uma só pessoa.

Consideramos então aqui dois recortes para a distribuição da relevância:

a. Cada autor teve o mesmo poder de atribuição de relevância para a sua própria tese (desconsiderando-se os candidatos não relevantes);
b. A quantidade de descritores relevantes para cada tese foi suficiente como amostra para analisar a distribuição da relevância em cada texto.

A primeira consideração acima faz com que todos os autores, desde o mais severo, que avaliou poucos candidatos como relevantes, ao mais benévolo, que atribuiu a todos como extremamente relevantes, sejam considerados igualmente. A atuação deles finalizaria então todo o processo de determinação dos SNs como descritores, e a relevância atribuída por eles faria parte de um peso total único (igual para cada autor) repartido dentre tais descritores que eles avaliaram relevantes.

Quanto à segunda consideração acima, embora tenha sido concluído no item anterior que a média de candidatos devesse ser 4 vezes maior que a de 20 enviada aos autores, este número foi suficiente para cobrir, como mínimo, todos os descritores extremamente relevantes (ver Tabela 19), sendo considerados estes como amostra suficiente (de 25%) do total de possíveis descritores.

Nos subitens a seguir, inicialmente é apresentada uma análise de distribuição das relevâncias considerando-se a sequência linear (divididas em 10 partes) dos SNs e, em seguida, é finalizada a análise deste capítulo considerando-se as densidades de relevância dos SNs em cada uma de suas três partes textuais: introdução, desenvolvimento e conclusão.

4.5.1 Análise da distribuição da relevância no texto dividido em 10 partes iguais

Como descrito no item 1.8 da metodologia, na página 61, para cada tese, o valor de relevância atribuído a cada descritor (0,25, 0,50 ou 1,00) foi dividido igualmente entre cada uma de suas ocorrências em suas respectivas posições (mensuradas em % em

relação à quantidade total de SNs extraídos). Em seguida, tais posições, com suas respectivas frações de valores de relevâncias já espalhadas ao longo do texto, foram agrupadas e somadas a cada 10% do texto, conforme é apresentado no APÊNDICE J. Uma vez que cada autor atribuiu um total diferente de relevâncias, como já mencionado aqui, esses valores foram normalizados para 100% em cada tese um função do seu respectivo total.

Uma vez que cada uma das 10 partes de cada tese teve seu valor normalizado, suas porcentagens foram consolidadas com peso igual. A Tabela 20 apresenta os dados consolidados e detalhados por seção do *corpus*.

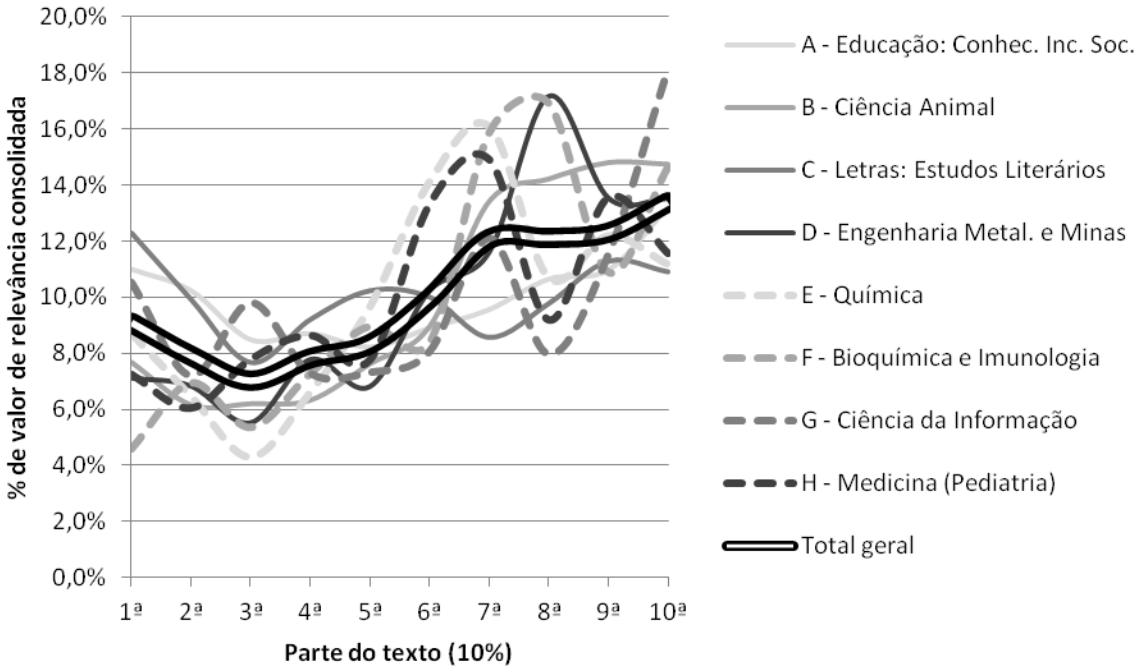
Tabela 20 - Distribuição dos valores de relevância em 10 partes nas teses do *corpus*

Seção do *corpus*	Parte de cada tese (10%)									
	1ª	**2ª**	**3ª**	**4ª**	**5ª**	**6ª**	**7ª**	**8ª**	**9ª**	**10ª**
A - Educação: Conhec. Inc. Soc.	11,0%	10,2%	8,5%	8,7%	8,2%	8,9%	9,6%	10,7%	11,0%	13,2%
B - Ciência Animal	7,7%	6,2%	6,2%	6,3%	7,6%	8,9%	13,4%	14,2%	14,8%	14,7%
C - Letras: Estudos Literários	12,3%	9,9%	7,7%	9,2%	10,2%	10,0%	8,6%	9,8%	11,3%	10,9%
D - Engenharia Metal. e Minas	7,1%	6,8%	5,5%	7,8%	6,8%	10,2%	11,6%	17,1%	13,5%	13,5%
E - Química	8,7%	6,5%	4,3%	6,6%	9,7%	14,1%	16,1%	10,7%	12,2%	11,2%
F - Bioquímica e Imunologia	4,6%	7,0%	5,4%	7,3%	9,0%	8,4%	15,9%	16,9%	10,9%	14,7%
G - Ciência da Informação	10,5%	7,2%	9,8%	7,3%	7,3%	8,1%	12,1%	8,0%	11,5%	18,1%
H - Medicina (Pediatria)	7,2%	6,1%	7,8%	8,6%	7,8%	13,3%	14,9%	9,2%	13,5%	11,6%
Total geral	**9,1%**	**7,9%**	**7,0%**	**7,8%**	**8,3%**	**10,0%**	**12,1%**	**12,1%**	**12,3%**	**13,4%**
Ciências Naturais	7,2%	6,5%	5,8%	7,2%	8,0%	10,7%	14,1%	14,0%	13,3%	13,4%
Ciências Sociais	11,3%	9,6%	8,5%	8,6%	8,7%	9,1%	9,7%	9,9%	11,2%	13,4%

Fonte: Elaborado pelo autor

O principal objetivo desta pesquisa toma forma inicial através desses resultados apresentados na Tabela 20 que podem ser visualizados no Gráfico 20 e no Gráfico 21 a seguir.

Gráfico 20 - Distribuição dos valores de relevância em 10 partes nas teses do *corpus*

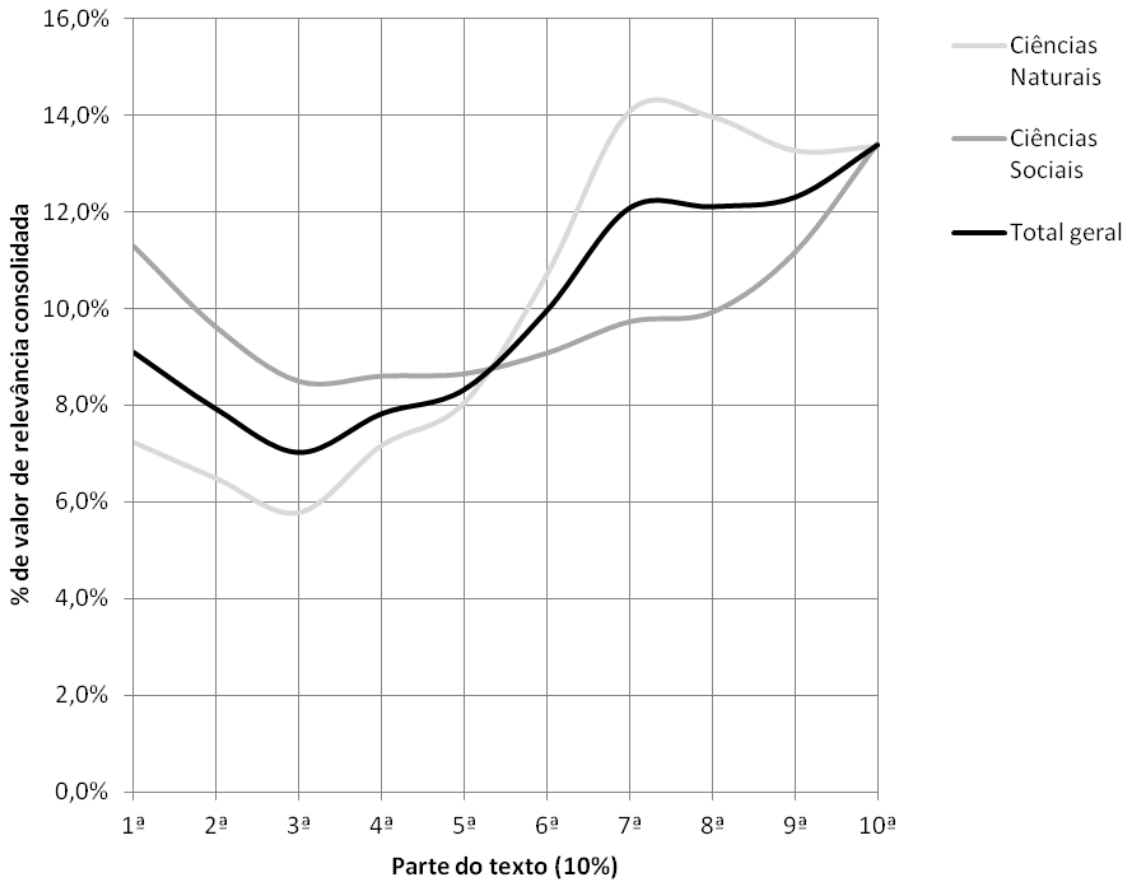


Fonte: Elaborado pelo autor

O comportamento da distribuição dos valores das relevâncias, embora possa parecer caótico nesse primeiro momento, apresenta um padrão geral que se inicia em torno dos 9%, abaixa até 7% na 3ª parte do texto e depois sobe constantemente até atingir mais de 13% no final. Entre o mínimo e o máximo geral, para as 10 partes, houve uma variação de cerca do dobro de valor de relevância.

Como já foi apresentado em outras partes da análise de dados neste capítulo, existe uma distinção entre as teses dos programas mais relacionados às ciências sociais e as relacionadas às ciências naturais. A seguir, o Gráfico 21 apresenta os valores consolidados para esses dois grandes grupos.

Gráfico 21 - Distribuição dos valores de relevância em 10 partes nas teses das ciências naturais e das ciências sociais



Fonte: Elaborado pelo autor

O comportamento da distribuição da relevância dos descritores ao longo de um texto relacionado às ciências sociais possui mais harmonia (somente com um momento de inflexão[46]) e menor variação entre seus extremos (cerca de uma vez e meia). Assim como o comportamento geral do *corpus*, ele inicia medianamente, desce até a 3ª parte e sobe continuamente até o final, chegando a 13,4% também. Já as teses pertencentes às ciências naturais possuem maior oscilação (com três momentos de inflexão) e uma maior variação entre seus extremos (cerca de duas vezes e meia). Suas partes iniciais apresentam baixos valores de relevância e, a partir da segunda metade do texto, atingem valores altos (acima dos 10%) chegando a 14%.

Embora Santos (1996) acredite que a dicotomia entre as ciências naturais e sociais tenha deixado de ter sentido, a análise da distribuição do valor da relevância dos descritores revela mais uma diferença entre esses dois conjuntos. Já foi analisada aqui a diferença entre tais conjuntos relativa ao tamanho médio de SNs extraídos das teses

[46] Inflexão é considerada aqui como o momento que uma sequência muda entre crescente e decrescente.

(Gráfico 7, ver na página 69). Tais diferenças podem ainda estar associadas a comportamentos linguísticos decorrentes da afirmativa de que, nas ciências sociais, não há consenso paradigmático (KUHN citado por SANTOS, 1996, p. 37). As ciências naturais, por sua vez, como já analisadas mais especificamente no programa de pós-graduação em Química, podem chegar a possuir um vocabulário controlado adotado internacionalmente.

A média da relevância para cada uma das 10 partes seria 10%, logo podemos analisar que cada parte tem menor ou maior relevância quanto mais distante seja seu valor em relação a essa média. O grupo relativo às ciências sociais apresenta uma distribuição da relevância mais coerente com a distribuição esperada em textos científicos, que concentra seus termos relevantes no início (1ª parte de 10) e no final do mesmo (9ª e 10ª partes). Já o comportamento das teses relativas às ciências naturais começa a apresentar maiores relevâncias somente a partir da metade de seus textos, atingindo seus máximos nas quatro partes finais.

A 3ª parte de ambos os grupos, ciências naturais e sociais, apresentou a menor taxa de relevância de descritores. Como vimos no Gráfico 8 da página 73, a terceira parte de um texto corresponde na média das teses analisadas ao início do desenvolvimento (que fica aproximadamente entre 10% e 90% do texto). Sem a pretensão de ser compravada aqui nesta pesquisa, pode-se lançar uma hipótese para esse fato. No início do desenvolvimento, a parte mais comumemente encontrada (através de exploração em algumas teses) é relativa ao referencial teórico. Como nesta parte, em alguns casos, são apresentadas questões mais gerais das suas respectivas áreas de pesquisa, podemos avaliar aqui que seus termos seriam pouco relevantes para a descrição da tese dentro do próprio programa de pesquisa. Tal suposição é coerente com o fato de um dos fatores de pontuação para a escolha dos candidatos ter sido, justamente, a ausência do SN nos demais documentos da mesma seção.

As ciências naturais, por possuírem então maior consenso paradigmático, apresentam maior probabilidade de semelhança de uso dos mesmos SNs em partes do texto que se referem aos seus aspectos conceituais gerais, como nas partes relativas a referenciais teóricos. Como essa maior homogeneidade favorece que seus respectivos termos sejam piores descritores, a taxa de valor de relevância tende a se concentrar nas posições do texto que tratam mais especificamente do assunto de cada tese. No Gráfico 21, o maior acúmulo do valor de relevância para as teses das ciências naturais ocorre na 7ª e 8ª partes. Como afirma Feltrim, Aluisio e Nunes (2000), podemos considerar que tais partes sejam relativas às seções de métodos e resultados, uma vez que estão um pouco antes da conclusão (que se inicia em média nos 92%, como apresentado aqui) e suficientemente distantes do início do desenvolvimento (em torno dos 10%), quando há maior chance de se encontrar a seção relativa ao referencial teórico. Seria necessária uma demarcação não

somente das partes relativas à introdução, ao desenvolvimento e à conclusão, como o feito aqui, mas também de tais seções. No entanto a denominação de tais seções, assim como a utilização das mesmas, varia muito de autor para autor, o que requer uma análise mais subjetiva do que a metodologia empregada aqui busca ser.

A última parte pode ser considerada a de maior taxa de relevância de descritores para o conjunto de todas as teses. Embora nas teses relativas às ciências naturais esta parte tenha sido a 7ª, a sua diferença para a última foi relativamente pequena. A seção de conclusão, como já apresentado aqui também no Gráfico 8, correspondeu, em média, à faixa dos 93% em diante no conjunto de todas as teses. Portanto, a parte do texto que mais possui descritores relevantes seria a conclusão, como será analisado no próximo subitem.

Análise da distribuição do valor de relevância na introdução, desenvolvimento e conclusão

Como apresentado no capítulo da metodologia, foram utilizados dois demarcadores para dividir cada tese em três partes: introdução, desenvolvimento e conclusão. Logo, a posição de cada SN pôde ser associada a cada uma delas. No entanto, diferentemente da divisão em 10 partes iguais apresentada no subitem anterior, aqui, cada parte apresentou um tamanho diferente das demais. Logo, a quantidade total dos valores de relevância foi ponderada para o tamanho de cada parte em número de SNs extraídos. Essa média de valores para cada parte estrutural de cada tese é apresentada no APÊNDICE L.

Os valores apresentados a seguir podem ser considerados como relativos à densidade de valor de relevância de cada parte, ou seja, com eles é possível analisar qual parte tem maior probabilidade que um de seus SNs seja um descritor relevante de toda a tese.

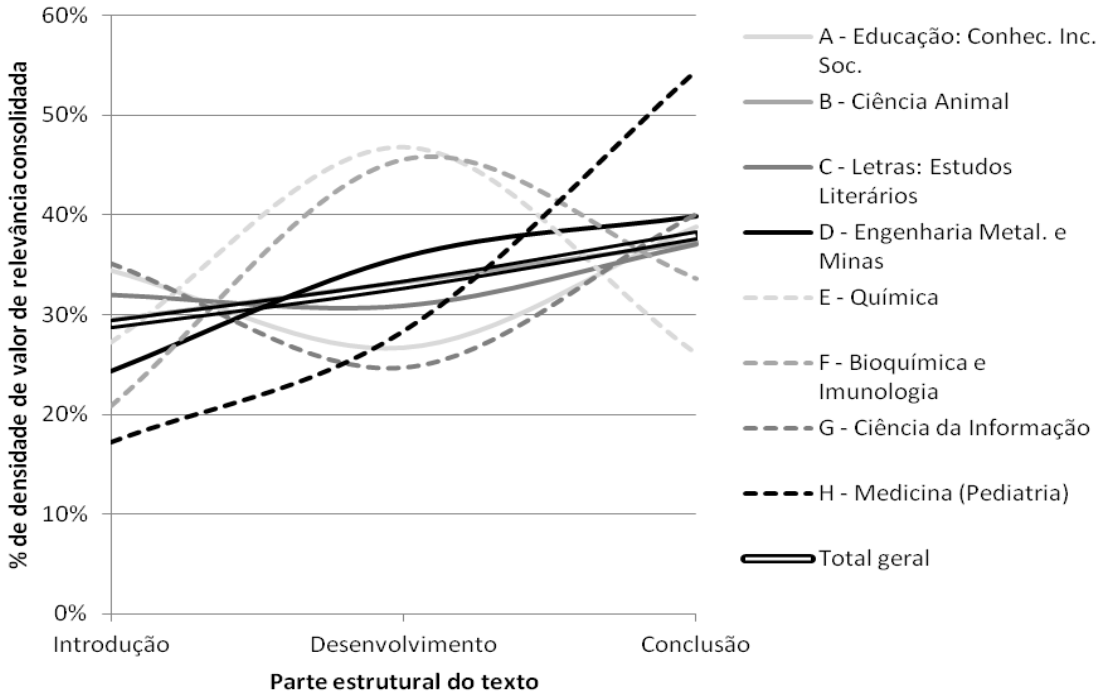
Tabela 21 - Distribuição dos valores da densidade de relevância dos sintagmas nominais por partes estruturais nas teses do corpus

Seção do *corpus*	Parte Estrutural		
	Introdução	**Desenvolvimento**	**Conclusão**
A - Educação: Conhec. Inc. Soc.	34,5%	26,7%	38,8%
B - Ciência Animal	29,4%	33,3%	37,3%
C - Letras: Estudos Literários	32,0%	30,9%	37,1%
D - Engenharia Metal. e Minas	24,3%	35,8%	39,9%
E - Química	27,2%	46,8%	26,1%
F - Bioquímica e Imunologia	20,8%	45,6%	33,6%
G - Ciência da Informação	35,2%	24,7%	40,1%
H - Medicina (Pediatria)	17,2%	28,3%	54,5%
Total geral	**29,0%**	**33,0%**	**38,0%**
Ciências Naturais	**24,9%**	**37,6%**	**37,5%**
Ciências Sociais	**33,9%**	**27,6%**	**38,6%**

Fonte: Elaborado pelo autor

Os dados da tabela anterior podem ser analisados no Gráfico 22 e no Gráfico 23 a seguir.

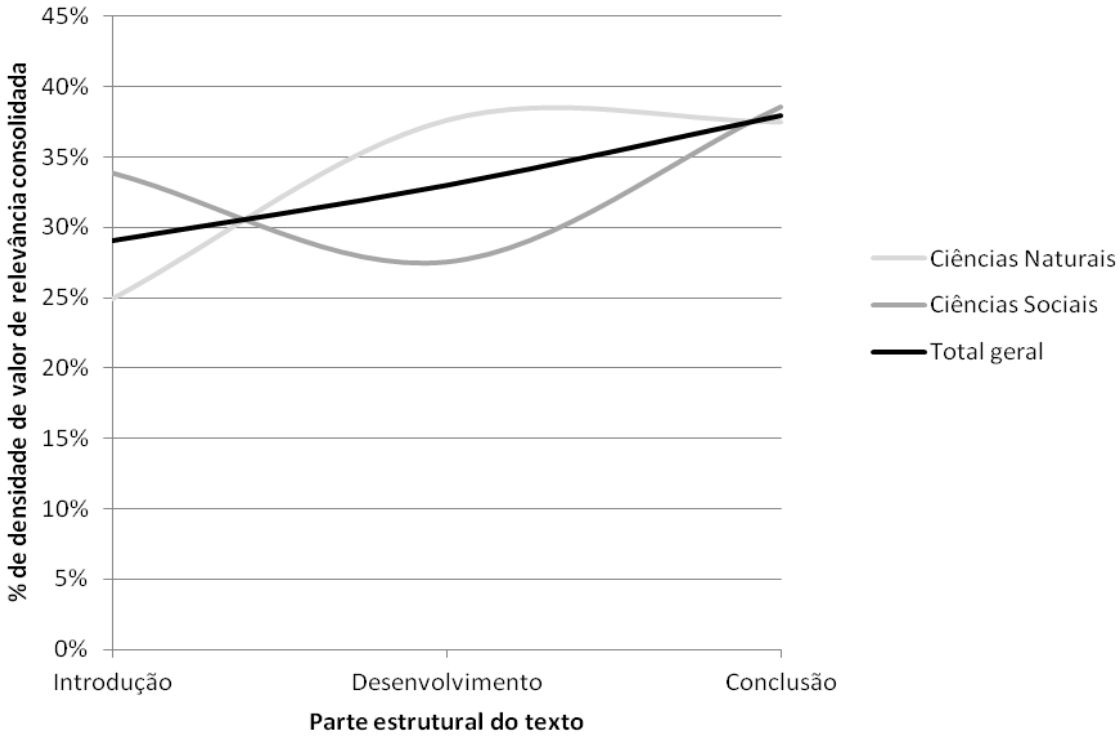
Gráfico 22 - Distribuição dos valores da densidade de relevância dos sintagmas nominais por partes estruturais nas teses do *corpus*



Fonte: Elaborado pelo autor

Na análise feita anteriormente, considerando uma divisão em 10 partes iguais (sem associação com partes estruturais), foi encontrada, no comportamento geral, uma inflexão para baixo na 3ª parte. Na análise por partes estruturais, tal comportamento, que considera todas as teses como um único conjunto geral, não apresentou inflexão. Além de diferir da análise anterior, tal resultado também difere do comportamento esperado de um texto científico (que tende a concentrar informações nas partes iniciais e finais). No entanto, é possível analisar que esse comportamento encontrado é resultante de dois comportamentos distintos entre as teses relativas às ciências naturais e às ciências sociais, como é apresentado no Gráfico 23, a seguir.

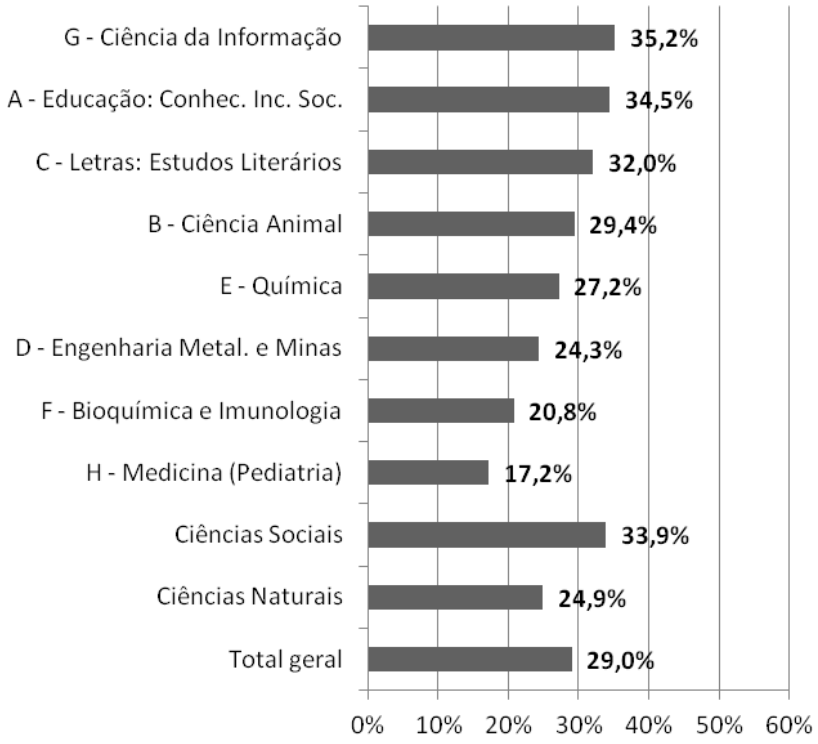
Gráfico 23 - Distribuição dos valores da densidade de relevância dos sintagmas nominais por partes estruturais nas teses das ciências naturais e das ciências sociais



Fonte: Elaborado pelo autor

Assim como analisado no subitem anterior, as teses relativas às ciências sociais apresentaram um comportamento mais próximo de um texto científico: com concentrações de densidades de valores de relevância na introdução e conclusão. Já os trabalhos relativos às ciências naturais apresentaram praticamente o comportamento inverso, considerando-se como eixo o comportamento geral. Para as ciências naturais, assim como apresentado anteriormente no Gráfico 21, a concentração da média de relevância dos SNs como descritores ocorre na parte estrutural de desenvolvimento, juntamente com a de conclusão. O fato da densidade de valor de relevância dos SNs na introdução das teses nesse grupo ser menor pode ainda estar atribuído ao maior consenso de uso de termos em cada área, como foi apresentado aqui no subitem anterior. Dentre as teses das ciências naturais, a seção do *corpus* que apresentou menor % na parte estrutural da introdução foi a de Medicina (Pediatria), como é apresentado no Gráfico 24.

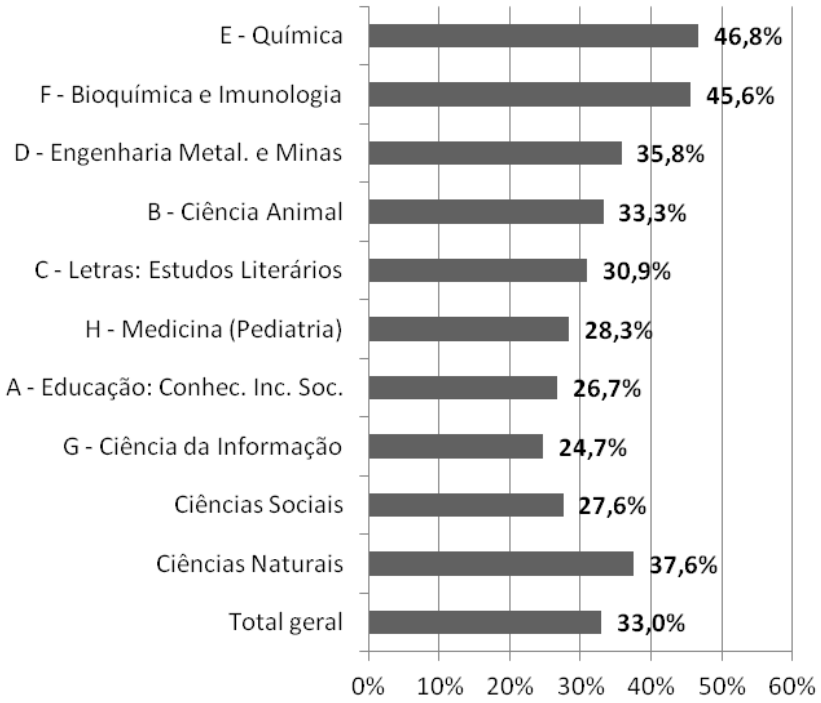
Gráfico 24 - Valores da densidade de relevância dos sintagmas nominais para a parte estrutural da Introdução



Fonte: Elaborado pelo autor

As seções do *corpus* que apresentaram maior concentração na parte de desenvolvimento foram as relacionadas ao programa de pós-graduação em Química e ao programa de Bioquímica e Imunologia, como é apresentado no Gráfico 25 a seguir. Estes foram também, dentre os programas relacionados às ciências naturais, os que apresentaram menores % de SNs na parte de desenvolvimento, conforme foi apresentado no Gráfico 8. Podemos concluir que, para esses programas, além de haver uma maior probabilidade em se encontrar SNs relevantes como descritores nas partes de desenvolvimento, há uma maior facilidade para se executar tal tarefa, dado que essas partes são relativamente menores que nos outros programas das ciências naturais.

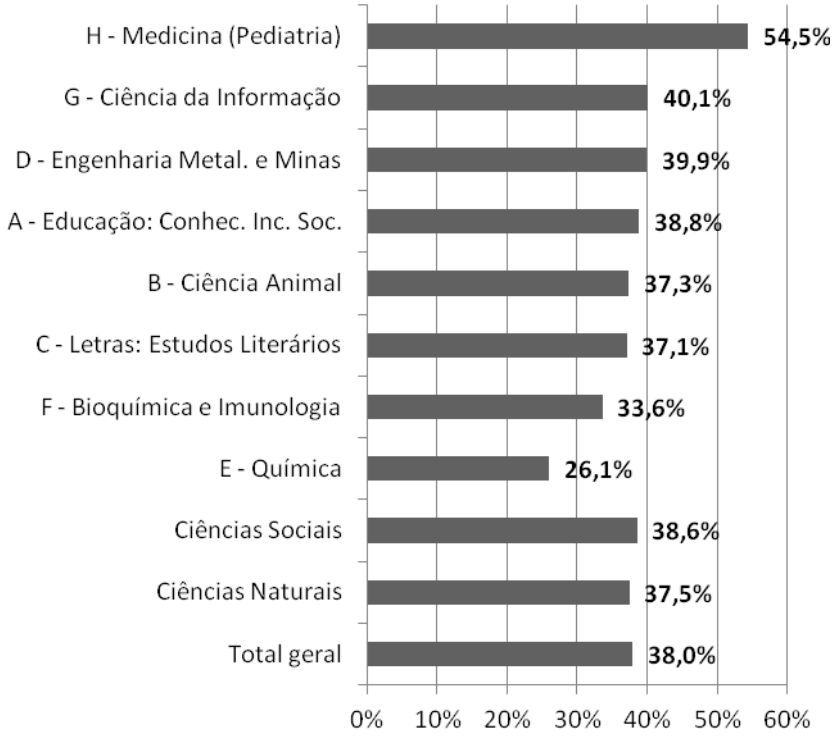
Gráfico 25 - Valores da densidade de relevância dos sintagmas nominais para a parte estrutural do Desenvolvimento



Fonte: Elaborado pelo autor

Mais uma vez, a seção relativa ao programa de pós-graduação em Medicina (Pediatria) apresentou um comportamento diferente das demais, com alta concentração na conclusão, como é apresentado no Gráfico 26 a seguir. Embora o tamanho desta amostra tenha sido o menor, 7 teses, 5 delas apresentaram maior densidade de valor de relevância dos SNs como descritores na parte da conclusão. Esse fato confirma que, para esse programa, a parte estrutural da conclusão é a mais densa dentre todas as outras partes de todos os programas, sendo a mais indicada para a extração de SNs como descritores.

Gráfico 26 - Valores da densidade de relevância dos sintagmas nominais para a parte estrutural da Conclusão



Fonte: Elaborado pelo autor.

O programa de pós-graduação que apresentou uma menor densidade de valores de relevância na parte de desenvolvimento foi o de Ciência da Informação, conforme foi apresentado no Gráfico 25. É possível caracterizá-lo, dentre os demais programas de pós-graduação aqui analisados, como aquele que mais se comportou de acordo com a distribuição esperada para textos científicos (com concentrações na introdução e conclusão). Para SRIs que consideram como critério tal distribuição padrão para textos científicos, como o apresentado por Galeas, Kretschmer e Freisleben (2009), a seção do *corpus* relativa ao programa de pós-graduação em Ciência da Informação seria, então, a mais indicada.

As distribuições apresentadas no subitem anterior são formalizadas através de equações no subitem a seguir, com o objetivo de possibilitar pesquisas futuras para tais SRIs que consideram a distribuição dos descritores como critério para a recuperação de informação, assim como para indexação automática.

4.5.2 Análise polinomial da distribuição dos valores de relevância

A distribuição dos descritores em um texto pode ser usada tanto para o processo de indexação automática como para a recuperação de informação que considere tal critério. Umas das técnicas apresentada no capítulo de revisão da literatura utiliza a expansão matemática pela série de Fourier (GALEAS; KRETSCHMER; FREISLEBEN, 2009). Neste subitem, é apresentada para cada seção do *corpus* uma formalização matemática mais simples: uma equação polinomial que pôde ser obtida através de recursos gráficos disponíveis no Microsoft Excel com a adição de linhas de tendência.

As funções polinomiais apresentadas aqui permitem formalizar o comportamento da distribuição de valores de relevância de descritores para cada seção do *corpus* e também verificar o grau de complexidade que cada distribuição apresentou. O grau polinomial encontrado para cada seção do *corpus* foi o menor possível, atendendo ao mínimo de 90% de variabilidade[47] ou o grau máximo de 6[48]. Os polinômios são apresentados na Tabela 22 e do Gráfico 27 ao Gráfico 34 a seguir.

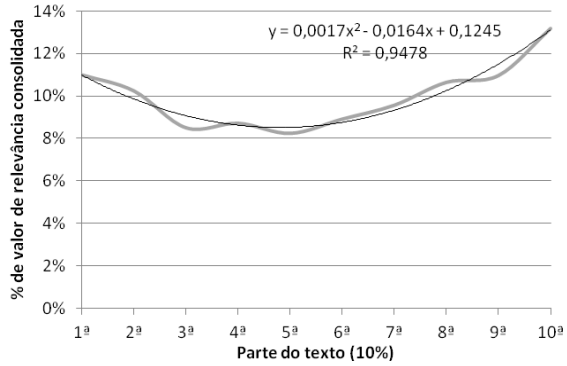
Tabela 22 - Equação da % do valor de relevância (y) de uma parte (x, de 1 a 10) em uma tese do *corpus*

Seção do *corpus*	R^2	Grau	Equação
A - Educação: Conhec. Inc. Soc.	94,8%	2	$y = 0{,}0017x^2 - 0{,}0164x + 0{,}1245$
B - Ciência Animal	97,3%	3	$y = -0{,}0007x^3 + 0{,}0121x^2 - 0{,}0517x + 0{,}1199$
C - Letras: Estudos Literários	94,4%	6	$y = -3E\text{-}05x^6 + 0{,}001x^5 - 0{,}0118x^4 + 0{,}0674x^3 - 0{,}1838x^2 + 0{,}2011x + 0{,}0495$
D - Engenharia Metal. e Minas	90,0%	5	$y = 2E\text{-}05x^5 - 0{,}0008x^4 + 0{,}009x^3 - 0{,}0394x^2 + 0{,}0654x + 0{,}036$
E - Química	92,7%	6	$y = -3E\text{-}05x^6 + 0{,}0012x^5 - 0{,}0169x^4 + 0{,}11x^3 - 0{,}343x^2 + 0{,}4572x - 0{,}1222$
F - Bioquímica e Imunologia	90,9%	6	$y = 6E\text{-}05x^6 - 0{,}0018x^5 + 0{,}0209x^4 - 0{,}1166x^3 + 0{,}3207x^2 - 0{,}3964x + 0{,}221$
G - Ciência da Informação	85,2%	6	$y = 4E\text{-}05x^6 - 0{,}0011x^5 + 0{,}0137x^4 - 0{,}0834x^3 + 0{,}2616x^2 - 0{,}3987x + 0{,}3116$
H - Medicina (Pediatria)	67,5%	6	$y = -1E\text{-}05x^6 + 0{,}0004x^5 - 0{,}0057x^4 + 0{,}0369x^3 - 0{,}1136x^2 + 0{,}1541x - 0{,}0015$

Fonte: Elaborado pelo autor.

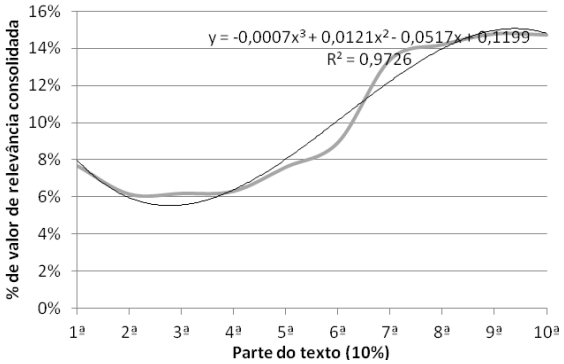
[47] A variabilidade é indicada pelo R^2 que foi utilizado aqui também na Tabela 19 na página 63.
[48] A limitação do grau 6 foi a mesma apresentada pelo recurso de linha de tendência do Microsoft Excel.

Gráfico 27 - Distribuição dos valores de relevância em 10 partes: seção A - Educação: Conhec. Inc. Soc.



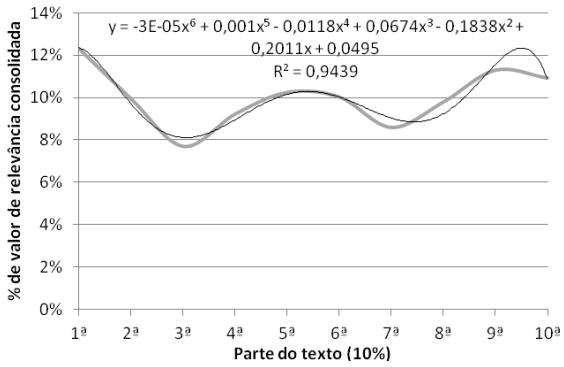
Fonte: Elaborado pelo autor.

Gráfico 28 - Distribuição dos valores de relevância em 10 partes: seção B - Ciência Animal



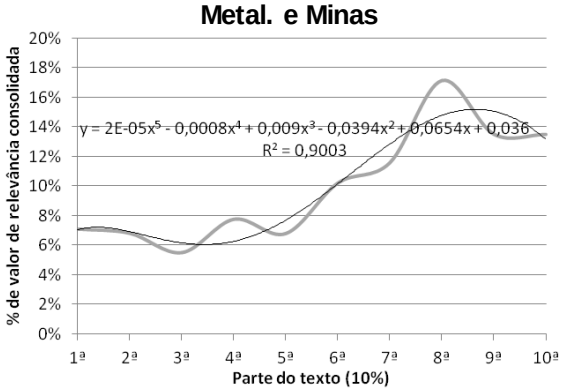
Fonte: Elaborado pelo autor.

Gráfico 29 - Distribuição dos valores de relevância em 10 partes: seção C - Letras: Estudos Literários



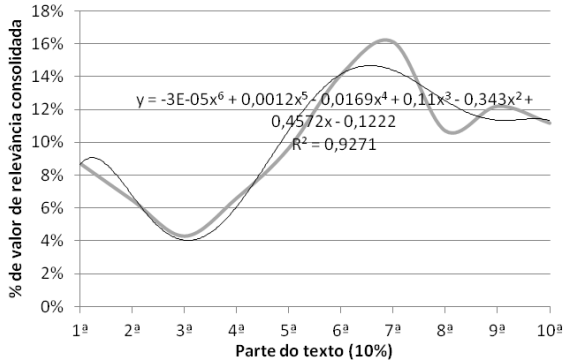
Fonte: Elaborado pelo autor.

Gráfico 30 - Distribuição dos valores de relevância em 10 partes: seção D - Engenharia Metal. e Minas



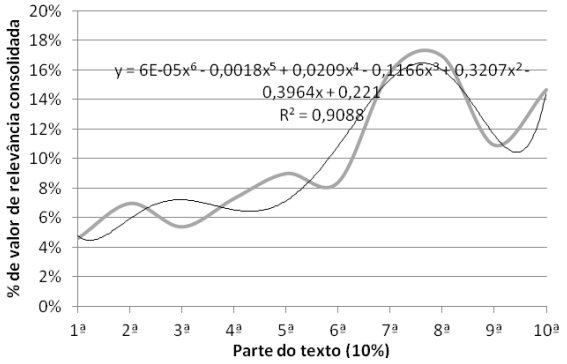
Fonte: Elaborado pelo autor.

Gráfico 31 - Distribuição dos valores de relevância em 10 partes: seção E - Química



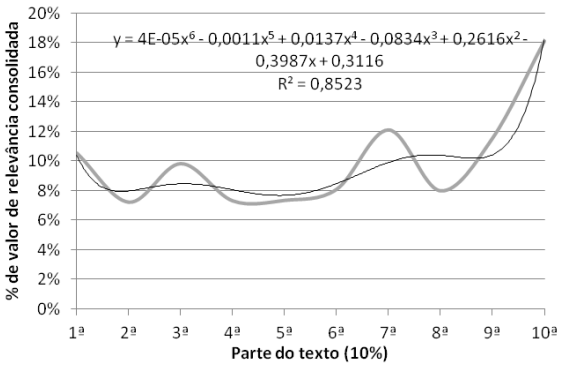
Fonte: Elaborado pelo autor.

Gráfico 32 - Distribuição dos valores de relevância em 10 partes: seção F - Bioquímica e Imunologia



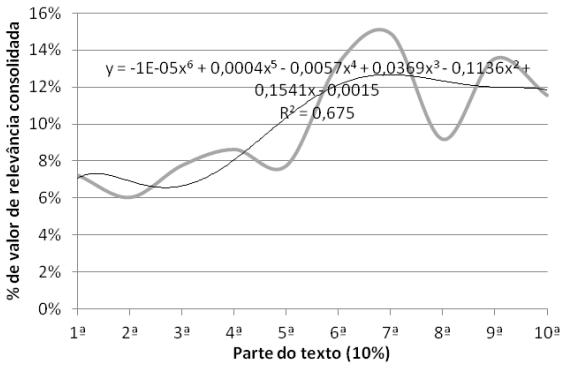
Fonte: Elaborado pelo autor.

Gráfico 33 - Distribuição dos valores de relevância em 10 partes: seção G - Ciência da Informação



Fonte: Elaborado pelo autor.

Gráfico 34 - Distribuição dos valores de relevância em 10 partes: seção H - Medicina (Pediatria)

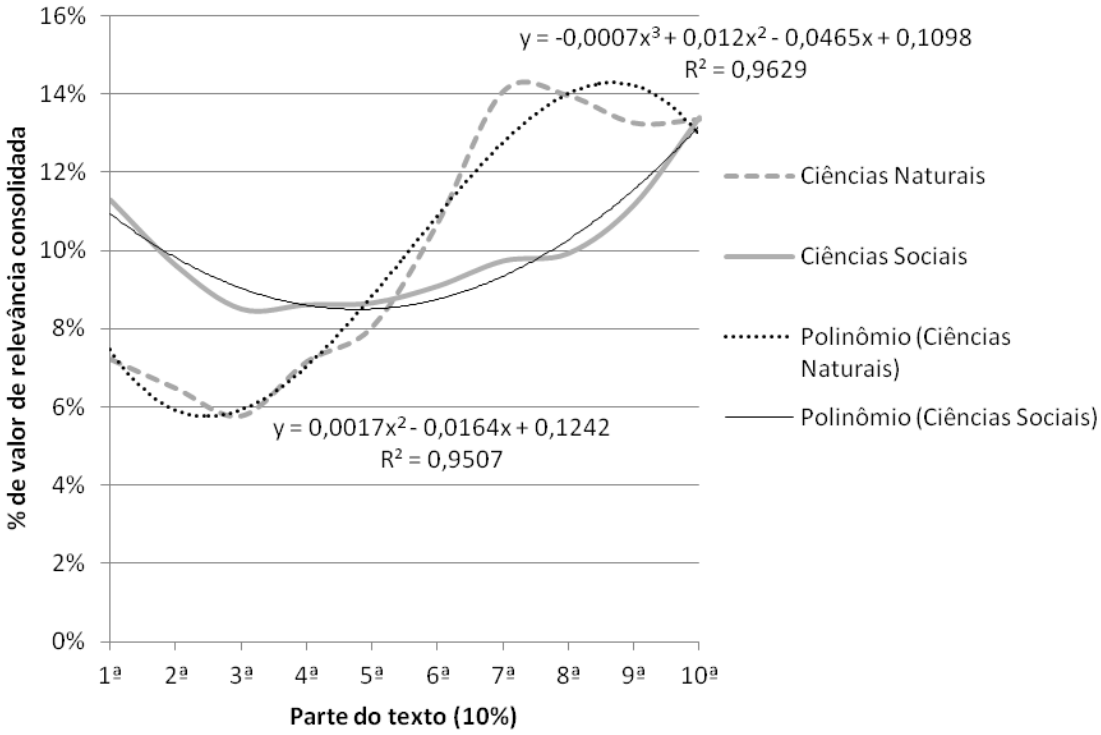


Fonte: Elaborado pelo autor.

As seções que apresentaram menores graus polinomiais foram as mesmas que apresentaram maiores quantidades de teses como amostra (seções A e B). Para as seções menores do *corpus* (G e H) a variabilidade chegou a abaixo de 90%, respeitando-se o limite máximo de grau polinomial em 6. A formalização de uma equação mais robusta para os mesmos exigiria uma quantidade amostral maior que a utilizada.

Os gráficos dos polinômios permitem a visualização de duas formas distintas: a côncava e a convexa. É possível perceber em todos os gráficos relativos às ciências naturais uma área côncava e mais elevada ao final de cada distribuição. Já os gráficos relativos às ciências sociais apresentam uma curvatura geral mais convexa para toda a sua distribuição, como é apresentado no Gráfico 35 a seguir.

Gráfico 35 - Polinômio da distribuição dos valores de relevância em 10 partes nas teses das ciências naturais e das ciências sociais



Fonte: Elaborado pelo autor.

A distribuição dos valores de relevância nas teses dos programas de pós-graduação relacionados às ciências naturais pode ser caracterizada por uma função polinomial de terceiro grau dada pela Equação 4, a seguir.

Equação 4 - Função da % do valor de relevância (y) de uma parte (x, de 1 a 10) em uma tese em ciências naturais

$$\boxed{y = -0{,}0007x^3 + 0{,}012x^2 - 0{,}0465x + 0{,}1098} \quad (R^2=96{,}3\%)$$

Fonte: Elaborado pelo autor

A distribuição dos valores de relevância nas teses dos programas de pós-graduação relacionados às ciências sociais pode ser caracterizada por uma função polinomial de segundo grau dada pela Equação 5, a seguir.

Equação 5 - Função da % do valor de relevância (y) de uma parte (x, de 1 a 10) em uma tese em ciências sociais

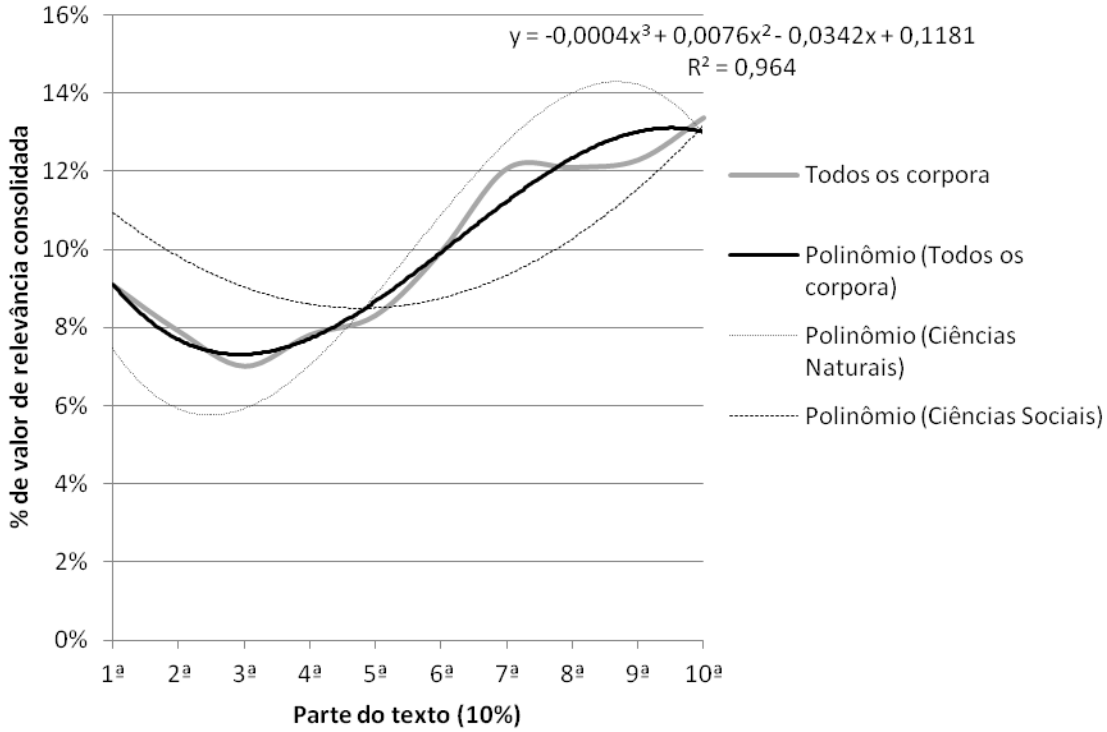
$$y = 0{,}0017x^2 - 0{,}0164x + 0{,}1242 \quad (R^2=95{,}1\%)$$

Fonte: Elaborado pelo autor.

A distribuição nas ciências sociais é mais simples que nas ciências naturais, devido ao seu menor grau polinomial. E, por apresentar um formato convexo, caracteriza-se, como já apresentado aqui anteriormente, como uma distribuição esperada para textos científicos. Já o fato das ciências naturais apresentarem baixa relevância na primeira metade de suas teses indica uma maior uniformidade linguística nesses textos, como já foi analisado aqui. Foi considerado que a primeira metade do discurso abriga os assuntos mais gerais e a segunda metade, aqueles mais específicos.

A função polinomial da distribuição no *corpus* é apresentada no Gráfico 36 a seguir e é resultante das funções relativas às ciências naturais e sociais.

Gráfico 36 - Polinômio da distribuição dos valores de relevância em 10 partes no *corpus*



Fonte: Elaborado pelo autor.

A distribuição dos valores de relevância nas teses de todos os programas de pós-graduação que representaram aqui todas as oito áreas de conhecimento da UFMG

pode ser caracterizada por uma função polinomial de terceiro grau dada pela Equação 6 a seguir.

Equação 6 - Função da % do valor de relevância (y) de uma parte (x, de 1 a 10) em uma tese na UFMG

$$y = -0{,}0004x^3 + 0{,}0076x^2 - 0{,}0342x + 0{,}1181 \quad R^2=96{,}4\%$$

Fonte: Elaborado pelo autor.

A análise de dados apresentada aqui demonstrou que todos os objetivos propostos nesta pesquisa foram alcançados. No capítulo a seguir, são apresentadas as conclusões e indicações de oportunidades de trabalhos futuros.

5 Conclusões

Para que os dados resultantes da pesquisa não ficassem restritos somente à própria área da pesquisa, ou somente ao processo de obtenção dos dados, buscou-se um contato mínimo com todas as outras áreas de conhecimento da instituição onde ela foi desenvolvida, resultando na adoção de 08 programas de pós-graduação para a constituição do *corpus* de pesquisa. Essa decisão permitiu que a pesquisa, além de contribuir para a Ciência da Informação, contribuísse para todas as demais áreas de conhecimento da UFMG onde o presente trabalho foi realizado.

A principal conclusão dessa pesquisa foi comprovar que existe um comportamento característico de distribuição de termos relevantes ao longo de um texto científico. Como o seu comportamento apresentou variações significativas, com certas partes do texto chegando a quase o dobro de valor de relevância de outras, é possível que ele seja usado como um critério para o processo de indexação automática.

O tamanho médio das teses entre as oito áreas de conhecimento da UFMG chegaram a variar quase três vezes entre o menor e o maior tamanho, que foram relativos, respectivamente, aos programas de pós-graduação em Ciência Animal e em Letras: Estudos Literários. O tamanho médio de todas as teses dos programas relacionados às ciências naturais foram menores que os dos relacionados às sociais. O tempo de processamento foi proporcional à quantidade de termos extraídos, logo o tempo de resposta para a indexação automática foi mais lento para os programas relacionados às ciências sociais.

A variação do tamanho médio das partes estruturais (introdução, desenvolvimento e conclusão) entre os diferentes programas de pós-graduação influencia na probabilidade em se encontrar um termo relevante de acordo com tais partes. Os programas que apresentaram menor quantidade de SNs nessas áreas foram os de Engenharia Metalúrgica e o de Ciência da Informação, sendo, portanto, os que apresentam menores custos para a indexação que considera somente estas partes do texto.

Dentre as pesquisas relatadas aqui que utilizaram a extração de SNs, esta extraiu um total quatro vezes maior que todas as demais juntas. A média de SNs extraídos por documento nesta pesquisa foi oito vezes maior que a segunda maior média. Mesmo com as dimensões dos documentos usados nesta pesquisa, o tempo total de processamento chegou a ser menor que em outras, este fato foi devido, sobretudo, à maior disponibilidade de recursos computacionais atuais. Podemos concluir que, com o crescente avanço de recursos de processamento, apontado por Moore (citado por LANCASTER, 1968), as pesquisas de indexação automática podem tender a adotar documentos cada vez maiores, assim como coleções também cada vez maiores.

Aproximadamente 12% de extrações foram consideradas nessa pesquisa como desnecessárias e então excluídas. A média dessas exclusões foi ainda maior para os programas de pós-graduação relacionados às ciências naturais, que possuem uma linguagem mais especializada, como no caso da Química, que utiliza um vocabulário controlado da língua inglesa e apresentou uma média de 17% de exclusões. O programa de pós-graduação em Letras: Estudos Literários apresentou a menor taxa, cerca de 11%, revelando uma maior proeficiência de seus autores na língua.

A ferramenta Ogma, que utiliza um vocabulário geral de nossa língua, pode, por exemplo, considerar a inserção de novos termos em sua base de dados, em um processo interativo com o usuário, como no modelo probabilístico, com o intuito de diminuir essas taxas de exclusões. A qualidade da extração automática de SNs propiciada pela ferramenta Ogma pode ser ainda avaliada em comparação com outras ferramentas, como o Palavras. No entanto, sua facilidade de uso contribuiu de forma significativa diante dos demais desafios desta pesquisa para um período de tempo relativamente curto. Outra possível melhoria no Ogma é a correção da falha que causa a inconsistência entre os SNs somente extraídos com o parâmetro *s* e os extraídos com o parâmetro *tral*.

O protótipo para a seleção automática de SNs como candidatos a descritores foi desenvolvido através de procedimentos que podem ser reproduzidos de forma totalmente automática, sendo que a maioria deles já está desenvolvida em *macros* do Microsoft Word e Excel. Essas *macros* podem ser facilmente passadas para um arquivo executável, de modo que, somado ao programa Ogma, se possa obter um produto final que possa ter como entrada qualquer conjunto de documentos textuais. Os detalhes para essa implementação (tais como demarcação automática das partes estruturais) são indicados para pesquisas futuras.

A atribuição de pesos dada na Equação 2, e utilizada aqui no processo de seleção automática de SNs como candidatos a descritores, pode ser melhorada com a normalização logarítmica do fator f_{ij} para (1 + log f_{ij}), como foi apresentado na terceira equação do Quadro 3. Essa possibilidade, embora precise ser verificada, tem probabilidades de produzir melhores resultados uma vez que tal fator mascarou os demais fatores, como o da frequência invertida dos documentos (log N/n_i) e o de classificação do SN. Este último fator apresentou um comportamento aleatório, o quê indica que seus valores atribuídos na Tabela 6 devam ser revisados. Outra alternativa seria desconsiderar esse fator e verificar o impacto na relevância atribuída aos candidatos.

Embora os custos para a realização tenham sido significativos, a etapa na qual os próprios autores avaliam os candidatos a descritores de suas próprias teses possibilitou uma validação o mais próxima possível daquela realizada por indexadores especialistas.

Caso não houvesse uma validação, a credibilidade dos resultados da distribuição poderiam ser comprometidos.

Apesar das exclusões de alguns SNs extraídos, os SNs candidatos a descritores que foram enviados para avaliação pelos autores ainda apresentaram termos que poderiam ser desconsiderados como, por exemplo, referências a datas, nomes próprios, expressões metalinguísticas (como “grifos do autor”, “tradução nossa”), etc. Mais uma vez, é recomendável que o processador de linguagem natural utilizado possa aceitar novos termos e regras para a determinação de suas *stoplists*, ou que estas sejam elaboradas adicionalmente, como foi feito através de *macros* nesta pesquisa.

A metodologia para a seleção de candidatos a descritores, baseada em Souza (2005) obteve um êxito de 77,9% de aprovação de relevância pelos autores, valor um pouco menor do que Souza (2005) conseguiu na sua própria pesquisa, 88,9%. Ou seja, nesta pesquisa, cerca de 6 em cada 8 candidatos a descritores são relevantes, sendo que 3 destes são extremamentes relevantes, 2 razoavelmente e 1 moderadamente relevante. Dadas as necessidades de melhorias já apontadas, esses valores de relevância são aceitáveis.

Assim como apresentado antes, os programas mais relacionados às ciências naturais apresentaram menores valores de relevância como descritor, uma vez que possuem um vocabulário mais especializado. Outro fator que influenciou de maneira positiva, porém equivocadamente, na pontuação de relevância dada pelos autores foi o fato de eles considerarem os descritores como representantes de assuntos de suas teses, ao invés de considerar os descritores que melhor representariam seus textos frente aos demais de sua área. Para uma nova pesquisa nesse sentido, seria importante frisar esse último conceito junto aos entrevistados.

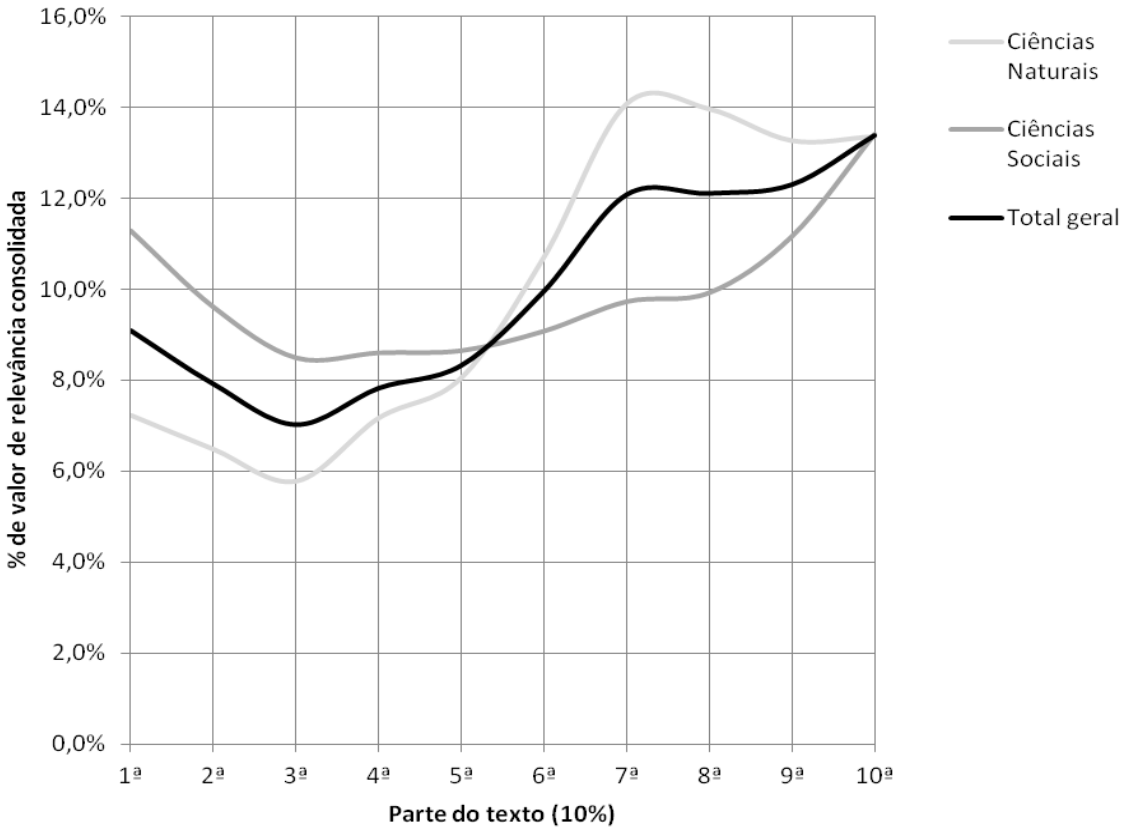
Outro êxito no emprego da metodologia de Souza (2005) foi a constatação da variação da relevância atribuída pelo autor de forma decrescente, assim como a pontuação dada ao candidato a descritor. No entanto, enquanto a primeira foi linear, a segunda foi exponencial. Isto reforça a necessidade em se adotar uma normalização logarítmica para o fator f_{ij}, como apresentado anteriormente, o que provocaria que a pontuação sofresse uma variação linear também.

De acordo com a projeção linear decrescente dos valores de relevância dados pelos autores, foi possível estimar, de modo mediano, a quantidade ideal de candidatos que poderiam ser enviados de modo a se obter uma quantidade mais exaustiva de descritores relevantes. Para essa pesquisa, ao invés dos 20 candidatos enviados a cada autor, o ideal seria 80 candidatos, aproximadamente. Para os programas de pós-graduação, foi possível estimar somente para dois deles a tal quantidade ideal com um certo nível de confiabilidade: 50 para Letras e 40 para Ciência da Informação. Os demais apresentaram um

comportamento muito distante de ser linear de modo a se fazer uma estimativa. Embora um questionário com maiores quantidades de questões possa implicar em menores adesões, o relato positivo de alguns entrevistados apontando a facilidade do seu preenchimento favorece esse possível aumento.

As teses do *corpus* relativas às ciências naturais apresentaram um comportamento semelhante de distribuição de termos relevantes ao longo dos textos de suas teses de doutorado, assim como, entre si, as teses relativas às ciências sociais também apresentaram semelhanças. No entanto, as ciências naturais apresentou um comportamento distinto das ciências sociais, como é apresentado novamente no mesmo Gráfico 21 já apresentado:

Gráfico 21 - Distribuição dos valores de relevância em 10 partes nas teses das ciências naturais e das ciências sociais



Fonte: Elaborado pelo autor.

As ciências sociais possuem um comportamente de distribuição de termos relevantes para indexação mais similar ao esperado para textos científicos. As ciências naturais apresentam uma concentração maior de termos relevantes na segunda metade do texto. Este fato pode ser atribuído ao mesmo motivo que leva estes textos a serem menores. Em função das ciências naturais tenderem a apresentar maior consenso de utilização de

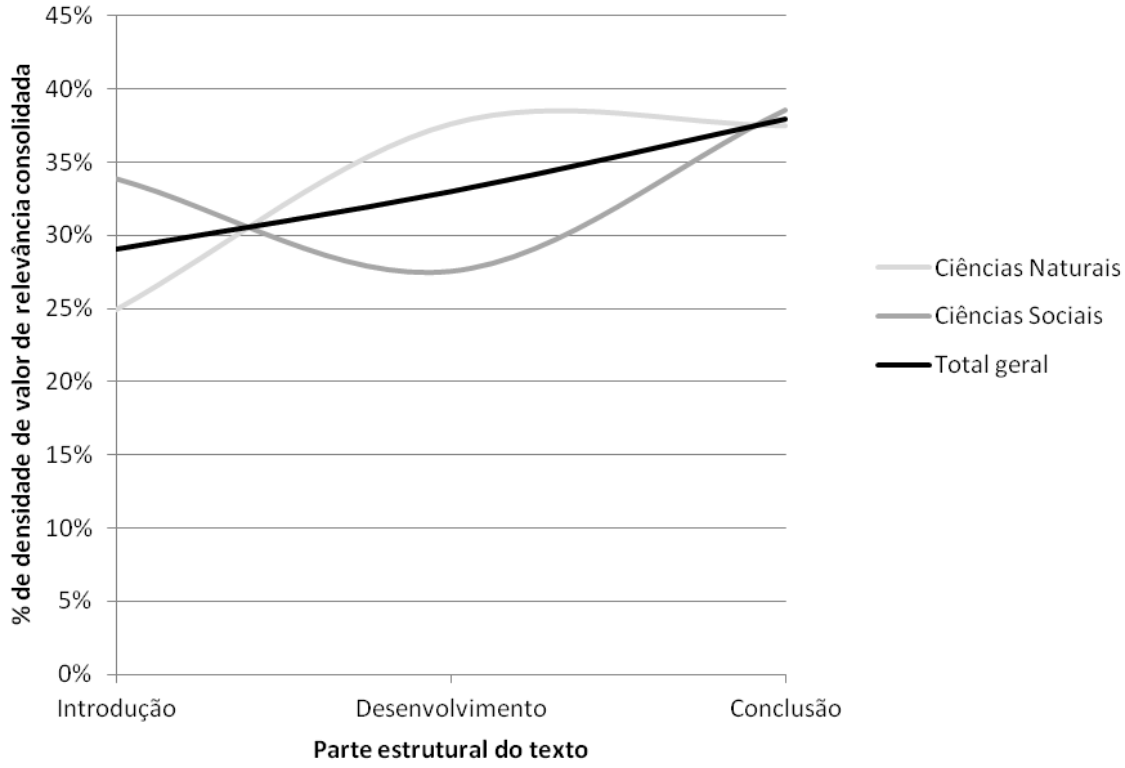
termos, seus discursos são mais concisos, necessitando uma menor quantidade de argumentações para apresentar assuntos gerais da área. Uma vez que tais assuntos costumam se encontrar na primeira metade do texto, nesta parte há uma tendência de utilização de termos que seriam usados pelos demais autores, tornando-os então termos menos relevantes em função de sua menor especificidade. No entanto, assim como esperado para textos científicos, as ciências naturais também apresentam um declínio logo após a introdução e que dura até aproximadamente 30% do seu texto, quando ocorre sua posição de menor relevância para termos de indexação.

Dentre todos os programas, o de pós-graduação em Química, que foi considerado aqui o de vocabulário mais consensual, apresenta o menor valor em 30% da posição ao longo de seus textos. Tal valor chega a 4%, diferindo do seu ápice aos 70% da posição que chega a 16% de relevância. Essa diferença é de 4 vezes mais que a menor. Para esses textos com maior concentração de relevância em poucos pontos, um indexador necessitará de menos esforço para eleger termos caso utilize somente amostras por posição.

Como trabalho futuro, uma segmentação mais detalhada dos textos, com marcação de partes como a de referencial teórico, pode comprovar a hipótese de que nessa parte há menor quantidade de termos relevantes para a indexação de textos das ciências naturais. Para isso, é desejável o desenvolvimento de ferramentas que permitam a demarcação dessas partes de forma automática.

A distribuição de relevância encontrou o mesmo comportamento quando analisada pelas partes estruturais de introdução, desenvolvimento e conclusão. As teses de ciências sociais apresentaram comportamento mais esperado para textos científicos, enquanto as teses de ciências naturais concentraram sua relevância após a metade do texto, como é apresentado novamente no Gráfico 23 a seguir.

Gráfico 23 - Distribuição dos valores da densidade de relevância dos sintagmas nominais por partes estruturais nas teses das ciências naturais e das ciências sociais



Fonte: Elaborado pelo autor.

De todas as partes, as que apresentam menores densidades de relevância são as relativas à introdução nos textos das ciências naturais e à parte de desenvolvimento nos de ciências sociais. O primeiro, pelo fato de tender a possuir um vocabulário mais consensual para assuntos gerais, como já foi concluído, e o outro por necessitar mais espaço para argumentar seus assuntos gerais da área, levando a textos mais longos e que diluem a média da relevância de seus termos como descritores.

O objetivo principal desta pesquisa, especificado na página 22, foi analisar se há um comportamento característico de distribuição de termos relevantes ao longo de um texto científico que possa contribuir como um critério para o processo de indexação automática do mesmo. Além desse objetivo ser alcançado, o comportamento da distribuição foi caracterizado através das equações matemáticas da Tabela 23 a seguir:

Tabela 23 – Equações finais do comportamento da distribuição do valor de relevância

Seção do corpus	R^2	Grau	Equação
A - Educação: Conhec. Inc. Soc.	94,8%	2	$y = 0,0017x^2 - 0,0164x + 0,1245$
B - Ciência Animal	97,3%	3	$y = -0,0007x^3 + 0,0121x^2 - 0,0517x + 0,1199$
C - Letras: Estudos Literários	94,4%	6	$y = -3E-05x^6 + 0,001x^5 - 0,0118x^4 + 0,0674x^3 - 0,1838x^2 + 0,2011x + 0,0495$
D - Engenharia Metal. e Minas	90,0%	5	$y = 2E-05x^5 - 0,0008x^4 + 0,009x^3 - 0,0394x^2 + 0,0654x + 0,036$
E – Química	92,7%	6	$y = -3E-05x^6 + 0,0012x^5 - 0,0169x^4 + 0,11x^3 - 0,343x^2 + 0,4572x - 0,1222$
F - Bioquímica e Imunologia	90,9%	6	$y = 6E-05x^6 - 0,0018x^5 + 0,0209x^4 - 0,1166x^3 + 0,3207x^2 - 0,3964x + 0,221$
G - Ciência da Informação	85,2%	6	$y = 4E-05x^6 - 0,0011x^5 + 0,0137x^4 - 0,0834x^3 + 0,2616x^2 - 0,3987x + 0,3116$
H - Medicina (Pediatria)	67,5%	6	$y = -1E-05x^6 + 0,0004x^5 - 0,0057x^4 + 0,0369x^3 - 0,1136x^2 + 0,1541x - 0,0015$
Ciências Naturais	96,3%	3	$y = -0,0007x^3 + 0,012x^2 - 0,0465x + 0,1098$
Ciências Sociais	95,1%	2	$y = 0,0017x^2 - 0,0164x + 0,1242$
Todas as teses	**96,4%**	**3**	$\mathbf{y = -0,0004x^3 + 0,0076x^2 - 0,0342x + 0,1181}$

Fonte: Elaborado pelo autor.

Com essas formulações, na tabela anterior, comprovamos um comportamento de distribuição de relevância dos descritores ao longo de teses da UFMG, contribuindo assim para possíveis processos de indexação automática que considerem tal critério.

De acordo com as caracterizações das distribuições acima, foi encontrada uma tendência para melhores resultados para amostras com maior quantidade de documentos.

O principal objetivo aqui foi avaliar a distribuição dos valores de relevância dos descritores por posição no texto. A divisão em 10 partes permitiu um acompanhamento o suficientemente detalhado de modo a perceber variações significativas da distribuição em cada seção do *corpus*. Outras divisões podem ser consideradas: maiores e menores que 10.

A expectativa dessa pesquisa era encontrar um comportamento similar para todos os textos do *corpus*, com maiores ênfases no início e final dos textos, como o encontrado de forma mais exemplar na seção da Ciência da Informação. O comportamento distinto para as teses relativas às ciências naturais abriu espaço para novas análises, como até mesmo a estilística. Um dos objetivos dessas análises poderia ser validar se realmente há um maior consenso do emprego de terminologias da área quando os documentos são

relativos às ciências naturais, o que poderia favorecer à maior concentração de valores de relevância na segunda metade dos textos, como constatado mais acentuadamente no programa de pós-graduação em Química, no qual existe um tipo de vocabulário controlado internacional.

O objetivo principal dessa pesquisa foi formalizado com equações que atribuem valores de relevância de acordo com partes do texto, seja num conjunto de dez, seja num conjunto de três partes. Os valores de variabilidade gerais dessas equações foram acima de 90% para a maioria das seções do *corpus* (considerando a limitação de grau da equação polinomial).

Como pesquisas futuras, além das já citadas neste capítulo, há possibilidades de:

- Representação da distribuição das teses de diferentes áreas de conhecimento usando séries de Fourier como é proposto por Galeas, Kretschmer e Freisheben (2009);
- Aplicação da metodologia com amostras de outros programas de pós-graduação;
- Classificação automática de textos com base na distribuição de relevância de seus descritores;
- Análise de fatores que delimitam distintos comportamentos linguísticos apontados pelas diferentes distribuições de relevância de seus descritores;
- Análise quantitativa do valor de relevância dos descritores de acordo com seus comportamentos de distribuição no texto, como aglomerações;

O critério de posicionamento apresentado aqui pode ser, como trabalho futuro, avaliado na indexação automática de teses, tal como foi realizado aqui, e comparado com o mesmo processo sem a inserção desse critério. Os mesmos autores poderiam ser consultados, de modo a verificar se o valor de relevância médio dado por eles apresenta melhorias.

6 Referências

AITCHISON, T. M.; HARDING, P. Automatic indexing and classification for mechanised information retrieval. In: EURIM: CONFÉRENCE EUROPÉENNE SUR LA RECHERCHE DANS L'ORGANISATION DES SERVICES D'INFORMATION ET DES BIBLIOTHÈQUES, 5., 1982, Versailles. **Proceedings...** London: Aslib, 1983 *apud* SAYÃO, L. F. **SALF**: um algoritmo para indexação automática utilizando vocabulário controlado. 1985. 177 f. Dissertação (Mestrado em Ciência da Informação) – Universidade Federal do Rio de Janeiro, IBICT, Rio de Janeiro, 1985.

ARTANDI, S. Machine indexing: linguistic and semiotic implications. **Journal of the American Society for Information Science**, v. 27, n. 4, p. 235-239, jul./aug. 1976 *apud* SAYÃO, L. F. **SALF**: um algoritmo para indexação automática utilizando vocabulário controlado. 1985. 177 f. Dissertação (Mestrado em Ciência da Informação) – Universidade Federal do Rio de Janeiro, IBICT, Rio de Janeiro, 1985.

BABBIE, E. ***Métodos de pesquisa de survey***. Belo Horizonte: UFMG, 1999.

BAEZA-YATES, R.; RIBEIRO-NETO, B. **Modern information retrieval**. New York: ACM Press, 1999. 511p.

BAEZA-YATES, R.; RIBEIRO-NETO, B. **Modern information retrieval**: the concepts and technology behind search. 2. ed. London: Pearson Education Limited, 2011. 913 p.

BARNES, C.I.; CONSTANTINI, L.; PERSCHKE, S. Automatic indexing using the SLC-II system. **Information Processing & Management**, v. 14, n. 2, p.107-119, 1978. Disponível em: < http://www.sciencedirect.com/science/article/pii/0306457378900687>. Acesso em: 19 ago. 2012 *apud* SAYÃO, L. F. **SALF**: um algoritmo para indexação automática utilizando vocabulário controlado. 1985. 177 f. Dissertação (Mestrado em Ciência da Informação) – Universidade Federal do Rio de Janeiro, IBICT, Rio de Janeiro, 1985.

BASTOS, S. B. **Análise comparativa entre indexação automática e manual da literatura brasileira de ciência da informação**. 1984. 204 f. Dissertação (Mestrado em Biblioteconomia e Documentação) - Faculdade de Ciência da Informação, Universidade de Brasília, 1984.

BAXENDALE, P. B. Machine-made index for technical literature: an experiment. **IBM Journal of Research and Development**, v. 2, n. 4, p. 354-361, 1958 *apud* SAYÃO, L. F. **SALF**: um algoritmo para indexação automática utilizando vocabulário controlado. 1985. 177 f. Dissertação (Mestrado em Ciência da Informação) – Universidade Federal do Rio de Janeiro, IBICT, Rio de Janeiro, 1985.

BERSTEIN, L. M.; WILLIANSON, R. E. Testing of natural language retrieval system for a full text knowledge base. **Journal of the American Society for Information Science**, v. 35, n. 4, p. 235-47, 1984 *apud* SAYÃO, L. F. **SALF**: um algoritmo para indexação automática utilizando vocabulário controlado. 1985. 177 f. Dissertação (Mestrado em Ciência da Informação) – Universidade Federal do Rio de Janeiro, IBICT, Rio de Janeiro, 1985.

BICK, E. **The Parsing System Palavras***: a*utomatic grammatical analysis of portuguese in a constraint grammar framework. Aarhus: Aarhus University Press, 2000.

BOOKSTEIN, A. Implication of boolean structure for probabilistic retrieval. In: PROC OF THE EIGHT ANNUAL INTERNATIONAL ACM/SIGIR CONFERENCE ON RESEARCH AND DEVELOPMENT IN INFORMATION RETRIEVAL*, 8., 1985,* Montreal, Canada. **Proceedings...** New York: ACM, 1985. p. 11-17. Disponível em: <http://dl.acm.org/citation.cfm?id=253505>. Acesso em: 20 nov. 2011 *apud* SAYÃO, L. F. **SALF**: um algoritmo para indexação automática utilizando vocabulário controlado. 1985. 177 f. Dissertação (Mestrado em Ciência da Informação) – Universidade Federal do Rio de Janeiro, IBICT, Rio de Janeiro, 1985.

BOOKSTEIN, A. On the perils of merging boolean and weighted retrieval systems. **Journal of the American Society for Information Sciences**, v. 29, n. 3, p.156-158, 1978 *apud* SAYÃO, L. F. **SALF**: um algoritmo para indexação automática utilizando vocabulário controlado. 1985. 177 f. Dissertação (Mestrado em Ciência da Informação) – Universidade Federal do Rio de Janeiro, IBICT, Rio de Janeiro, 1985.

BOOKSTEIN, A.; SWASON, D. R. Probabilistic models for automatic indexing. **Journal of the American Society for Information Science**, v. 25, n. 5, p. 312-316, sep./oct. 1974. Disponível em: <http://onlinelibrary.wiley.com/doi/10.1002/asi.4630250505/abstract>. Acesso em: 15 nov. 2011 *apud* SAYÃO, L. F. **SALF**: um algoritmo para indexação automática utilizando vocabulário controlado. 1985. 177 f. Dissertação (Mestrado em Ciência da Informação) – Universidade Federal do Rio de Janeiro, IBICT, Rio de Janeiro, 1985.

BORGES, G. S. B. **Indexação automática de documentos textuais***:* proposta de critérios essenciais. 2009. 111 f. Dissertação (Mestrado em Ciência da Informação) – Escola de Ciência da Informação, Universidade Federal de Minas Gerais, Belo Horizonte, 2009.

BORKO, H. Information science: what is it?. **American Documentation**, v.19, n.1, p. 3-5, jan. 1968 *apud* SAYÃO, L. F. **SALF**: um algoritmo para indexação automática utilizando vocabulário controlado. 1985. 177 f. Dissertação (Mestrado em Ciência da Informação) – Universidade Federal do Rio de Janeiro, IBICT, Rio de Janeiro, 1985.

BORKO, H. Toward a theory of indexing. **Information Processing and Management**, v. 13, n. 6, p. 355-365, 1977. Disponível em: < http://www.sciencedirect.com/science/article/pii/0306457377900553>. Acesso em: 04 mar. 2012 *apud* SAYÃO, L. F. **SALF**: um algoritmo para indexação automática utilizando vocabulário controlado. 1985. 177 f. Dissertação (Mestrado em Ciência da Informação) – Universidade Federal do Rio de Janeiro, IBICT, Rio de Janeiro, 1985.

BORKO, H. Automatic indexing: a tutorial. In: ACM SIGIR FORUM, 81., 1982, Los Angeles. **Proceedings...** Los Angeles: CA, 1982. p. 9-13. Disponível em: <http://dl.acm.org/citation.cfm?id=1095456>. Acesso em: 10 jan. 2012 *apud* SAYÃO, L. F. **SALF**: um algoritmo para indexação automática utilizando vocabulário controlado. 1985. 177 f. Dissertação (Mestrado em Ciência da Informação) – Universidade Federal do Rio de Janeiro, IBICT, Rio de Janeiro, 1985.

BORKO, H.; BERNIER, C. **Indexing concepts and methods**. New York: Academic Press. 1978 *apud* SAYÃO, L. F. **SALF**: um algoritmo para indexação automática utilizando vocabulário controlado. 1985. 177 f. Dissertação (Mestrado em Ciência da Informação) – Universidade Federal do Rio de Janeiro, IBICT, Rio de Janeiro, 1985.

BRASIL. Comitê Gestor da Internet. **Pesquisa sobre o uso das tecnologias de informação e comunicação no Brasil**: TIC Domicílios e TIC Empresas 2010. São Paulo: Comitê Gestor da Internet no Brasil, 2011.

BRAGA, L. M. **Palavras de títulos e resumos como acesso ao conteúdo do documento**: uma análise numérica. 1982. 181 p. Dissertação (Mestrado em Ciência da Informação) – IBICT, Universidade Federal do Rio de Janeiro, Rio de Janeiro, 1982.

BROZOZOWSKI, J. P. MASQUARADE: searching the full text of abstracts using automatic indexing. ***Journal of Information Science***, v. 6, p. 67-73, fev. 1983. Disponível em: < http://jis.sagepub.com/content/6/2-3/67.refs>. Acesso em: 04 mar. 2012 *apud* SAYÃO, L. F. **SALF**: um algoritmo para indexação automática utilizando vocabulário controlado. 1985. 177 f. Dissertação (Mestrado em Ciência da Informação) – Universidade Federal do Rio de Janeiro, IBICT, Rio de Janeiro, 1985.

BUSH, V. As we may think. **Atlantic Monthly**, v. 176, n. 1, p. 101-108. jul. 1945. Disponível em: <http://www.theatlantic.com/magazine/archive/1945/07/as-we-may-think/303881/>. Acesso em: 04 abr. 2012.

CASTELLS, M. **A sociedade em rede**. 3. ed. São Paulo: Paz e Terra, 1999.

CINTRA, A. M. M. **Para entender as linguagens documentarias**. 2. ed. São Paulo: Polis, 2002. 92 p.

CORRÊA, R. *et al*. Indexação e recuperação de teses e dissertações por meio de sintagmas nominais. **AtoZ**, Curitiba, v. 1, n. 1, ago. 2011. Disponível em: <http://www.atoz.ufpr.br/index.php/atoz/article/view/2>. Acesso em: 05 abr. 2011.

COUTINHO, L. F. **A atividade de indexação**: uma construção social da realidade. 2012. 94 f. Dissertação (Mestrado em Ciência da Informação) – Universidade Federal de Minas Gerais, Programa de Pós-Graduação em Ciência da Informação, Belo Horizonte, 2012.

DIAS, E. W.; NAVES, M. M. L. **Análise de assunto**: teoria e prática. Brasília: Thesaurus, 2007. 116 p.

DILLON, M. Thesaurus-based automatic book indexing. **Information Processing & Management**, v. 18, n. 4, p. 167-78, 1982. *apud* SAYÃO, L. F. **SALF**: um algoritmo para indexação automática utilizando vocabulário controlado. 1985. 177 f. Dissertação (Mestrado em Ciência da Informação) – Universidade Federal do Rio de Janeiro, IBICT, Rio de Janeiro, 1985.

DILLON, M.; FEDERHART, P. Statistical recognition of content terms in general texts. **Journal of the American Society for Information Science**, v. 34, n. 1, p. 3-10, 1984. *apud* SAYÃO, L. F. **SALF**: um algoritmo para indexação automática utilizando vocabulário controlado. 1985. 177 f. Dissertação (Mestrado em Ciência da Informação) – Universidade Federal do Rio de Janeiro, IBICT, Rio de Janeiro, 1985.

DILLON, M.; GRAY, A. Fully automatic syntax-based indexing. **Journal of the American Society for INformation Science**. v. 34, n. 2, p. 99-108, 1983. *apud* SAYÃO, L. F. **SALF**: um algoritmo para indexação automática utilizando vocabulário controlado. 1985. 177 f. Dissertação (Mestrado em Ciência da Informação) – Universidade Federal do Rio de Janeiro, IBICT, Rio de Janeiro, 1985.

DILLON, M. *et al*. The use of automatic indexing for authority control. **Journal of Library Automation**, v. 14, n. 4, p. 268-277, 1981. *apud* SAYÃO, L. F. **SALF**: um algoritmo para indexação automática utilizando vocabulário controlado. 1985. 177 f. Dissertação (Mestrado

em Ciência da Informação) – Universidade Federal do Rio de Janeiro, IBICT, Rio de Janeiro, 1985.

DILLON, M.; MCDONALD, L. K. Fully automatic book indexing. **Journal of Documentation**, v. 39, n. 3, p.135-154, sep. 1983 *apud* SAYÃO, L. F. **SALF**: um algoritmo para indexação automática utilizando vocabulário controlado. 1985. 177 f. Dissertação (Mestrado em Ciência da Informação) – Universidade Federal do Rio de Janeiro, IBICT, Rio de Janeiro, 1985.

DUBOIS, J. *et al*. **Dicionário de lingüística**. São Paulo: Cultrix, 1973. 657p.

DUNHAM, G. S.; PACAK, M. G.; PRATT, A. W. Automatic indexing of pathology data. **Journal of the American Society for Information Science**, p. 81-90, 1978 *apud* SAYÃO, L. F. **SALF**: um algoritmo para indexação automática utilizando vocabulário controlado. 1985. 177 f. Dissertação (Mestrado em Ciência da Informação) – Universidade Federal do Rio de Janeiro, IBICT, Rio de Janeiro, 1985.

ECO, U. **Como se faz uma tese em ciências humanas**. 13. ed. Lisboa: Presença, 2007. 238 p.

EDMUNDSON, H. P. A new method in automatic extracting. **Journal of ACM**, v. 16, n. 2, p.264-285, april 1969. Disponível em: < http://dl.acm.org/citation.cfm?id=321519>. Acesso em: 13 nov. 2011 *apud* SAYÃO, L. F. **SALF**: um algoritmo para indexação automática utilizando vocabulário controlado. 1985. 177 f. Dissertação (Mestrado em Ciência da Informação) – Universidade Federal do Rio de Janeiro, IBICT, Rio de Janeiro, 1985.

EDMUNDSON, H. P. Mathematical models of the texts. **Information Processing & Management**, v. 20, n. 12, p. 235-247, 1984 *apud* SAYÃO, L. F. **SALF**: um algoritmo para indexação automática utilizando vocabulário controlado. 1985. 177 f. Dissertação (Mestrado em Ciência da Informação) – Universidade Federal do Rio de Janeiro, IBICT, Rio de Janeiro, 1985.

FELTRIM, V. D.; ALUÍSIO, S. M.; NUNES, M. G. V. **Uma revisão bibliográfica sobre a estruturação de textos científicos em português**. São Carlos: ICMC-USP, 2000. Disponível em: <https://saga.faccat.br/p907/c_arquivo.php?chave=47&baixar+true.>. Acesso em: 13 jun. 2012.

FIELD, B. J. Towards automatic indexing: automatic assignment of controlled-language indexing and classification from free indexing. **Journal of Documentation**, v. 31, n. 4, 1975 *apud* SAYÃO, L. F. **SALF**: um algoritmo para indexação automática utilizando vocabulário controlado. 1985. 177 f. Dissertação (Mestrado em Ciência da Informação) – Universidade Federal do Rio de Janeiro, IBICT, Rio de Janeiro, 1985.

GALEAS, P., KRETSCHMER, R., FREISLEBEN, B. Document relevance assessment via term distribution analysis using fourier series expansion. In: ACM/IEEE-CS JOINT INTERNATIONAL CONFERENCE ON DIGITAL LIBRARIES, 9., 2009, New York. **Proceedings...** New York, USA: [s.n.], 2009. p. 277–284.

GRAVES, R. W.; HELANDER, D. P. A feasibility study of automatic indexing and information retrieval. **IEEE Transactions on Engineering Writing and Speech**, v. 32, n. 2, p. 58-59, 1970 *apud* SAYÃO, L. F. **SALF**: um algoritmo para indexação automática utilizando vocabulário controlado. 1985. 177 f. Dissertação (Mestrado em Ciência da Informação) – Universidade Federal do Rio de Janeiro, IBICT, Rio de Janeiro, 1985.

HALLER, J. Análise automática de textos em sistemas de informação. **Revista de Biblioteconomia de Brasília,** n. 11, v. 1, p.105-113, jan./jun. 1983.

HALLER, J. Processamento de textos em linguagem natural. In: CONGRESSO NACIONAL DE INFORMÁTICA, 15., 1982, Rio de Janeiro. **Anais...** Rio de Janeiro: [s.n.], 1982. 9 p.

HJΦRLAND, B. Toward a theory of aboutness, subject, topicality, theme, domain, field, content... and relevance. **Journal of the American Society for Information Science and Technology**, v. 52, n. 9, p. 774-778, 2001.

KOBASHI, N. Y. *A* **Elaboração de informações documentárias**: em busca de uma metodologia. 1994. Tese (Doutorado em Ciência da Informação) – Departamento de Biblioteconomia e Documentação, Universidade de São Paulo, São Paulo, 1994.

KOBASHI, N. Y.; FERNANDES, J. C. Pragmática linguística e organização da informação. In: ENCONTRO NACIONAL DE PESQUISA EM CIÊNCIA DA INFORMAÇÃO, 10., 2009, Paraíba. **Anais Eletrônicos...** João Pessoa: UFPB, 2009. Disponível em: <http://dci2.ccsa.ufpb.br:8080/jspui/handle/123456789/491>. Acesso em nov. de 2011.

KURAMOTO, H. **Proposition d'un système de recherche d'Information assistée par ordinateur avec application à la langue portugaise**. 1999. Tese (Doutorado em Ciências da Informação e da Comunicação) – Université Lumière Lyon 2, Paris, França, 1999.

KURAMOTO, H. Uma abordagem alternativa para o tratamento e a recuperação de informação textual : os sintagmas nominais. **Revista Ciência da Informação**, v. 25, n. 2, 1996.

LADEIRA, A. P. **Processamento de linguagem natural**: caracterização da produção científica dos pesquisadores brasileiros. Tese (Doutorado em Ciência da Informação) – Escola de Ciência da Informação, Universidade Federal de Minas Gerais, Belo Horizonte, 2010.

LANCASTER, F. W. **Indexação e resumos**: teoria e prática. 2. ed. Brasília: Briquet de Lemos, 2004.

LANCASTER, F. W. **Information retrieval systems:** characteristics, testing and evaluation. New York: Willy, 1968.

LAVILLE, C.; DIONNE, J. **A construção do saber: manual de metodologia de pesquisa em ciências humanas**. Porto Alegre: Artes médicas, 1999.

LEVINE, D. M.; BERENSON, M. L.; STEPHAN, D. **Estatística**: teoria e aplicações usando Microsoft Excel em português. Rio de Janeiro: LTC, 2000.

LIMA, G. A. B. O. **Protótipo Mapa Hipertextual - MHTX**: um modelo para organização hipertextual de documentos acadêmicos por meio do uso de mapas conceituais, analise facetada e sistemas hipertextuais. Belo Horizonte, 2004. Disponível em: <www.gercinalima.com/mhtx>. Acesso em: 31 out. 2010.

LUHN, H. P. A statistical approach to mechanized enconding and searching of literature information. **IBM Journal of Research and Development**, v. 1, n. 4, p. 309-317, oct. 1957 *apud* SAYÃO, L. F. **SALF**: um algoritmo para indexação automática utilizando vocabulário

controlado. 1985. 177 f. Dissertação (Mestrado em Ciência da Informação) – Universidade Federal do Rio de Janeiro, IBICT, Rio de Janeiro, 1985.

LUHN, H. P. The automatic creation of literature abstracts. **IBM Journal of Research and Development**, v. 2, p. 159-165, 1958a *apud* SAYÃO, L. F. **SALF**: um algoritmo para indexação automática utilizando vocabulário controlado. 1985. 177 f. Dissertação (Mestrado em Ciência da Informação) – Universidade Federal do Rio de Janeiro, IBICT, Rio de Janeiro, 1985.

LUHN, H. P. An experiment in auto-abstracting: auto-abstracts of área 5. In: INTERNATIONAL CONFERENCE ON SCIENTIFIC INFORMATION, 1958b, New York. **Proceedings...**, New York:Yorktown Heights, 18 p. *apud* SAYÃO, L. F. **SALF**: um algoritmo para indexação automática utilizando vocabulário controlado. 1985. 177 f. Dissertação (Mestrado em Ciência da Informação) – Universidade Federal do Rio de Janeiro, IBICT, Rio de Janeiro, 1985.

LUHN, H. P. **Auto-encoding of documents for information retrieval system**. London: Pergamon Press, 1959 *apud* SAYÃO, L. F. **SALF**: um algoritmo para indexação automática utilizando vocabulário controlado. 1985. 177 f. Dissertação (Mestrado em Ciência da Informação) – Universidade Federal do Rio de Janeiro, IBICT, Rio de Janeiro, 1985.

LUHN, H. P. Automatic intelligence systems: some basic problems and prerequisites for their solution. In: TOMESKI, E. A.; WESTCOTT, R. (Ed.). **Clarification, unification and integration of storage and retrieval.** New York: Management Dynamics, 1961. p. 3-20. *apud* SAYÃO, L. F. **SALF**: um algoritmo para indexação automática utilizando vocabulário controlado. 1985. 177 f. Dissertação (Mestrado em Ciência da Informação) – Universidade Federal do Rio de Janeiro, IBICT, Rio de Janeiro, 1985.

LYONS, J. **Linguagem e Lingüística**: uma introdução. Rio de Janeiro: Livros Tecnicos e Cientificos, 1987. 322 p.

MAIA, L. C. G. **Uso de sintagmas nominais na classificação automática de documentos**. Tese (Doutorado em Ciência da Informação). Orientador Prof. Dr. Renato Rocha Souza. UFMG, ECI, 2008.

MARON, M. E. Automatic indexing: na experimental inquiry. **Journal of the Association for Computing Machinery**, v. 8, p. 404-417, 1961. *apud* SAYÃO, L. F. **SALF**: um algoritmo para indexação automática utilizando vocabulário controlado. 1985. 177 f. Dissertação (Mestrado em Ciência da Informação) – Universidade Federal do Rio de Janeiro, IBICT, Rio de Janeiro, 1985.

MARON, M.; KUHNS, J. On relevance, probabilistic indexing and information retrieval. **Journal of ACM**, v. 7, n. 3, p. 216-244, 1960. *apud* SAYÃO, L. F. **SALF**: um algoritmo para indexação automática utilizando vocabulário controlado. 1985. 177 f. Dissertação (Mestrado em Ciência da Informação) – Universidade Federal do Rio de Janeiro, IBICT, Rio de Janeiro, 1985.

MARTINS, A. V. **Um método para indexação automática de textos**. 1983. 100 f. Dissertação (Mestrado em Sistemas e Computação) – Instituto Militar de Engenharia, Rio de Janeiro, 1983.

MOYNE, J. A. Information retrieval and natural language. In: AMERICAN SOCIETY FOR INFORMATION SCIENCE, 1969, New York. **Proceedings...** New York: [s.n.], 1969. p. 259-

263. *apud* SAYÃO, L. F. **SALF**: um algoritmo para indexação automática utilizando vocabulário controlado. 1985. 177 f. Dissertação (Mestrado em Ciência da Informação) – Universidade Federal do Rio de Janeiro, IBICT, Rio de Janeiro, 1985.

NISHIDA, F.; TAKAMATSU, S.; FUJITA, Y. Semiautomatic indexing of structured information text. **Journal of Chemical Information and Computer Sciences**, v. 24, n. 1, p. 15-20, 1984 *apud* SAYÃO, L. F. **SALF**: um algoritmo para indexação automática utilizando vocabulário controlado. 1985. 177 f. Dissertação (Mestrado em Ciência da Informação) – Universidade Federal do Rio de Janeiro, IBICT, Rio de Janeiro, 1985.

ORTEGA, C. D. ; LARA, M. L. G. A noção de estrutura e os registros de informação dos sistemas documentários. **Transinformação**, v. 22, p. 7-17, 2010.

ORTEGA, C. D. Relações históricas entre Biblioteconomia, Documentação e Ciência da Informação. **DataGramaZero**, v. 5, n. 5, out. 2004.

OTHERO, G. A. **A gramática da frase em português**: algumas reflexões para a formalização da estrutura frasal em português. Porto Alegre: EDIPUCRS, 2009. 160 p.

PERINI, M. A. *et al*. O SN em português: a hipótese mórfica. **Revista de Estudos de Linguagem - UFMG**, Belo Horizonte, p. 43-56, jul./dez. 1996.

ROBERTSON, S.; SPARCK JONES, K. Relevance weighting of search terms. **Journal of the American Society for Information Sciences**, v. 27, n. 3, p. 129-146, 1976 *apud* SAYÃO, L. F. **SALF**: um algoritmo para indexação automática utilizando vocabulário controlado. 1985. 177 f. Dissertação (Mestrado em Ciência da Informação) – Universidade Federal do Rio de Janeiro, IBICT, Rio de Janeiro, 1985.

ROBREDO, J. A. indexação automática como mecanismo básico no processo de transferência da informação. In: CONGRESSO LATINO-AMERICANO DE BIBLIOTECONOMIA E DOCUMENTAÇÃO, 1., Salvador, 1980. **Anais...** Salvador: FEBAB, 1980, 19 p.

ROBREDO, J. A. Otimização dos processos de indexação dos documentos e de recuperação da informação mediante o uso de instrumentos de controle terminológico. **Ciência da Informação**, v. 11, n. 1, p. 3-18, 1982a.

ROBREDO, J. A. indexação automática de textos: o presente já entrou no futuro. In: MACHADO, U. D. (Ed). **Estudos avançados em Biblioteconomia e Ciência da Informação**, Brasília: ABDF, 1982b. p. 236-74

ROBREDO, J. A.; FERREIRA, J. A. Conceituação de um programa para indexação automática de textos. **Revista de Biblioteconomia de Brasília**, v. 8, n. 2, p. 254-263, jul./dez. 1980.

SALTON, G. Designing automatic information system; results obtained with the SMART programs. *Social Science Information.* Vol. 6(2):111-17, Feb 1967 *apud* SAYÃO, L. F. **SALF**: um algoritmo para indexação automática utilizando vocabulário controlado. 1985. 177 f. Dissertação (Mestrado em Ciência da Informação) – Universidade Federal do Rio de Janeiro, IBICT, Rio de Janeiro, 1985.

SALTON, G. **Automatic information, organization and retrieval**. New York: McGraw-Hill, 1968 *apud* SAYÃO, L. F. **SALF**: um algoritmo para indexação automática utilizando

vocabulário controlado. 1985. 177 f. Dissertação (Mestrado em Ciência da Informação) – Universidade Federal do Rio de Janeiro, IBICT, Rio de Janeiro, 1985.

SALTON, G. Automatic indexing using bibliographic citations. **Journal of Documentation**, v. 27, n. 2, p. 98-110, jun. 1971a. *apud* SAYÃO, L. F. **SALF**: um algoritmo para indexação automática utilizando vocabulário controlado. 1985. 177 f. Dissertação (Mestrado em Ciência da Informação) – Universidade Federal do Rio de Janeiro, IBICT, Rio de Janeiro, 1985.

SALTON, G. **The SMART retrieval systems**: experiments in automatic document processing. New York: Prentice-Hall, Englewood Cliffs, 1971b. *apud* SAYÃO, L. F. **SALF**: um algoritmo para indexação automática utilizando vocabulário controlado. 1985. 177 f. Dissertação (Mestrado em Ciência da Informação) – Universidade Federal do Rio de Janeiro, IBICT, Rio de Janeiro, 1985.

SALTON, G. A new comparison between conventional indexing and automatic text processing. **Journal of the American Society for Information Science**, v. 23, n. 2, p. 75-84, 1972 *apud* SAYÃO, L. F. **SALF**: um algoritmo para indexação automática utilizando vocabulário controlado. 1985. 177 f. Dissertação (Mestrado em Ciência da Informação) – Universidade Federal do Rio de Janeiro, IBICT, Rio de Janeiro, 1985.

SALTON, G. Automatic indexing: a summary. In: EURIM: CONFÉRENCE EUROPÉENNE SUR LA RECHERCHE DANS L'ORGANISATION DES SERVICES D'INFORMATION ET DES BIBLIOTHÈQUES, 5., 1982, Versailles. **Proceedings...** London: Aslib, 1982. p. 66-77. *apud* SAYÃO, L. F. **SALF**: um algoritmo para indexação automática utilizando vocabulário controlado. 1985. 177 f. Dissertação (Mestrado em Ciência da Informação) – Universidade Federal do Rio de Janeiro, IBICT, Rio de Janeiro, 1985.

SALTON, G. The measurement of the term importance in automatic indexing. **Journal of the American Society for Information Science***,* v. 32, n. 3, p. 175-186. may 1981 *apud* SAYÃO, L. F. **SALF**: um algoritmo para indexação automática utilizando vocabulário controlado. 1985. 177 f. Dissertação (Mestrado em Ciência da Informação) – Universidade Federal do Rio de Janeiro, IBICT, Rio de Janeiro, 1985.

SALTON, G.; LESK, M. E. Computer evaluation of indexing and text processing. **Journal of the ACM**, v. 15, n. 1, p. 8-36, jan. 1968 *apud* SAYÃO, L. F. **SALF**: um algoritmo para indexação automática utilizando vocabulário controlado. 1985. 177 f. Dissertação (Mestrado em Ciência da Informação) – Universidade Federal do Rio de Janeiro, IBICT, Rio de Janeiro, 1985.

SALTON, G., McGILL, M. **Introduction to Modern Information Retrieval**. McGraw-Hill Book Co.: New York, 1983 *apud* SAYÃO, L. F. **SALF**: um algoritmo para indexação automática utilizando vocabulário controlado. 1985. 177 f. Dissertação (Mestrado em Ciência da Informação) – Universidade Federal do Rio de Janeiro, IBICT, Rio de Janeiro, 1985.

SALTON, G.; WONG, A., YANG, C. A vector space model for automatic indexing. **Communications of the ACM***,* v. 18, n. 11, p. 613-620, 1975 *apud* SAYÃO, L. F. **SALF**: um algoritmo para indexação automática utilizando vocabulário controlado. 1985. 177 f. Dissertação (Mestrado em Ciência da Informação) – Universidade Federal do Rio de Janeiro, IBICT, Rio de Janeiro, 1985.

SALTON, G.; YANG, G. S.; YU, C. T. A Theory of term importance in automatic text analysis. **Journal of the American Society for Information Science**, v. 26, n. 1, p. 33-44, jan./fev. 1975 *apud* SAYÃO, L. F. **SALF**: um algoritmo para indexação automática utilizando vocabulário controlado. 1985. 177 f. Dissertação (Mestrado em Ciência da Informação) – Universidade Federal do Rio de Janeiro, IBICT, Rio de Janeiro, 1985.

SANTOS, B. S. **Um discurso sobre as ciências**. Porto: Afrontamento, 1996.

SANTOS, D. Caminhos percorridos no mapa da portuguesificação: a Linguateca em perspectiva. **Linguamática**, v. 1, n. 1, 2009, p. 25-59. Disponível em: <http://linguamatica.com/index.php/linguamatica/article/viewFile/20/9>.

SANTOS, D.; SARMENTO, L. O projecto AC/DC: acesso a corpora/disponibilização de corpora. In: MENDES, A.; FREITAS, T. (Ed.). Encontro Nacional da Associação Portuguesa de Linguística, 18., 2002, Lisboa. **Actas...** Lisboa: APL, 2002. p. 705-717. Disponível em: <http://www.linguateca.pt/documentos/SantosSarmentoAPL2002.pdf>. Acesso em: 22 maio 2012.

SARACEVIC, T. Ciência da informação: origem, evolução e relações. **Perspectivas em Ciência da Informação.** Belo Horizonte, v.1, n.1, p. 41-62, jan./jun. 1996.

SARDINHA, T. B. **Lingüística de Corpus**. Barueri, SP: Manole, 2004. 410p.

SAYÃO, L. F. **SALF**: um algoritmo para indexação automática utilizando vocabulário controlado. 1985. 177 f. Dissertação (Mestrado em Ciência da Informação) – Universidade Federal do Rio de Janeiro, IBICT, Rio de Janeiro, 1985.

SCHULTZ, C. K. **H. P. Luhn**: pionner of information science – selected works. New York: American Documentation Institute, Spartan Books, 1968.

SHAH, P. K. *et al.* Information extraction from full text scientific articles: where are the keywords?. **BMC Bioinfomatics,** v. 4, n. 20, 2003.

SILVA, B. **Origem e evolução do descritor**. Rio de Janeiro: Fundação Getúlio Vargas, 1972.

SOUZA, R. R. **Uma proposta de metodologia para escolha automática de descritores utilizando sintagmas nominais**. 2005. 197 f. Tese (Doutorado em Ciência da Informação) – Escola de Ciência da Informação, Universidade Federal de Minas Gerais, Belo Horizonte, 2005.

SOUZA, R. R.; RAGHAVAN, K. S. **A methodology for noun phrase-based automatic indexing**. A ser editado, 2006.

SPARCK JONES, K. A statistical interpretation of term specificity and its application to retrieval. **Journal of Documentation**, v. 28, n. 1, p. 11-20, 1972 *apud* SAYÃO, L. F. **SALF**: um algoritmo para indexação automática utilizando vocabulário controlado. 1985. 177 f. Dissertação (Mestrado em Ciência da Informação) – Universidade Federal do Rio de Janeiro, IBICT, Rio de Janeiro, 1985.

SPARCK JONES, K. Collection properties influencing automatic term classification performance. **Information Storage and Retrieval**, v. 9, p. 499-513, 1973 *apud* SAYÃO, L.

F. **SALF**: um algoritmo para indexação automática utilizando vocabulário controlado. 1985. 177 f. Dissertação (Mestrado em Ciência da Informação) – Universidade Federal do Rio de Janeiro, IBICT, Rio de Janeiro, 1985.

SPARCK JONES, K. The role of automatic indexing in oprational on-line retrieval systems. In: FID CONGRES, 38, Edinburg, 1978. **Proceedings...** London: ASLIB, 1980, p. 33-38 *apud* SAYÃO, L. F. **SALF**: um algoritmo para indexação automática utilizando vocabulário controlado. 1985. 177 f. Dissertação (Mestrado em Ciência da Informação) – Universidade Federal do Rio de Janeiro, IBICT, Rio de Janeiro, 1985.

SPARCK JONES, K. Experiments in relevance weighting of search terms. **Information Processing & Management**, v. 15, n.13, p. 133-144, 1979 *apud* SAYÃO, L. F. **SALF**: um algoritmo para indexação automática utilizando vocabulário controlado. 1985. 177 f. Dissertação (Mestrado em Ciência da Informação) – Universidade Federal do Rio de Janeiro, IBICT, Rio de Janeiro, 1985.

STOKOLOV, N. V. On automatic support to indexing a life science data base. **Information Processing & Management**, v. 18, n. 6, p. 313-321, 1982 *apud* SAYÃO, L. F. **SALF**: um algoritmo para indexação automática utilizando vocabulário controlado. 1985. 177 f. Dissertação (Mestrado em Ciência da Informação) – Universidade Federal do Rio de Janeiro, IBICT, Rio de Janeiro, 1985.

SWANSON, D. R. Library goals and the role of automation. **Spec. Libraries**, v. 53, p. 466-71, 1962 *apud* SAYÃO, L. F. **SALF**: um algoritmo para indexação automática utilizando vocabulário controlado. 1985. 177 f. Dissertação (Mestrado em Ciência da Informação) – Universidade Federal do Rio de Janeiro, IBICT, Rio de Janeiro, 1985.

SWANSON, D. R. Automation indexing and classification. In: NATO ADVANCED STUDY INSTITUTE ON AUTOMATIC ANALYSIS, 1963, Venice. **Proceedings...** New York: [s.n.], 1963. p. 125-128 *apud* SAYÃO, L. F. **SALF**: um algoritmo para indexação automática utilizando vocabulário controlado. 1985. 177 f. Dissertação (Mestrado em Ciência da Informação) – Universidade Federal do Rio de Janeiro, IBICT, Rio de Janeiro, 1985.

TRACHTENBERG, A. Automatic document classification using information theoretical methods. In: LUHN, H. P. (Ed.) **Automation and Scientific Communication**. [s.l.]: [s.n.], 1963. p. 349-50 *apud* SAYÃO, L. F. **SALF**: um algoritmo para indexação automática utilizando vocabulário controlado. 1985. 177 f. Dissertação (Mestrado em Ciência da Informação) – Universidade Federal do Rio de Janeiro, IBICT, Rio de Janeiro, 1985.

TRASK, R. L. **Dicionário de Linguagem e Lingüística**. São Paulo: Contexto, 2004. 364 p.

UNIVERSIDADE FEDERAL DE MINAS GERAIS. **Biblioteca de Teses e Dissertações da UFMG**. Belo Horizonte. Disponível em: < http://www.bibliotecadigital.ufmg.br/dspace/browse-date>. Acesso em novembro de 2011.

VAN DER MEULEN, W. A.; JANSEN, P. J. F. C. Automatic versus manual indexing. **Information Processing and Management**. v. 13, n. 1, p. 13-21, 1977 *apud* SAYÃO, L. F. **SALF**: um algoritmo para indexação automática utilizando vocabulário controlado. 1985. 177 f. Dissertação (Mestrado em Ciência da Informação) – Universidade Federal do Rio de Janeiro, IBICT, Rio de Janeiro, 1985.

VAN RIJSBERGEN, C. J. A theoretical basis for the use of co-ocurrence data in information retrieval. **Journal of Documentation**, v. 27, n. 2, p. 69-82, jun. 1971 *apud* SAYÃO, L. F.

SALF: um algoritmo para indexação automática utilizando vocabulário controlado. 1985. 177 f. Dissertação (Mestrado em Ciência da Informação) – Universidade Federal do Rio de Janeiro, IBICT, Rio de Janeiro, 1985.

VAN RIJSBERGEN, C. J. **Information Retrieval**. London: Butterwords, 1979 *apud* SAYÃO, L. F. **SALF**: um algoritmo para indexação automática utilizando vocabulário controlado. 1985. 177 f. Dissertação (Mestrado em Ciência da Informação) – Universidade Federal do Rio de Janeiro, IBICT, Rio de Janeiro, 1985.

VON STAA, A. PRAXPAL: um indexador semi-automático. In: CONGRESSO DA SOCIEDADE BRASILEIRA DE COMPUTAÇÂO, 3., 1983, Campinas. **Anais...** São Paulo: [s.n.], 1983. p. 433-443.

WERSIG, G. Information science: the study of postmodern knowledge usage. **Information Processing & Management**. New York, v. 29, n. 2, p. 229-239, 1993.

WIUPS. *World Internet Users and Population Stats*, 2011. Disponível em <http://www.internetworldstats.com/stats.htm>. Acessado em novembro de 2011.

YU, C. T.; SALTON, G. Precision weighting: an effective indexing method. **Journal of Association for Computing Machiminery**, v. 23, p. 76-88, 1976.

ZAHER, C. L. *et al*. Automação da informação em Física no Brasil. In: SEMINÁRIO SOBRE INFORMÁTICA, 1968, Rio de Janeiro. **Anais...** Rio de Janeiro: IBBD, 1969, p. 39-52.

ZIPF, G. K. **Selected studies of the principle of relative frequency in language**. Cambridge, USA: Havard University Press, 1932.

APÊNDICE A - QUANTIDADE DE TESES NA BIBLIOTECA DE TESES E DISSERTAÇÕES DA UFMG

Programa CAPES de Pós-Graduação	Quantidade de Teses
Pós-Graduação em Educação: Conhecimento e Inclusão Social	214
Pós-Graduação em Ciência Animal	128
Pós-Graduação em Letras: Estudos Literários	105
Pós-Graduação em Engenharia Metalúrgica e de Minas	91
Pós-Graduação em Estudos Lingüísticos	90
Pós-Graduação em Engenharia Elétrica	88
Pós-Graduação em Química	76
Pós-Graduação em Física	75
Pós-Graduação em Ciência da Computação	72
Pós-Graduação em Bioquímica e Imunologia	61
Pós-Graduação em Ciência da Informação	58
Pós-Graduação em Medicina (Pediatria)	56
Pós-Graduação em Demografia	46
Pós-Graduação em Parasitologia	43
Pós-Graduação em Odontologia	39
Pós-Graduação em Zootecnia	35
Pós-Graduação em Engenharia Mecânica	34
Pós-Graduação em Cirurgia	33
Pós-Graduação em Saneamento Meio Ambiente e Recursos Hídricos	31
Pós-Graduação em Geografia	30
Pós-Graduação em Ciências Biológicas (Fisiologia e Farmacologia)	29
Pós-Graduação em Saúde Pública	28
Pós-Graduação em Engenharia de Estruturas	27
Pós-Graduação em Ciências Farmacêuticas	27
Pós-Graduação em Economia	26
Pós-Graduação em Ciências Biológicas (Farmacologia Bioquímica e Molecular)	23
Pós-Graduação em História	21
Pós-Graduação em Clínica Médica	21
Pós-Graduação em Medicina (Medicina Tropical)	21
Pós-Graduação em Filosofia	20

Programa CAPES de Pós-Graduação	Quantidade de Teses
Pós-Graduação em Geologia	19
Pós-Graduação em Ciência de Alimentos	18
Pós-Graduação em Patologia	18
Pós-Graduação em Medicina (Ginecologia e Obstetrícia)	18
Pós-Graduação em Enfermagem	17
Pós-Graduação em Artes Visuais	16
Pós-Graduação em Matemática	16
Pós-Graduação em Ciência Política	14
Pós-Graduação em Biologia Celular	14
Pós-Graduação em Oftalmologia	14
Pós-Graduação em Administração	13
Pós-Graduação em Bioinformática	13
Pós-Graduação em Medicina (Gastroenterologia)	12
Pós-Graduação em Sociologia e Política	12
Pós-Graduação em Sociologia	11
Pós-Graduação em Ciências Biológicas (Microbiologia)	10
Pós-Graduação em Estatística	9
Pós-Graduação em Comunicação Social	8
Pós-Graduação em Direito	7
Pós-Graduação em Engenharia Química	4
Pós-Graduação em Biologia Vegetal	4
Pós-Graduação em Ecologia (Conservação e Manejo da Vida Silvestre)	4
Pós-Graduação em Educação Física	1
Pós-Graduação em Ciências e Técnicas Nucleares	1
Pós-Graduação em Medicina Veterinária	-
Pós-Graduação em Música	-
Pós-Graduação em Arquitetura e Urbanismo	-
Pós-Graduação em Psicologia	-
Pós-Graduação em Engenharia de Produção	-
Pós-Graduação em Ciências da Reabilitação	-
Pós-Graduação em Construção Civil	-
Pós-Graduação em Ciências Agrárias	-
Programa CAPES não informado	-

Programa CAPES de Pós-Graduação	Quantidade de Teses
Pós-Graduação em Análise e Modelagem de Sistemas Ambientais	-
Pós-Graduação em Genética	-
Pós-Graduação em Antropologia	-
Pós-Graduação em Ambiente Construído e Patrimônio Sustentável	-
Pós-Graduação em Ciências Contábeis	-
TOTAL	**1.921**

Fonte: adaptado de BIBLIOTECA DIGITAL DA UFMG, 2012[49].

[49] BIBLIOTECAS DE TESES E DISSERTAÇÕES DA UFMG. Disponível em: <http://www.bibliotecadigital.ufmg.br/dspace/browse-date>. Acesso em: 31 mar. 2012.

APÊNDICE B - EXEMPLO DE E-MAIL ENVIADO PARA OS AUTORES SOLICITANDO SUA PARTICIPAÇÃO NA PESQUISA

Título do E-mail: Participação em Pesquisa - Medicina (Pediatria)

Cara Márcia ,

Sou aluno de mestrado da Escola de Ciência da Informação na UFMG. Meu tema de pesquisa é em Indexação Automática. Pretendo usar sua tese no meu Corpus de pesquisa. Gostaria de saber se você tem interesse em participar. Para isso será necessário apenas a sua escolha entre duas listagens de palavras-chaves geradas automaticamente com base no texto da sua tese de doutorado em Medicina (Pediatria) publicada na Biblioteca Digital da UFMG .

Caso tenha interesse, retornarei sua mensagem as duas listagens (de aproximadamente 10 termos cada) para a sua escolha de qual representa melhor sua tese.

Obrigado desde já!

Atenciosamente,

Luiz Mesquita.

O conteúdo desta mensagem é de responsabilidade do seu remetente e a solicitação de envio foi realizada através de opção disponível no Currículo Lattes.
Informações de envio:
Remetente: Luiz Antônio Lopes Mesquita
E-mail: <retirado>
Data/Hora: 28/05/2012 16:39:51
Endereço IP de Origem: 177.19.22.132

APÊNDICE C - LISTA DOS TERMOS RETIRADOS (*STOPWORDS*) NO PROCESSO DE LIMPEZA DOS SINTAGMAS NOMINAIS EXTRAÍDOS PELO OGMA

a	às	dessas	isso	nesses
a ela	assim	dessas	isto	nesta
a elas	através	desse	já	nesta
a ele	but	desse	maior	nesta
a eles	cada	desses	maiores	nestas
ainda	certa	desses	maioria	nestas
além disso	certo	desta	mais	nestas
algo	como	desta	menos	neste
algum	considerada	destas	mesma	neste
algum	consideradas	destas	mesmas	neste
alguma	considerado	deste	mesmo	nestes
alguma	considerados	deste	mesmos	nestes
algumas	da	destes	meu	nestes
algumas	dados	destes	meus	NN
alguns	daquela	dito	minha	NN%
alguns	daquela	diversas	minhas	NNN
ambas	daquelas	diversos	muita	NNN%
ambos	daquelas	do	muitas	NNNN
and	daquele	dos	muito	NNNN%
apenas	daquele	duas	muitos	no
apesar da	daqueles	ela	N	nos
apesar das	daqueles	elas	N%	nossa
apesar do	das	ele	na	nossas
apesar dos	de	eles	não	nosso
apud	dela	enquanto	nas	nossos
aquela	dela	entanto	nenhum	o
àquela	delas	então	nenhuma	os
aquelas	delas	essa	nenhumas	outra
àquelas	dele	essas	nenhuns	outras
aquele	dele	esse	nessa	outro
àquele	deles	esses	nessa	outros
aqueles	deles	esta	nessas	pouca
àqueles	dentre	estas	nessas	poucas
aqui	dentre	este	nesse	pouco
as	dessa	estes	nesse	poucos
às	dessa	fim	nesses	próprio

próprios	se	sua	também	umas
quais	seu	sua	tanto	uns
quaisquer	seu	suas	the	várias
qual	seu	suas	toda	várias
qualquer	seus	suas	todas	vários
quando	seus	tais	todo	vários
quase	seus	tais	todos	vez
que	sobretudo	tal	um	vezes
que	sua	talvez	uma	

Obs.: “N” corresponde a um dígito qualquer de 0 a 9.

APÊNDICE D - MACRO DO MICROSOFT OFFICE WORD 2007 PARA LIMPEZA DOS SINTAGMAS NOMINAIS EXTRAÍDOS PELO OGMA

```
Sub AbrirLimparSalvarComoSl()

Dim Grupo
Dim ItensGrupo
Dim ItensStopWords
Dim ItensLimpaInicio1
Dim ItensLimpaInicio2
Dim ItensLimpaInicio3
Dim ItensFalsosSintagmas
Dim ItensLimpaNumeros

ItensLimpaInicio1 = Array("^p^#%^p", "^p^#^#%^p", "^p^#^#^#%^p", "^p^#^#^#^#%^p", "^p^#^#^#^#^p", "^p^#^#^#^p", "^p^#^#^p", "^p^#^p", "^pa ", "^pa ela ", "^pa elas ", "^pa ele ", "^pa eles ", "^palgo ", "^palgum ", "^palguma ", "^palgumas ", "^palguns ", "^pambas ", "^pambas ", "^pambos ", "^pambos ", "^papenas ", "^papesar da ", "^papesar das ", "^papesar do ", "^papesar dos ", "^paquela ", "^paquela ", "^pàquela ", "^pàquela ", "^paquelas ", "^paquelas ", "^pàquelas ", "^pàquelas ", "^paquele ", "^paquele ", "^pàquele ", "^pàquele ", "^paqueles ", "^paqueles ", "^pàqueles ", "^pàqueles ", "^paqui ", "^pas ", "^pàs ", "^pàs ", "^passim ", "^pcada ", "^pcerta ", "^pcerto ", "^pcomo ", "^pconsiderada ", "^pconsideradas ", "^pconsiderado ", "^pconsiderados ", "^pda ", "^pda ", "^pdaquela ", "^pdaquelas ", "^pdaquele ", "^pdaqueles ", "^pdas ", "^pdas ", "^pde ", "^pde ", "^pdela ", "^pdela ", "^pdelas ", "^pdelas ", "^pdele ", "^pdele ", "^pdeles ", "^pdeles ", "^pdentre ")
ItensLimpaInicio2 = Array("^pdentre ", "^pdessa ", "^pdessas ", "^pdesse ", "^pdesses ", "^pdesta ", "^pdestas ", "^pdeste ", "^pdestes ", "^pdo ", "^pdo ", "^pdos ", "^pdos ", "^pela ", "^pelas ", "^pele ", "^peles ", "^penquanto ", "^pentão ", "^pessa ", "^pessas ", "^pesse ", "^pesses ", "^pesta ", "^pestas ", "^peste ", "^pestes ", "^pisso ", "^pisto ", "^pjá ", "^pjá ", "^pmaior ", "^pmaiores ", "^pmais ", "^pmenor ", "^pmenores ", "^pmenos ", "^pmesma ", "^pmesmas ", "^pmesmo ", "^pmesmos ", "^pmeu ", "^pmeus ", "^pminha ", "^pminhas ", "^pmuita ", "^pmuitas ", "^pmuito ", "^pmuitos ", "^pna ", "^pna ", "^pnão ", "^pnas ", "^pnas ", "^pnenhum ", "^pnenhuma ", "^pnenhumas ", "^pnenhuns ", "^pnessa ", "^pnessa ", "^pnessas ", "^pnessas ", "^pnesse ", "^pnesse ", "^pnesses ", "^pnesses ", "^pnesta ", "^pnesta ", "^pnesta ", "^pnestas ", "^pnestas ", "^pnestas ", "^pneste ", "^pneste ", "^pneste ", "^pnestes ", "^pnestes ")
ItensLimpaInicio3 = Array("^pnestes ", "^pno ", "^pno ", "^pnos ", "^pnos ", "^pnossa ", "^pnossas ", "^pnosso ", "^pnossos ", "^po ", "^pos ", "^poutra ", "^poutras ", "^poutro ", "^poutros ", "^ppouca ", "^ppoucas ", "^ppouco ", "^ppoucos ", "^ppróprio ", "^ppróprios ", "^pquaisquer ", "^pqualquer ", "^pquando ", "^pquase ", "^pse ", "^pseu ", "^pseu ", "^pseus ", "^pseus ", "^psua ", "^psua ", "^psuas ", "^psuas ", "^ptais ", "^ptais ", "^ptal ", "^ptal ", "^ptalvez ", "^ptambém ", "^ptanto ", "^ptanto ", "^ptoda ", "^ptodas ", "^ptodo ", "^ptodos ", "^pum ", "^puma ", "^pumas ", "^puns ", "^pvárias ", "^pvárias ", "^pvários ", "^pvários ")
ItensFalsosSintagmas = Array("^palém disso^p", "^palgum^p", "^palguma^p", "^palgumas^p", "^palguns^p", "^pand^p", "^papud^p", "^pàs^p", "^patravés^p", "^pbut^p", "^pdaquela^p", "^pdaquelas^p", "^pdaquele^p", "^pdaqueles^p", "^pdela^p", "^pdelas^p", "^pdele^p", "^pdeles^p", "^pdentre^p", "^pdessa^p", "^pdessas^p", "^pdesse^p", "^pdesses^p", "^pdesta^p", "^pdestas^p", "^pdeste^p", "^pdestes^p", "^pdito^p", "^pduas^p", "^pentanto^p", "^pfim^p", "^pnessa^p", "^pnessas^p", "^pnesse^p", "^pnesses^p", "^pnesta^p", "^pnestas^p", "^pneste^p", "^pnestes^p", "^pque^p", "^psobretudo^p", "^ptais^p", "^pthe^p", "^pvárias^p", "^pvários^p", "^pvez^p", "^pvezes^p")
ItensLimpaNumeros = Array("^p^#%^p", "^p^#^#%^p", "^p^#^#^#%^p", "^p^#^#^#^#%^p", "^p^#^#^#^#^p", "^p^#^#^#^p", "^p^#^#^p", "^p^#^p")

'Grupo = "A"
'ItensGrupo = Array("01", "02", "03", "04", "05", "06", "07", "08", "09", "10", "11", "12", "13", "14", "15", "16", "17", "18", "19", "20", "21", "22", "23", "24")

For Each Iten In ItensGrupo     ' Itere através de cada elemento.
```

```
    Documents.Open FileName:="G:\UFMG\ECI\DISSERTAÇÃO\CORPUS\Teses\" & Grupo & "\" &
Grupo & "" & Iten & "-s.txt", ConfirmConversions:=False, ReadOnly _
        :=False, AddToRecentFiles:=False, PasswordDocument:="", PasswordTemplate _
        :="", Revert:=False, WritePasswordDocument:="", WritePasswordTemplate:="" _
        , Format:=wdOpenFormatAuto, XMLTransform:="", Encoding:=1252

    Selection.Find.ClearFormatting
    Selection.Find.Replacement.ClearFormatting

 For Each Item In ItensLimpaInicio1
    With Selection.Find
        .Text = Item
        .Replacement.Text = "^p"
        .Forward = True
        .Wrap = wdFindContinue
        .Format = False
        .MatchCase = False
        .MatchWholeWord = False
        .MatchWildcards = False
        .MatchSoundsLike = False
        .MatchAllWordForms = False
    End With
    Selection.Find.Execute Replace:=wdReplaceAll
 Next Iten

 For Each Item In ItensLimpaInicio2
    With Selection.Find
        .Text = Item
        .Replacement.Text = "^p"
        .Forward = True
        .Wrap = wdFindContinue
        .Format = False
        .MatchCase = False
        .MatchWholeWord = False
        .MatchWildcards = False
        .MatchSoundsLike = False
        .MatchAllWordForms = False
    End With
    Selection.Find.Execute Replace:=wdReplaceAll
 Next Iten

 For Each Item In ItensLimpaInicio3
    With Selection.Find
        .Text = Item
        .Replacement.Text = "^p"
        .Forward = True
        .Wrap = wdFindContinue
        .Format = False
        .MatchCase = False
        .MatchWholeWord = False
        .MatchWildcards = False
        .MatchSoundsLike = False
        .MatchAllWordForms = False
    End With
    Selection.Find.Execute Replace:=wdReplaceAll
 Next Iten

 For Each Item In ItensLimpaNumeros
    With Selection.Find
        .Text = Item
```

```
        .Replacement.Text = "^p"
        .Forward = True
        .Wrap = wdFindContinue
        .Format = False
        .MatchCase = False
        .MatchWholeWord = False
        .MatchWildcards = False
        .MatchSoundsLike = False
        .MatchAllWordForms = False
    End With
    Selection.Find.Execute Replace:=wdReplaceAll
Next Iten

For Each Item In ItensFalsosSintagmas
    With Selection.Find
        .Text = Item
        .Replacement.Text = "^p"
        .Forward = True
        .Wrap = wdFindContinue
        .Format = False
        .MatchCase = False
        .MatchWholeWord = False
        .MatchWildcards = False
        .MatchSoundsLike = False
        .MatchAllWordForms = False
    End With
    Selection.Find.Execute Replace:=wdReplaceAll
Next Iten


    ActiveDocument.SaveAs FileName:="G:\UFMG\ECI\DISSERTAÇÃO\CORPUS\Teses\" & Grupo &
"\" & Grupo & "" & Iten & "-sl.txt", FileFormat:=wdFormatText, _
        LockComments:=False, Password:="", AddToRecentFiles:=True, WritePassword _
        :="", ReadOnlyRecommended:=False, EmbedTrueTypeFonts:=False, _
        SaveNativePictureFormat:=False, SaveFormsData:=False, SaveAsAOCELetter:= _
        False, Encoding:=1252, InsertLineBreaks:=False, AllowSubstitutions:=False _
        , LineEnding:=wdCRLF
    ActiveDocument.Close

Next Iten


End Sub
```

APÊNDICE E - MACROS DO MICROSOFT OFFICE EXCEL 2007 PARA A DETERMINAÇÃO DOS SINTAGMAS NOMINAIS COMO CANDIDATOS A DESCRITORES

```
Public LocalizacaoGrupo As String
Public Grupo As String
Public ItensGrupo As Variant
Public TamanhoGrupo As Integer

Sub ImportarSL()

For Each Item In ItensGrupo

    Sheets.Add After:=Sheets(Sheets.Count)
    Sheets(Sheets.Count).Select
    Sheets(Sheets.Count).Name = Grupo & Item & "-sl"
    Sheets(Grupo & Item & "-sl").Select
    With ActiveSheet.QueryTables.Add(Connection:= _
        "TEXT;" & LocalizacaoGrupo & Grupo & "\" & Grupo & Item & "-sl.txt", Destination:=Range( _
        "$A$1"))
        .Name = Grupo & Item & "-sl"
        .FieldNames = True
        .RowNumbers = False
        .FillAdjacentFormulas = False
        .PreserveFormatting = True
        .RefreshOnFileOpen = False
        .RefreshStyle = xlInsertDeleteCells
        .SavePassword = False
        .SaveData = True
        .AdjustColumnWidth = True
        .RefreshPeriod = 0
        .TextFilePromptOnRefresh = False
        .TextFilePlatform = 1252
        .TextFileStartRow = 1
        .TextFileParseType = xlDelimited
        .TextFileTextQualifier = xlTextQualifierDoubleQuote
        .TextFileConsecutiveDelimiter = False
        .TextFileTabDelimiter = True
        .TextFileSemicolonDelimiter = False
        .TextFileCommaDelimiter = False
        .TextFileSpaceDelimiter = False
        .TextFileColumnDataTypes = Array(1)
        .TextFileTrailingMinusNumbers = True
        .Refresh BackgroundQuery:=False
    End With

    Sheets(Grupo & Item & "-sl").Select
    Columns("A:A").Select
    Selection.Insert Shift:=xlToRight, CopyOrigin:=xlFormatFromLeftOrAbove
    Rows("1:1").Select
    Selection.Insert Shift:=xlDown, CopyOrigin:=xlFormatFromLeftOrAbove
    Range("A1").FormulaR1C1 = "Posicao"
    Range("B1").FormulaR1C1 = "Sintagma"
    Range("A2").FormulaR1C1 = "1"
    Range("A3").FormulaR1C1 = "2"
    Range("A2:A3").AutoFill Destination:=Range("A2:A" & Range("B1",
ActiveSheet.Range("B1048576").End(xlUp)).Count)
```

```
Next Item


End Sub
Sub ImportarTRAL()

For Each Item In ItensGrupo

    Sheets.Add After:=Sheets(Sheets.Count)
    Sheets(Sheets.Count).Select
    Sheets(Sheets.Count).Name = Grupo & Item & "-tral"
    Sheets(Grupo & Item & "-tral").Select
    With ActiveSheet.QueryTables.Add(Connection:= _
        "TEXT;" & LocalizacaoGrupo & Grupo & "\" & Grupo & Item & "-tral.txt", Destination:=Range( _
        "$A$1"))
        .Name = Grupo & Item & "-tral"
        .FieldNames = True
        .RowNumbers = False
        .FillAdjacentFormulas = False
        .PreserveFormatting = True
        .RefreshOnFileOpen = False
        .RefreshStyle = xlInsertDeleteCells
        .SavePassword = False
        .SaveData = True
        .AdjustColumnWidth = True
        .RefreshPeriod = 0
        .TextFilePromptOnRefresh = False
        .TextFilePlatform = 1252
        .TextFileStartRow = 1
        .TextFileParseType = xlDelimited
        .TextFileTextQualifier = xlTextQualifierDoubleQuote
        .TextFileConsecutiveDelimiter = False
        .TextFileTabDelimiter = True
        .TextFileSemicolonDelimiter = False
        .TextFileCommaDelimiter = False
        .TextFileSpaceDelimiter = False
        .TextFileOtherDelimiter = "/"
        .TextFileColumnDataTypes = Array(1, 1, 1, 1)
        .TextFileTrailingMinusNumbers = True
        .Refresh BackgroundQuery:=False
    End With

Next Item


End Sub
Sub LimpaSintagmaErroSlxTral()

    Sheets.Add After:=Sheets(Sheets.Count)
    Sheets(Sheets.Count).Name = Grupo & ".erros"


For Each Item In ItensGrupo


    Sheets(Grupo & Item & "-sl").Select
    Range("C1").FormulaR1C1 = "CSN"
    Range("C2").FormulaR1C1 = "=VLOOKUP(RC[-1],'" & Grupo & Item & "-tral'!C[-2]:C[1],4,FALSE)"
    Range("C2").AutoFill Destination:=Range("C2:C" & Range("B1",
ActiveSheet.Range("B1048576").End(xlUp)).Count)
```

```
    Range("D1").FormulaR1C1 = "ERRO"
    Range("D2").FormulaR1C1 = "=ISERROR(RC[-1])"
    Range("D2").AutoFill Destination:=Range("D2:D" & Range("B1",
ActiveSheet.Range("B1048576").End(xlUp)).Count)
    Columns("A:D").Select
    ActiveWorkbook.Worksheets(Grupo & Item & "-sl").Sort.SortFields.Clear
    ActiveWorkbook.Worksheets(Grupo & Item & "-sl").Sort.SortFields.Add Key:=Range("C2:C" &
Range("B1", ActiveSheet.Range("B1048576").End(xlUp)).Count _
        ), SortOn:=xlSortOnValues, Order:=xlDescending, DataOption:=xlSortNormal
    ActiveWorkbook.Worksheets(Grupo & Item & "-sl").Sort.SortFields.Add Key:=Range("A2:A" &
Range("B1", ActiveSheet.Range("B1048576").End(xlUp)).Count _
        ), SortOn:=xlSortOnValues, Order:=xlAscending, DataOption:=xlSortNormal
    With ActiveWorkbook.Worksheets(Grupo & Item & "-sl").Sort
        .SetRange Range("A1:D" & Range("B1", ActiveSheet.Range("B1048576").End(xlUp)).Count)
        .Header = xlYes
        .MatchCase = False
        .Orientation = xlTopToBottom
        .SortMethod = xlPinYin
        .Apply
    End With

    Range("E2").FormulaR1C1 = "=COUNTIF(C[-1],TRUE)"
    QuantidadeErro = Range("E2").Value
    Sheets(Grupo & ".Corpus").Range("I" & Item + 1).Value = QuantidadeErro
    Columns("D:E").Delete Shift:=xlToLeft

    Sheets(Grupo & Item & "-sl").Select
    Rows("2:" & QuantidadeErro + 1).Cut

    Sheets(Grupo & ".erros").Select
    PrimeiraLinhaLivre = Range("B1", ActiveSheet.Range("B1048576").End(xlUp)).Count + 1
    Range("A" & PrimeiraLinhaLivre).Select
    ActiveSheet.Paste
    Application.CutCopyMode = False
    Range("A" & PrimeiraLinhaLivre & ":A" & PrimeiraLinhaLivre + QuantidadeErro - 1).Value = Grupo
& Item

    Sheets(Grupo & Item & "-sl").Select
    Columns("A:C").Select
    ActiveWorkbook.Worksheets(Grupo & Item & "-sl").Sort.SortFields.Clear
    ActiveWorkbook.Worksheets(Grupo & Item & "-sl").Sort.SortFields.Add Key:=Range("A2:A" &
Range("B1", ActiveSheet.Range("B1048576").End(xlUp)).Count _
        ), SortOn:=xlSortOnValues, Order:=xlAscending, DataOption:=xlSortNormal
    With ActiveWorkbook.Worksheets(Grupo & Item & "-sl").Sort
        .SetRange Range("A1:C" & Range("B1", ActiveSheet.Range("B1048576").End(xlUp)).Count)
        .Header = xlYes
        .MatchCase = False
        .Orientation = xlTopToBottom
        .SortMethod = xlPinYin
        .Apply
    End With
    Range("A2").FormulaR1C1 = "1"
    Range("A3").FormulaR1C1 = "2"
    Range("A2:A3").AutoFill Destination:=Range("A2:A" & Range("B1",
ActiveSheet.Range("B1048576").End(xlUp)).Count)

    Columns("C:C").Select
    Selection.Delete Shift:=xlToLeft

Next Item
```

```
    Sheets(Grupo & ".erros").Select
    Columns("C:C").Delete Shift:=xlToLeft
    Columns("A:B").EntireColumn.AutoFit

End Sub

Sub RetiraMarcaMeioFim()


    Marca = "lambori"

For Each Item In ItensGrupo

    Sheets(Grupo & Item & "-sl").Select
    Set CelulaMarca = ActiveSheet.Columns.Find(Marca, LookAt:=xlPart, LookIn:=xlValues)
    If Not CelulaMarca Is Nothing Then
      'Else
          'CelulaMarca.Select
          Sheets(Grupo & ".Corpus").Range("J" & Item + 1).FormulaR1C1 = CelulaMarca.Row - 1
           Sheets(Grupo & Item & "-sl").Rows(CelulaMarca.Row & ":" & CelulaMarca.Row).Delete
Shift:=xlUp
          Set CelulaMarca = ActiveSheet.Columns.Find(Marca, LookAt:=xlPart, LookIn:=xlValues)
             If Not CelulaMarca Is Nothing Then
                'CelulaMarca.Select
                Sheets(Grupo & ".Corpus").Range("K" & Item + 1).FormulaR1C1 = CelulaMarca.Row - 1
                 Sheets(Grupo & Item & "-sl").Rows(CelulaMarca.Row & ":" & CelulaMarca.Row).Delete
Shift:=xlUp
             End If
       End If

    Range("A2").FormulaR1C1 = "1"
    Range("A3").FormulaR1C1 = "2"
    Range("A2:A3").AutoFill Destination:=Range("A2:A" & Range("B1",
ActiveSheet.Range("B1048576").End(xlUp)).Count)

Next Item

End Sub
Sub CalculaQuantidadeSintagmas()



For Each Item In ItensGrupo

    Sheets.Add After:=Sheets(Sheets.Count)
    Sheets(Sheets.Count).Name = Grupo & Item
    Sheets(Grupo & Item).Select

    Range("D1").Select
    ActiveWorkbook.PivotCaches.Create(SourceType:=xlDatabase, SourceData:= _
       Grupo & Item & "-sl!R1C1:R1048576C2", Version:=xlPivotTableVersion12).CreatePivotTable _
       TableDestination:=Grupo & Item & "!R1C4", TableName:="Tabela dinâmica4", _
       DefaultVersion:=xlPivotTableVersion12
    Sheets(Grupo & Item).Select
    Cells(1, 4).Select
    ActiveWorkbook.ShowPivotTableFieldList = True
    With ActiveSheet.PivotTables("Tabela dinâmica4").PivotFields("Sintagma")
       .Orientation = xlRowField
       .Position = 1
```

```
    End With
    ActiveSheet.PivotTables("Tabela dinâmica4").AddDataField ActiveSheet. _
        PivotTables("Tabela dinâmica4").PivotFields("Posicao"), "Contar de Posicao", _
        xlCount
    ActiveWorkbook.ShowPivotTableFieldList = False
    Columns("D:E").Select
    Selection.Copy
    Range("A1").Select
    Selection.PasteSpecial Paste:=xlPasteValues, Operation:=xlNone, SkipBlanks _
        :=False, Transpose:=False
    Columns("D:E").Select
    Application.CutCopyMode = False
    Selection.Delete Shift:=xlToLeft

    Columns("A:B").EntireColumn.AutoFit

    Range("A1").FormulaR1C1 = "Sintagma"
    Range("B1").FormulaR1C1 = "Quantidade"

    Ultimalinha = Range("A1", ActiveSheet.Range("A1048576").End(xlUp)).Count
    Rows(Ultimalinha - 1 & ":" & Ultimalinha).Delete Shift:=xlUp

    Columns("A:B").Select
    ActiveWorkbook.Worksheets(Grupo & Item).Sort.SortFields.Clear
    ActiveWorkbook.Worksheets(Grupo & Item).Sort.SortFields.Add Key:=Range("B2:B" & Range("A1",
ActiveSheet.Range("A1048576").End(xlUp)).Count), _
        SortOn:=xlSortOnValues, Order:=xlDescending, DataOption:=xlSortNormal
    With ActiveWorkbook.Worksheets(Grupo & Item).Sort
        .SetRange Range("A1:B" & Range("A1", ActiveSheet.Range("A1048576").End(xlUp)).Count)
        .Header = xlYes
        .MatchCase = False
        .Orientation = xlTopToBottom
        .SortMethod = xlPinYin
        .Apply
    End With

    Range("C2").FormulaR1C1 = "=MAX(C[-1])"
    Sheets(Grupo & ".Corpus").Range("H" & Item + 1).Value = Range("C2").Value
    Range("C3").FormulaR1C1 = "=COUNTIF(C[-1],1)"
    Sheets(Grupo & ".Corpus").Range("G" & Item + 1).Value = Range("C3").Value
    Range("C4").FormulaR1C1 = "=SUM(C[-1])"
    Sheets(Grupo & ".Corpus").Range("E" & Item + 1).Value = Range("C4").Value
    Range("C5").FormulaR1C1 = "=COUNT(C[-1])"
    Sheets(Grupo & ".Corpus").Range("F" & Item + 1).Value = Range("C5").Value

    Sheets(Grupo & Item & "-sl").Select
    Ultimalinha = Range("A1", ActiveSheet.Range("A1048576").End(xlUp)).Count - 1
    Sheets(Grupo & ".Corpus").Range("L" & Item + 1).Value = Ultimalinha
    Sheets(Grupo & Item).Select
    Range("C2:C6").FormulaR1C1 = ""

Next Item

End Sub
Sub CalculaDocumentoscomSintagma()



For Each Item In ItensGrupo
```

```
    Sheets(Grupo & Item).Select

    Range("C1").FormulaR1C1 = "Documentos"

    FormulaDocumentos = "="
    For Each ItemFormula In ItensGrupo
        FormulaDocumentos = FormulaDocumentos & "+COUNTIF('" & Grupo & ItemFormula & "'!C[-
2],RC[-2])"
    Next ItemFormula

    Range("C2").FormulaR1C1 = FormulaDocumentos
    Range("C2").AutoFill Destination:=Range("C2:C" & Range("A1",
ActiveSheet.Range("A1048576").End(xlUp)).Count)
    Columns("C:C").Select
    Selection.Copy
    Selection.PasteSpecial Paste:=xlPasteValues, Operation:=xlNone, SkipBlanks _
        :=False, Transpose:=False
    Application.CutCopyMode = False

Next Item

End Sub
Sub CalculaCategoriaSintagma()


For Each Item In ItensGrupo

    Sheets(Grupo & Item).Select

    Range("D1").FormulaR1C1 = "CSN"
    Range("D2").FormulaR1C1 = "=VLOOKUP(RC[-3],'" & Grupo & Item & "-tral'!C[-3]:C,4,FALSE)"
    Range("D2").AutoFill Destination:=Range("D2:D" & Range("A1",
ActiveSheet.Range("A1048576").End(xlUp)).Count)

    Range("E1").FormulaR1C1 = "CSN Valor"
    Range("E2").FormulaR1C1 = "=VLOOKUP(RC[-1],'" & Grupo & ".Corpus!R4C1:R10C2,2,FALSE)"
    Range("E2").AutoFill Destination:=Range("E2:E" & Range("A1",
ActiveSheet.Range("A1048576").End(xlUp)).Count)
    Columns("E:E").Select
    Selection.Copy
    Selection.PasteSpecial Paste:=xlPasteValues, Operation:=xlNone, SkipBlanks _
        :=False, Transpose:=False
    Columns("D:D").Select
    Application.CutCopyMode = False
    Selection.Delete Shift:=xlToLeft

    'Sheets(Grupo & Item & "-tral").Select
    'Application.DisplayAlerts = False
    'ActiveWindow.SelectedSheets.Delete
    'Application.DisplayAlerts = True

Next Item


End Sub
Sub PontuacaoSintagma()


X = 0
For Each Item In ItensGrupo
```

```
    Sheets(Grupo & Item).Select
    X = X + 1

    Range("E1").FormulaR1C1 = "Pontuação"
    Range("E2").FormulaR1C1 = _
        "=(RC[-3]/" & Grupo & ".Corpus!R" & X + 1 & "C8)*LOG(" & Grupo & ".Corpus!R2C2/RC[-2])*RC[-
1]"
    Range("E2").AutoFill Destination:=Range("E2:E" & Range("A1",
ActiveSheet.Range("A1048576").End(xlUp)).Count)

    Columns("E:E").Select
    Selection.Copy
     Selection.PasteSpecial Paste:=xlPasteValues, Operation:=xlNone, SkipBlanks _
        :=False, Transpose:=False
    Columns("A:E").Select
    Application.CutCopyMode = False
    ActiveWorkbook.Worksheets(Grupo & Item).Sort.SortFields.Clear
    ActiveWorkbook.Worksheets(Grupo & Item).Sort.SortFields.Add Key:=Range("E2:E" & Range("A1",
ActiveSheet.Range("A1048576").End(xlUp)).Count), _
        SortOn:=xlSortOnValues, Order:=xlDescending, DataOption:=xlSortNormal
    With ActiveWorkbook.Worksheets(Grupo & Item).Sort
        .SetRange Range("A1:E" & Range("A1", ActiveSheet.Range("A1048576").End(xlUp)).Count)
        .Header = xlYes
        .MatchCase = False
        .Orientation = xlTopToBottom
        .SortMethod = xlPinYin
        .Apply
    End With

Next Item

End Sub
Sub FormularioPesquisa()

X = 0
For Each Item In ItensGrupo

    X = X + 1

    Sheets.Add After:=Sheets(Sheets.Count)
    Sheets(Sheets.Count).Select
    Sheets(Sheets.Count).Name = Grupo & Item & "-q"
    Sheets(Grupo & Item & "-q").Select

    Range("A1").FormulaR1C1 = "Programa"
    Range("A2").FormulaR1C1 = "Grupo"
    Range("A3").FormulaR1C1 = "Tese"
    Range("A4").FormulaR1C1 = "Data da Publicação"
    Range("A5").FormulaR1C1 = "Autor"
    Range("A6").FormulaR1C1 = "Título"
    Range("A7").FormulaR1C1 = "Email"
    Range("A8").FormulaR1C1 = "Link Formulário"
    Range("A10").FormulaR1C1 = "Descritor Candidato"
    Range("B10").FormulaR1C1 = "Avaliação (1 a 7)"

    Range("B1").FormulaR1C1 = Sheets(Grupo & ".Teses").Range("B1").FormulaR1C1
    Range("B2").FormulaR1C1 = Grupo
    Range("B3").FormulaR1C1 = Item
    Sheets(Grupo & ".Teses").Select
```

```
    Range("B" & X + 2 & ":E" & X + 2).Select
    Selection.Copy
    Sheets(Grupo & Item & "-q").Select
    Range("B4").Select
    Selection.PasteSpecial Paste:=xlPasteValues, Operation:=xlNone, SkipBlanks _
        :=False, Transpose:=True

    Range("B3").HorizontalAlignment = xlLeft
    Range("B4").NumberFormat = "m/d/yyyy"
    Range("B4").HorizontalAlignment = xlLeft


    Sheets(Grupo & Item).Select
    Range("A2:A21").Select
    Application.CutCopyMode = False
    Selection.Copy
    Sheets(Grupo & Item & "-q").Select
    Range("A11").Select
    ActiveSheet.Paste
    Columns("A:A").EntireColumn.AutoFit
    Columns("B:B").EntireColumn.AutoFit
    Range("A1:A8").Select
    Selection.Font.Bold = False
    Selection.Font.Bold = True
    Range("A10:B10").Select
    Selection.Font.Bold = False
    Selection.Font.Bold = True

    Range("A10:B30").Select
    ActiveWorkbook.Worksheets(Grupo & Item & "-q").Sort.SortFields.Clear
    ActiveWorkbook.Worksheets(Grupo & Item & "-q").Sort.SortFields.Add Key:=Range("A11:A30") _
        , SortOn:=xlSortOnValues, Order:=xlAscending, DataOption:=xlSortNormal
    With ActiveWorkbook.Worksheets(Grupo & Item & "-q").Sort
        .SetRange Range("A10:B30")
        .Header = xlYes
        .MatchCase = False
        .Orientation = xlTopToBottom
        .SortMethod = xlPinYin
        .Apply
    End With


Next Item

End Sub

Sub PreparacaoTeses()

LocalizacaoGrupo = "H:\UFMG\ECI\DISSERTAÇÃO\CORPUS\Teses\"
Grupo = "H"
ItensGrupo = Array("01", "02", "03", "04", "05", "06", "07") ', "08") ', "09", "10") ', "11", "12") ', "13") ',
"14", "15", "16", "17", "18", "19", "20", "21", "22", "23", "24")
TamanhoGrupo = 7

Sheets.Add After:=Sheets(Sheets.Count)
Sheets(Sheets.Count).Name = Grupo & ".Corpus"


    Range("A1").FormulaR1C1 = "Corpus"
    Range("B1").FormulaR1C1 = "Número de Documentos (N)"
```

```
    Range("D1").FormulaR1C1 = "Documento"
    Range("E1").FormulaR1C1 = "Sintagmas Extraídos"
    Range("F1").FormulaR1C1 = "Sintagmas Identificados"
    Range("G1").FormulaR1C1 = "Sintagmas Únicos"
    Range("H1").FormulaR1C1 = "Maior Frequencia de um Sintagma"
    Range("I1").FormulaR1C1 = "Erros de Extração"
    Range("J1").FormulaR1C1 = "Posição do Início do Meio"
    Range("K1").FormulaR1C1 = "Posição do Início do Fim"
    Range("L1").FormulaR1C1 = "Posição Final"

    Range("A2").FormulaR1C1 = Grupo
    Range("B2").FormulaR1C1 = TamanhoGrupo

    Range("A4").FormulaR1C1 = "CSN"
    Range("A5").FormulaR1C1 = "1a"
    Range("A6").FormulaR1C1 = "1b"
    Range("A7").FormulaR1C1 = "2"
    Range("A8").FormulaR1C1 = "3"
    Range("A9").FormulaR1C1 = "4"
    Range("A10").FormulaR1C1 = "5"
    Range("B4").FormulaR1C1 = "Valor CSN"
    Range("B5").FormulaR1C1 = "0.2"
    Range("B6").FormulaR1C1 = "0.8"
    Range("B7").FormulaR1C1 = "1.1"
    Range("B8").FormulaR1C1 = "1.4"
    Range("B9").FormulaR1C1 = "1.2"
    Range("B10").FormulaR1C1 = "0.8"

    X = 0
    For Each Item In ItensGrupo
       X = X + 1
       Range("D" & X + 1).FormulaR1C1 = Grupo & Item
    Next Item

ImportarSL
ImportarTRAL
LimpaSintagmaErroSlxTral
RetiraMarcaMeioFim
CalculaQuantidadeSintagmas
CalculaDocumentoscomSintagma
CalculaCategoriaSintagma
PontuacaoSintagma
FormularioPesquisa

End Sub
```

APÊNDICE F - EXEMPLO DE QUESTIONÁRIO ENVIADO PARA OS ENTREVISTADOS

Questionário A01

Programa Pós-Graduação em Educação: Conhecimento e Inclusão Social

Grupo A

Tese 1

Data da Publicação 14/12/2011

Autor Marlice de Oliveira e Nogueira

Título Pais professores e a escolarização dos filhos

Email <trecho retirado>@uol.com.br

Para cada sintagma nominal abaixo determine o grau de relevância do mesmo como descritor de sua tese.

---

*Obrigatório

bom aluno*

	1	2	3	4	5	6	7	
Não Relevante	○	○	○	○	○	○	○	Extremamente Relevante

caso dos pais*

	1	2	3	4	5	6	7	
Não Relevante	○	○	○	○	○	○	○	Extremamente Relevante

dois filhos*

	1	2	3	4	5	6	7	
Não Relevante	○	○	○	○	○	○	○	Extremamente Relevante

escola dos filhos*

	1	2	3	4	5	6	7	
Não Relevante	○	○	○	○	○	○	○	Extremamente Relevante

escolar dos filhos*

	1	2	3	4	5	6	7	
Não Relevante	○	○	○	○	○	○	○	Extremamente Relevante

escolares dos filhos*

	1	2	3	4	5	6	7	
Não Relevante	○	○	○	○	○	○	○	Extremamente Relevante

escolarização dos filhos*

	1	2	3	4	5	6	7	
Não Relevante	○	○	○	○	○	○	○	Extremamente Relevante

escolha do estabelecimento*

	1	2	3	4	5	6	7	
Não Relevante	○	○	○	○	○	○	○	Extremamente Relevante

famílias do grupo*

	1	2	3	4	5	6	7	
Não Relevante	○	○	○	○	○	○	○	Extremamente Relevante

famílias fortemente orientadas para o sucesso*

	1	2	3	4	5	6	7	
Não Relevante	○	○	○	○	○	○	○	Extremamente Relevante

maio e junho de 2009*

	1	2	3	4	5	6	7	
Não Relevante	○	○	○	○	○	○	○	Extremamente Relevante

meses de abril*

	1	2	3	4	5	6	7	
Não Relevante	○	○	○	○	○	○	○	Extremamente Relevante

professor de ciências*

	1	2	3	4	5	6	7	
Não Relevante	○	○	○	○	○	○	○	Extremamente Relevante

professora de ciências*

1	2	3	4	5	6	7

Não Relevante								Extremamente Relevante

professora de geografia*

	1	2	3	4	5	6	7	
Não Relevante								Extremamente Relevante

professora de matemática*

	1	2	3	4	5	6	7	
Não Relevante								Extremamente Relevante

professores do grupo*

	1	2	3	4	5	6	7	
Não Relevante								Extremamente Relevante

professores do município*

	1	2	3	4	5	6	7	
Não Relevante								Extremamente Relevante

questionário aplicado*

	1	2	3	4	5	6	7	
Não Relevante								Extremamente Relevante

total 114*

	1	2	3	4	5	6	7	
Não Relevante								Extremamente Relevante

Enviar

APÊNDICE G - MACROS DO MICROSOFT OFFICE EXCEL 2007 PARA A CONSOLIDAÇÃO DE VALORES ASSOCIADOS POR POSIÇÃO

```
Public LocalizacaoGrupo As String
Public Grupo As String
Public ItensGrupo As Variant
Public TamanhoGrupo  As Integer

Sub CalculaValorAvaliado()

    Sheets(Grupo & ".Corpus").Select

    Range("A12").FormulaR1C1  = "Avaliação do Autor"
    Range("A13").FormulaR1C1  = "1"
    Range("A14").FormulaR1C1  = "2"
    Range("A15").FormulaR1C1  = "3"
    Range("A16").FormulaR1C1  = "4"
    Range("A17").FormulaR1C1  = "5"
    Range("A18").FormulaR1C1  = "6"
    Range("A19").FormulaR1C1  = "7"
    Range("B12").FormulaR1C1  = "Valor da Avaliação do Autor"
    Range("B13").FormulaR1C1  = "0"
    Range("B14").FormulaR1C1  = "0.25"
    Range("B15").FormulaR1C1  = "0.25"
    Range("B16").FormulaR1C1  = "0.5"
    Range("B17").FormulaR1C1  = "0.5"
    Range("B18").FormulaR1C1  = "1"
    Range("B19").FormulaR1C1  = "1"

For Each Item In ItensGrupo

    Sheets(Grupo & Item).Select

    Range("F1").FormulaR1C1  = "Avaliação do Autor"
    Range("G1").FormulaR1C1  = "Valor da Avaliação do Autor"
    Range("H1").FormulaR1C1  = "Valor do SN (unidade)"

    Range("F2").FormulaR1C1  =  "=VLOOKUP(RC[-5],'" & Grupo & Item & "-
q'!R11C1:R30C2,2,FALSE)"
    Range("F2").AutoFill Destination:=Range("F2:F21"),  Type:=xlFillDefault

    Range("G2").FormulaR1C1  = "=VLOOKUP(RC[-1]," & Grupo & ".Corpus!R13C1:R19C2,2,FALSE)"
    Range("G2").AutoFill Destination:=Range("G2:G21"),  Type:=xlFillDefault

    Range("H2").FormulaR1C1  = "=RC[-1]/RC[-6]"
    Range("H2").AutoFill Destination:=Range("H2:H21"),  Type:=xlFillDefault



Next Item

End Sub
Sub AtribuiValorPosicao()

X = 0

For Each Item In ItensGrupo

X = X + 1
```

```
    Sheets(Grupo & Item & "-sl").Select

    Range("C1").FormulaR1C1 = "Posição (1/10)"
    Range("D1").FormulaR1C1 = "Posição (1/3)"
    Range("E1").FormulaR1C1 = "Valor"

    Range("C2").FormulaR1C1 = "=ROUNDUP(RC[-2]/MAX(C[-2]),1)"
    Range("C2").AutoFill Destination:=Range("C2:C" & Range("B1",
ActiveSheet.Range("B1048576").End(xlUp)).Count)

    Range("D2").FormulaR1C1 = "=IF(RC[-3]<" & Grupo & ".Corpus!R" & X + 1 & "C10,1,IF(RC[-3]<" &
Grupo & ".Corpus!R" & X + 1 & "C11,2,3))"
    Range("D2").AutoFill Destination:=Range("D2:D" & Range("B1",
ActiveSheet.Range("B1048576").End(xlUp)).Count)

    Range("E2").FormulaR1C1 = "=IF(ISERROR(VLOOKUP(RC[-3],'" & Grupo & Item &
"'!R2C1:R21C8,8,FALSE)),0,VLOOKUP(RC[-3],'" & Grupo & Item & "'!R2C1:R21C8,8,FALSE))"
    Range("E2").AutoFill Destination:=Range("E2:E" & Range("B1",
ActiveSheet.Range("B1048576").End(xlUp)).Count)

    Columns("C:E").Select
    Selection.Copy
    Selection.PasteSpecial Paste:=xlPasteValues, Operation:=xlNone, SkipBlanks _
        :=False, Transpose:=False
    Application.CutCopyMode = False

Next Item

End Sub
Sub CalculaValorPosicaoGeral()

For Each Item In ItensGrupo

    Sheets(Grupo & Item & "-sl").Select


    Range("G1").Consolidate Sources:= _
        """ & LocalizacaoGrupo & Grupo & "\[" & Grupo & ".xlsx]" & Grupo & Item & "-sl'!C3:C5",
Function:= _
        xlSum, TopRow:=False, LeftColumn:=True, CreateLinks:=False
    Columns("H:H").Delete Shift:=xlToLeft
    Range("G12").FormulaR1C1 = "Total"
    Range("H12").FormulaR1C1 = "=SUM(R[-10]C:R[-1]C)"
    Range("H1").FormulaR1C1 = "Soma de Valores de Relevância"
    Range("J1").Consolidate Sources:= _
        """ & LocalizacaoGrupo & Grupo & "\[" & Grupo & ".xlsx]" & Grupo & Item & "-sl'!C4:C5",
Function:= _
        xlAverage, TopRow:=False, LeftColumn:=True, CreateLinks:=False
    Range("K1").FormulaR1C1 = "Média de Valores de Relevância"


Next Item



End Sub
Sub AnaliseGeral()

    Sheets.Add After:=Sheets(Sheets.Count)
```

```
    Sheets(Sheets.Count).Select
    Sheets(Sheets.Count).Name = Grupo & ".Analise"
    Sheets(Grupo & ".Analise").Select

    Sheets(Grupo & ".Corpus").Range("D1:L25").Copy
    Sheets(Grupo & ".Analise").Select
    Range("A1").Select
    Selection.PasteSpecial Paste:=xlPasteValues, Operation:=xlNone, SkipBlanks _
        :=False, Transpose:=True
    Application.CutCopyMode = False

    Range("A11").FormulaR1C1 = "SN Candidatos"
    Range("A12").FormulaR1C1 = "1º"
    Range("A13").FormulaR1C1 = "2º"
    Range("A12:A13").AutoFill Destination:=Range("A12:A31"), Type:=xlFillDefault

X = 0

For Each Item In ItensGrupo

    X = X + 1

    Sheets(Grupo & ".Analise").Range("A11").Offset(0, X).Select
    ActiveCell.FormulaR1C1 = Grupo & Item
    Sheets(Grupo & Item).Range("A2:A21").Copy
    Sheets(Grupo & ".Analise").Range("A12").Offset(0, X).Select
    Selection.PasteSpecial Paste:=xlPasteValues, Operation:=xlNone, SkipBlanks _
        :=False, Transpose:=False
    Application.CutCopyMode = False

Next Item


    Sheets(Grupo & ".Analise").Select
    Range("A33").FormulaR1C1 = "Avaliação do SN Candidato"
    Range("A34").FormulaR1C1 = "1º"
    Range("A35").FormulaR1C1 = "2º"
    Range("A34:A35").AutoFill Destination:=Range("A34:A53"), Type:=xlFillDefault

X = 0

For Each Item In ItensGrupo

    X = X + 1

    Sheets(Grupo & ".Analise").Range("A33").Offset(0, X).Select
    ActiveCell.FormulaR1C1 = Grupo & Item

    Sheets(Grupo & Item).Range("G2:G21").Copy

    Sheets(Grupo & ".Analise").Range("A34").Offset(0, X).Select
    Selection.PasteSpecial Paste:=xlPasteValues, Operation:=xlNone, SkipBlanks _
        :=False, Transpose:=False
    Application.CutCopyMode = False

Next Item


    Range("A54").FormulaR1C1 = "Total"
```

```
X = 0

For Each Item In ItensGrupo

   X = X + 1

   Sheets(Grupo & ".Analise").Range("A54").Offset(0, X).Select
   ActiveCell.FormulaR1C1 = "=SUM(R[-20]C:R[-1]C)"

Next Item

X = X + 1

   Sheets(Grupo & ".Analise").Range("A33").Offset(0, X).Select
   ActiveCell.FormulaR1C1 = "Média"

Y = 1

While Y < 22

   Sheets(Grupo & ".Analise").Range("A33").Offset(0, X).Select
   ActiveCell.Offset(Y, 0).FormulaR1C1 = "=AVERAGE(RC[-" & X - 1 & "]:RC[-1])"
   Y = Y + 1

Wend


   Range("A56").FormulaR1C1 = "Posição (1/10)"
   Range("A57").FormulaR1C1 = "10%"
   Range("A58").FormulaR1C1 = "20%"
   Range("A57:A58").AutoFill Destination:=Range("A57:A66"), Type:=xlFillDefault
   Range("A67").FormulaR1C1 = "Total"


 X = 0

For Each Item In ItensGrupo

   X = X + 1

   Sheets(Grupo & ".Analise").Range("A56").Offset(0, X).Select
   ActiveCell.FormulaR1C1 = Grupo & Item

   Sheets(Grupo & Item & "-sl").Range("H2:H12").Copy

   Sheets(Grupo & ".Analise").Range("A57").Offset(0, X).Select
   Selection.PasteSpecial Paste:=xlPasteValues, Operation:=xlNone, SkipBlanks _
      :=False, Transpose:=False
   Application.CutCopyMode = False

Next Item

X = X + 1

   Sheets(Grupo & ".Analise").Range("A56").Offset(0, X).Select
   ActiveCell.FormulaR1C1 = "Média"

Y = 1

While Y < 12
```

```
    Sheets(Grupo & ".Analise").Range("A56").Offset(0, X).Select
    ActiveCell.Offset(Y, 0).FormulaR1C1 = "=AVERAGE(RC[-" & X - 1 & "]:RC[-1])"
    Y = Y + 1

Wend


    Range("A69").FormulaR1C1 = "Posição (1/3)"
    Range("A70").FormulaR1C1 = "Introdução"
    Range("A71").FormulaR1C1 = "Desenvolvimento"
    Range("A72").FormulaR1C1 = "Conclusão"
    Range("A73").FormulaR1C1 = "Total"


 X = 0

For Each Item In ItensGrupo

    X = X + 1

    Sheets(Grupo & ".Analise").Range("A69").Offset(0, X).Select
    ActiveCell.FormulaR1C1 = Grupo & Item

    Sheets(Grupo & Item & "-sl").Range("K2:K4").Copy

    Sheets(Grupo & ".Analise").Range("A70").Offset(0, X).Select
    Selection.PasteSpecial Paste:=xlPasteValues, Operation:=xlNone, SkipBlanks _
        :=False, Transpose:=False
    Application.CutCopyMode = False

    Sheets(Grupo & ".Analise").Range("A73").Offset(0, X).Select
    ActiveCell.FormulaR1C1 = "=SUM(R[-3]C:R[-1]C)"


Next Item

X = X + 1

    Sheets(Grupo & ".Analise").Range("A69").Offset(0, X).Select
    ActiveCell.FormulaR1C1 = "Média"

Y = 1

While Y < 5

    Sheets(Grupo & ".Analise").Range("A69").Offset(0, X).Select
    ActiveCell.Offset(Y, 0).FormulaR1C1 = "=AVERAGE(RC[-" & X - 1 & "]:RC[-1])"
    Y = Y + 1

Wend




End Sub

Sub AnaliseTeses()
```

```
LocalizacaoGrupo = "H:\UFMG\ECI\DISSERTAÇÃO\CORPUS\Teses\"
Grupo = "H"
ItensGrupo = Array("01", "02", "03", "04", "05", "06", "07") ', "08") ', "09", "10", "11", "12", "13", "14",
"15", "16", "17", "18", "19", "20", "21", "22", "23", "24")
TamanhoGrupo  = 7



CalculaValorAvaliado
AtribuiValorPosicao
CalculaValorPosicaoGeral
AnaliseGeral

End Sub
```

APÊNDICE H - LISTA DAS TESES ANALISADAS COM DATA DE PUBLICAÇÃO NA BDTD/UFMG, AUTOR E TÍTULO

Seção do *corpus*	Área de Conhecimento	Programa de pós-graduação com maior nº de teses na mesma área de conhecimento	Quantidade de teses
A	Ciências Humanas	Pós-Graduação em Educação: Conhecimento e Inclusão Social	24
B	Ciências Agrárias	Pós-Graduação em Ciência Animal	16
C	Linguística, Letras e Artes	Pós-Graduação em Letras: Estudos Literários	13
D	Engenharias	Pós-Graduação em Engenharia Metalúrgica e de Minas	12
E	Ciências Exatas e da Terra	Pós-Graduação em Química	10
F	Ciências Biológicas	Pós-Graduação em Bioquímica e Imunologia	8
G	Ciências Sociais Aplicadas	Pós-Graduação em Ciência da Informação	8
H	Ciências da Saúde	Pós-Graduação em Medicina (Pediatria)	7

Fonte: Elaborado pelo autor.

Seção do *corpus*	No	Data da Publicação	Título	Autor (orientando)
A	1	14/12/2011	Pais professores e a escolarização dos filhos	Marlice de Oliveira e Nogueira
A	2	28/02/2012	Política, trabalho e intolerância: ensino primário e as práticas educativas em Minas Gerais (1930-1954)	Aline Choucair Vaz
A	3	09/09/2011	Currículo, gênero e nordestinidade: o que ensina o forró eletrônico?	Marlécio Maknamara
A	4	24/05/2011	A evolução do entendimento dos estudantes em eletricidade: um estudo longitudinal	Geide Rosa Coelho
A	5	20/12/2011	Discurso em salas de aula de ciências: uma estrutura de análise baseada na teoria da atividade,	Rodrigo Drumond

Seção do *corpus*	No	Data da Publicação	Título	Autor (orientando)
			sociolinguística e linguística textual	
A	6	06/09/2011	Diversificação dos modos de ser masculino e estatização da violência masculina na escrita literária e jornalística de Bernardo Guimarães	Matheus da Cruz e Zica
A	7	26/04/2011	Letramento escolar: eventos e apropriações de gêneros textuais por adolescentes	Valeria Barbosa de Resende
A	8	21/12/2011	O ensino de ciências por investigação na educação superior: um ambiente para o estudo da aprendizagem científica	Fabio Augusto Rodrigues e Silva
A	9	17/12/2010	Experiência e formação: o fazer teatral nas trajetórias docentes	Andrea Maria Favilla Lobo
A	10	15/12/2010	O trabalho docente no movimento de reformas educacionais no estado do Acre	Ednaceli Abreu Damasceno
A	11	14/12/2010	Tensões contemporâneas no processo de passagem da educação infantil para o ensino fundamental: um estudo de caso	Vanessa Ferraz Almeida Neves
A	12	16/12/2010	Aulas no ensino superior: uma visão sobre professores de disciplinas científicas na licenciatura em Química da UFMG	Ana Luiza de Quadros
A	13	08/02/2011	Caminhos da docência: trajetórias de mulheres professoras em Sabará Minas Gerais (1830-1904)	Cecilia Vieira do Nascimento
A	14	17/02/2011	Um estudo sobre a consistência de modelos mentais sobre mecânica de estudantes de ensino médio	Simone Aparecida Fernandes
A	15	21/09/2010	Reformas educacionais e gestão democrática no estado do Acre: repercussões no trabalho do núcleo gestor da escola	Lucia de Fatima Melo
A	16	15/10/2010	A constituição docente em matemática à distância: Entre saberes,experiências e narrativas	Diva Souza Silva
A	17	06/07/2010	Desenvolvimento profissional de professores de História: estudo de caso de um grupo colaborativo mediado pelas tecnologias de informação e comunicação aplicadas à educação	Andreia de Assis Ferreira
A	18	31/03/2010	Desenvolvimento profissional de professores: a influência da vivência em um grupo colaborativo	Paulo Henrique Dias Menezes

Seção do *corpus*	No	Data da Publicação	Título	Autor (orientando)
A	19	24/08/2010	Uma pedagogia da experiência do encontro bordada nas trocas: Associação de Mulheres do Bairro Bethânia - Ipatinga, MG	Maria Luciana Brandao Silva
A	20	13/04/2010	Saberes e práticas em redes de trocas: a temática africana e afro-brasileira em questão	Lorene dos Santos
A	21	06/05/2010	A relação pedagógica e a avaliação no espelho do portfólio: memórias docentes e discentes	Marcia Ambrosio Rodrigues Rezende
A	22	15/04/2010	Quando O SANTO chama: O terreiro de umbanda como contexto de aprendizagem na prática	Renata Silva Bergo
A	23	31/05/2010	As políticas de educação superior: novos modos de regulação e seus desdobramentos nos cursos de graduação em Odontologia (1995-2008)	Maria Ines Barreiros Senna
A	24	26/02/2010	Orkut.com.escol@: currículos e ciborguização juvenil	Shirlei Rezende Sales
B	1	21/06/2010	A representação social do saber de trabalhadores rurais sobre o controle de parasitos em propriedades produtoras de leite	Ana Cristina Passos de Paiva Bello
B	2	17/02/2011	Perfil eletroforético de proteínas e concentrações de leptina, insulina e IGF-I do plasma seminal de tourinhos Gir-Leiteiros na peripuberdade	Fernando Andrade Souza
B	3	25/11/2011	Clostrídios entéricos de leitões neonatos, desenvolvimento e avaliação de uma vacina experimental	Felipe Masiero Salvarani
B	4	29/01/2010	Prevalência de enteropatógenos em suínos de recria/terminação em Minas Gerais e desenvolvimento de modelo experimental murino de enteropatia proliferativa	Aline de Marco Viott
B	5	09/02/2010	Desenvolvimento reprodutivo e análise das proteínas do plasma seminal com afinidade à heparina, em tourinhos Gir selecionados para a produção de leite	Jorge Andre Matias Martins
B	6	08/05/2009	Avaliação histológica, histoquímica, morfométrica e radiográfica de traquéias de cães portadores de colapso traqueal	Paulo Eduardo Ferian
B	7	07/03/2008	Mamite bovina em rebanhos leiteiros da região sul do Estado de Minas Gerais	Geraldo Marcio da Costa
B	8	27/02/2009	Monitoramento sorológico e da presença do DNA pró-viral do lentivirus caprino (CAEV) no sangue e semen de reprodutores infectados	Juliano Cezar Minardi da Cruz
B	9	16/04/2009	Modelo de infecção gastrintestinal e o papel do LPS, urease e sistema de secreção do tipo 4 da Brucella melitensis em camundongos	Tatiane Alves da Paixao

Seção do *corpus*	No	Data da Publicação	Título	Autor (orientando)
B	10	05/07/2011	Ocorrência de arsênio, cádmio e chumbo em tecidos de aves, suínos, bovinos de corte e equinos no Brasil	Juarez Fabiano de Alkmim Filho
B	11	16/04/2009	Parâmetros reprodutivos, metabólitos e produção de leite de vacas mestiças Holandês X Zebu submetidas a dois manejos pré-parto""	Bruno Campos de Carvalho
B	12	19/02/2009	Caracterização molecular e imunológica do veneno de Tityus fasciolatus e sua ação sobre camundongos	Priscylla Tatiana Chalfun Guimaraes
B	13	16/02/2009	Detecção do vírus da anemia infecciosa das galinhas em Minas Gerais	Priscilla Rochele Barrios
B	14	27/02/2008	Formas de produção pecuária e distribuição da febre aftosa no departamento de Santa Cruz, Bolívia, 2000-2007	Hernan Oliver Daza Gutierrez
B	15	26/02/2008	Imunogenicidade de bacterinas anti-leptospiras para bovinos produzidas no Brasil, 2006/7	Rogerio Oliveira Rodrigues
B	16	07/03/2008	Purificação e caracterização parcial de inibidores de serino protease e sua influencia sobre a viabilidade espermática equina nos processos de resfriamento e congelamento	Andre Belico de Vasconcelos
C	1	27/02/2012	Literatura e biblioteca em Jorge Luis Borges e Italo Calvino	Maria Elisa Rodrigues Moreira
C	2	02/12/2011	Antonio Candido: crítica, reflexão e memória	Jose Quintao de Oliveira
C	3	06/02/2012	Do canto da voz ao batuque da letra: a presença africana em narrativas orais inscritas no Brasil	Josiley Francisco de Souza
C	4	13/02/2012	A narrativa memorialística dos álbuns de Antonio Guerra	Maria Tereza Gomes de Almeida Lima
C	5	23/02/2011	O poema concreto e a contribuição de Lacan: a não-relação endereçada	Rosangela Ramos Corgosinho
C	6	24/02/2011	A crítica entre a literatura e a História: o percurso da crítica literária de Sérgio Buarque de Holanda dos verdes anos à profissionalização do ofício	Mariana Thiengo
C	7	16/05/2011	Ensaísmo de Paulo Leminski: panorama de um pensamento movente	Renata Melo Moreira

Seção do *corpus*	No	Data da Publicação	Título	Autor (orientando)
C	8	25/02/2011	O vermelho da vida na escrita de Hilda Hilst	Ludmilla Zago Andrade
C	9	10/06/2011	Entre Guimarães Rosa, Manoel de Barros e Bartolomeu Campos Queirós: a criação de uma infância da escrita	Rosane da Silva Gomes
C	10	23/05/2011	Textualidades em negativo: a ficção de António Lobo Antunes	Denis Leandro Francisco
C	11	01/10/2011	Koxuk, a imagem do yâmîy na poética maxakali	Charles Antonio de Paula Bicalho
C	12	30/11/2011	A matemática em Georges Perec e Jorge Luis Borges: um estudo comparativo	Jacques Fux
C	13	08/07/2010	O que ajunta espalha: tempo e paradoxo em Grande sertão: veredas, de João Guimarães Rosa, e Nós, os do Makulusu, de José Luandino Vieira	Julio Cesar Machado de Paula
D	1	14/07/2011	Modificações superficiais de aço Ti-UBC por processos a plasma em configuração triodo: influência no comportamento ao desgaste e à corrosão	Carlos Alberto Llanes Leyva
D	2	09/11/2011	Reciclagem de resíduo gerado na extração de quartzito	Mario Luis Cabello Russo
D	3	07/04/2011	Obtenção e caracterização de aço fundido bainítico com elevada resistência á fadiga mecânica de alto ciclo	Denilson Jose do Carmo
D	4	22/02/2011	Desenvolvimento e caracterização de copolímeros obtidos a partir de monômeros acrílicos e metacrílicos visando a aplicação como excipientes farmacêuticos para preparação de matrizes inertes por compressão direta	Janaina Cecilia Oliveira Villanova
D	5	25/05/2009	Análise do envelhecimento acelerado e da ação inibidora do ácido ascórbico na degradação oxidativa do polietileno de ultra-elevada massa molar para aplicação biomédica	Magda Francisca Goncalves Rocha
D	6	11/04/2008	Efeito do nitrogênio e do cobre na formação da martensita em aços inoxidáveis austeníticos e sua influência sobre o fenômeno de delayed cracking	Marta Ribeiro dos Santos
D	7	06/05/2009	Avaliação do efeito de modificações superficiais a plasma no desempenho frente ao desgaste de um aço baixa liga: estudo da correlação entre profundidade de endurecimento e melhoria de desempenho	Sandra Goulart Santos

Seção do *corpus*	No	Data da Publicação	Título	Autor (orientando)
D	8	17/12/2009	Estudo do efeito da reticulação por genipin em suportesbiocompatíveis de quitosana-PVA	Viviane Mota Bispo
D	9	14/05/2008	Características físicas, estruturais e mecânicas de instrumentos endodônticos de NiTi ProTaper	Renata de Castro Martins
D	10	19/02/2009	Sensor de NO2 utilizando-se filmes moleculares de macrociclos de porfirinas	Nelicio Faria de Sales
D	11	15/03/2010	Gestão ambiental dos sedimentos de corrente do rio SãoFrancisco na região de Três Marias/ Minas Gerais	Debora Fernandes Almeida
D	12	26/02/2008	Caracterização de ametistas naturais	Eduardo Henrique Martins Nunes
E	1	27/02/2012	Estudos de nanotubos de carbono e de titanatos e suas aplicações em reações de oxidação	Eudes Lorencon
E	2	26/08/2011	Determinação de parâmetros físico-químicos do óleo diesel a partir de curvas de destilação utilizando técnicas quimiométricas	Helga Gabriela Aleme
E	3	04/03/2011	Síntese de novos derivados fulerênicos explorando a "reação click" e de um derivado C60-catiônico polar	Guilherme Rocha Pereira
E	4	25/08/2011	Degradação oxidativa de compostos orgânicos em meio aquoso por via catalítica heterogênia com magnetita e goethita dopadas com nióbio	Diana Quintao Lima de Oliveira
E	5	14/04/2011	Estudo de filmes finos e materiais particulados de TiO2 e de Ag/TiO2 produzidos pelo processo sol-gel	Marcelo Machado Viana
E	6	24/02/2011	Aplicação dos processos oxidativos, redutivos e (foto)eletroquímicos na degradação de fármacos em meio aquoso	Karla Moreira Vieira
E	7	10/02/2012	Complexos metálicos de hidrazonas, tiossemicarbazonas e lapachol: atividade farmacológica e avaliação de relações estrutura-atividade	Gabrieli Lessa Parrilha
E	8	24/08/2011	Estudo da interação pósitron-matéria em sólidos supramoleculares orgânicos e sistemas aromáticos substituídos	Fernando Castro de Oliveira
E	9	14/02/2012	Geoquímica dos solos e das águas da Península Fildes e Ilha Ardley - Antártica Marítima	Renato Pereira de Andrade

Seção do *corpus*	No	Data da Publicação	Título	Autor (orientando)
E	10	17/08/2012	Estudo do perfil farmacológico de novas tiossemicarbazonas e novos complexos de bismuto (III) e antimônio (III)	Debora Costa Reis
F	1	12/09/2011	Análise peptidômica de venenos animais	Breno Rates Azevedo
F	2	19/02/2009	Purificação e caracterização bioquímica do tripsinogênio, a- e y-tripsina bovina e análise termodinâmica em meio ácido por calorimetria diferencial de varredura	Alexandre Martins Costa Santos
F	3	30/03/2010	O papel do interferon do tipo I e sua sinalização na resposta imune inata contra a infecção pela Brucella abortus	Leonardo Augusto de Almeida
F	4	19/02/2011	Efeitos do envelhecimento na mucosa intestinal: indução e declínio da tolerância oral	Andrezza Fernanda Santiago
F	5	11/09/2010	Estudos do papel do gene Rad51 de tripanossomatídeos na recombinação e no reparo de DNA	Danielle Gomes Passos Silva
F	6	12/07/2010	Cálcio intracelular na proliferação de células hepáticas	Viviane Aguiar Andrade
F	7	19/04/2010	Avaliações imunogenéticas do desenvolvimento de anticorpos inibidores do fator VIII na hemofilia A	Daniel Goncalves Chaves
F	8	30/07/2009	Reparo de DNA em dois patógenos humanos: caracterização do gene IMP4 de Schistosoma mansini e estudos acerca do MMR, Sistema GO e taxa de mutação em Trypanosoma cruzi	Carolina Furtado Torres da Silva
G	1	10/10/2011	Migração conceitual entre Sistemas de Recuperação da Informação e Ciências Cognitivas: uma análise discursiva	Fernando Skackauskas Dias
G	2	14/12/2011	Sistema de Informação da Atenção Básica (SIAB) como instrumento de poder	Ricardo Bezerra Cavalcante
G	3	05/11/2010	Modelagem para organização e representação do conhecimento em ontologias de domínio: uma experiência na área da cultura do sorgo	Andres Manuel Villafuerte Oyola
G	4	15/04/2011	Qualidade da informação e produsage: semiótica, informação e o usuário antropofágico	Joana Ziller de Araujo Josephson
G	5	18/08/2011	Comportamento informacional na tomada de decisão: proposta de Modelo Integrativo	Frederico Cesar Mafra

Seção do *corpus*	No	Data da Publicação	Título	Autor (orientando)
				Pereira
G	6	05/11/2010	Processamento de linguagem natural: caracterizacao da produção científica dos pesquisadores brasileiros	Ana Paula Ladeira
G	7	04/02/2010	Análise de domínio organizacional na perspectiva arquivística: potencialidade no uso da metodologia DIRKS - Designing and Implementing Recordkeeping Systems	Celia da Consolacao Dias
G	8	30/11/2009	Conformação de regime de informação: a experiência do arranjo produtivo local de eletrônica de Santa Rita do Sapucaí - MG	Adriane Maria Arantes de Carvalho
H	1	08/11/2010	Significado de humanização da assistência para os profissionais de saúde que atendem na sala de emergência de um pronto-socorro	Mercia Aleide Ribeiro Leite
H	2	21/03/2011	Processos avaliativos no curso de medicina: desempenho dos estudantes em relação às competências em pediatria e sua significação pelo docente	Luiz Megale
H	3	28/03/2011	Telessaúde na atenção primária: uma experiência do distrito sanitário Centro-Sul de Belo Horizonte	Edson Jose Carpintero Rezende
H	4	26/02/2010	Declinações da dismorfofobia: estudo psicanalítico da distorção da imagem corporal	Musso Garcia Greco
H	5	07/04/2011	Avaliação da qualidade de vida dos adolescentes em tratamento oncológico no Hospital das Clínicas da Universidade Federal de Minas Gerais	Karla Emilia de Sa Rodrigues
H	6	15/03/2010	Deficiência de vitamina A e fatores associados em crianças e adolescente em dois municípios do semiárido de Minas Gerais	Romero Alves Teixeira
H	7	26/03/2010	Acesso venoso central percutâneo , via veia jugular externa, pelatécnica de Seldinger em crianças: é imprescindível a inserção do fio guia até a veia cava superior para o sucesso do cateterismo?	Paulo Custodio Furtado Cruzeiro

Fonte: Elaborado pelo autor.

APÊNDICE I - LISTA DOS SINTAGMAS NOMINAIS SELECIONADOS COMO CANDIDATOS A DESCRITORES

Legenda:

Doc.$_{cj}$ → documento (*corpus c* + número do documento *j*);

Pos. → posição da eleição do candidato;

f_{ijc} → frequência do sintagma nominal *i* no documento *j* do *corpus c*;

n_{ic} → número de documentos no *corpus 'c'* que contém o sintagma nominal *i*;

CNP_i → categoria do sintagma nominal (i);

$Score_{ijc}$ → pontuação como candidato do sintagma nominal (i) obtida a partir da Equação 2 na página 57;

$Relevância_{ijc}$ -> Avaliação da relevância do sintagma nominal como descritor dada pelo autor da tese (de 0 'Não Relevante' a 6 'Extremamente Relevante').

Doc.$_{cj}$	**Pos.**	**Sintagma nominal candidato (i)**	**f_{ijc}**	**n_{ic}**	**CNP_i**	**$Score_{ijc}$**	**$Relevância_{ijc}$ (autor)**
A01	1º	escolar dos filhos	62	1	1,4	0,3483	**4**
A01	2º	meses de abril	44	1	1,1	0,1942	**0**
A01	3º	professores do município	46	2	1,1	0,1587	**3**
A01	4º	questionário aplicado	46	1	0,8	0,1477	**3**
A01	5º	escolarização dos filhos	17	1	1,4	0,0955	**6**
A01	6º	escolha do estabelecimento	18	1	1,1	0,0794	**5**
A01	7º	maio e junho de 2009	18	1	1,1	0,0794	**0**
A01	8º	escolares dos filhos	13	1	1,4	0,0730	**4**
A01	9º	professora de ciências	20	2	1,1	0,0690	**0**

Doc.$_{cj}$	Pos.	Sintagma nominal candidato (i)	f_{ijc}	n_{ic}	CNP_i	$Score_{ijc}$	$Relevância_{ijc}$ (autor)
A01	10º	professor de ciências	19	2	1,1	0,0656	**0**
A01	11º	caso dos pais	11	1	1,4	0,0618	**0**
A01	12º	professores do grupo	21	3	1,1	0,0606	**3**
A01	13º	professora de geografia	17	2	1,1	0,0587	**0**
A01	14º	dois filhos	108	3	0,2	0,0567	**2**
A01	15º	professora de matemática	22	4	1,1	0,0547	**0**
A01	16º	famílias do grupo	12	1	1,1	0,0530	**4**
A01	17º	famílias fortemente orientadas para o sucesso	8	1	1,1	0,0353	**6**
A01	18º	escola dos filhos	6	1	1,4	0,0337	**5**
A01	19º	bom aluno	10	1	0,8	0,0321	**5**
A01	20º	total 114	10	1	0,8	0,0321	**3**
A02	1º	grifos meus	61	4	0,8	0,2121	**1**
A02	2º	dia do trabalho	24	1	1,1	0,2036	**5**
A02	3º	hemeroteca histórica da biblioteca pública	21	1	1,1	0,1781	**5**
A02	4º	estado novo	19	1	0,8	0,1172	**4**
A02	5º	livros de leitura	17	2	1,1	0,1127	**4**
A02	6º	acervo do museu da escola	8	1	1,4	0,0864	**5**
A02	7º	belo horizonte	90	15	0,8	0,0821	**3**
A02	8º	maio de 1951	8	1	1,1	0,0679	**1**
A02	9º	fig	40	1	0,2	0,0617	**3**
A02	10º	década de 1930	7	1	1,1	0,0594	**4**
A02	11º	museu da escola	7	1	1,1	0,0594	**5**

Doc.$_{cj}$	Pos.	Sintagma nominal candidato (i)	f_{ijc}	n_{ic}	CNP_i	$Score_{ijc}$	$Relevância_{ijc}$ (autor)
A02	12º	maria dos reis	5	1	1,4	0,0540	**0**
A02	13º	jornalfolha	33	1	0,2	0,0509	**3**
A02	14º	anos de 1930	6	1	1,1	0,0509	**4**
A02	15º	representações sobre o trabalho	6	1	1,1	0,0509	**6**
A02	16º	propaganda política	8	1	0,8	0,0493	**4**
A02	17º	vargas	77	7	0,2	0,0460	**4**
A02	18º	jornalestado	29	1	0,2	0,0447	**3**
A02	19º	intolerância	36	2	0,2	0,0434	**6**
A02	20º	centro de referência do professor	4	1	1,4	0,0432	**1**
A03	1º	forró eletrônico	130	1	0,8	1,1042	**6**
A03	2º	currículo do forró eletrônico	37	1	1,1	0,4321	**6**
A03	3º	músicas de forró eletrônico	17	1	1,1	0,1985	**5**
A03	4º	dispositivo pedagógico da nordestinidade	12	1	1,1	0,1401	**6**
A03	5º	discurso do forró eletrônico	10	1	1,1	0,1168	**6**
A03	6º	forrozeiro	51	1	0,2	0,1083	**4**
A03	7º	forró	48	1	0,2	0,1019	**4**
A03	8º	cultura da mídia	8	1	1,1	0,0934	**3**
A03	9º	processos de subjetivação	9	2	1,1	0,0822	**5**
A03	10º	aviões do forró	7	1	1,1	0,0818	**4**
A03	11º	louro	22	6	0,8	0,0815	**4**
A03	12º	madeira	12	2	0,8	0,0797	**2**
A03	13º	forrozeira	34	1	0,2	0,0722	**4**

Doc.$_{cj}$	**Pos.**	**Sintagma nominal candidato (i)**	**f**$_{ijc}$	**n**$_{ic}$	**CNP**$_{i}$	**Score**$_{ijc}$	**Relevância**$_{ijc}$ **(autor)**
A03	14º	currículo aqui investigado	8	1	0,8	0,0679	**0**
A03	15º	público forrozeiro	7	1	0,8	0,0595	**6**
A03	16º	discursos do forró eletrônico	5	1	1,1	0,0584	**6**
A03	17º	meio a relações de poder	5	1	1,1	0,0584	**4**
A03	18º	foucault	66	7	0,2	0,0543	**6**
A03	19º	gênero musical	8	2	0,8	0,0531	**3**
A03	20º	albuquerque	44	4	0,2	0,0527	**4**
A04	1º	ondas de dados	14	1	1,1	0,2104	**4**
A04	2º	entendimento dos estudantes	12	2	1,4	0,1795	**6**
A04	3º	engajamento cognitivo	16	1	0,8	0,1749	**6**
A04	4º	unidade de eletricidade	11	1	1,1	0,1654	**5**
A04	5º	coeficiente de separação entre as pessoas	8	1	1,4	0,1531	**5**
A04	6º	estrutura de covariância	10	1	1,1	0,1503	**6**
A04	7º	três ondas de dados	10	1	1,1	0,1503	**5**
A04	8º	etm de patologia	9	1	1,1	0,1353	**5**
A04	9º	itens da escala	9	1	1,1	0,1353	**6**
A04	10º	matriz de covariância	9	1	1,1	0,1353	**6**
A04	11º	estudo longitudinal	15	2	0,8	0,1282	**6**
A04	12º	corrente elétrica	11	1	0,8	0,1203	**3**
A04	13º	ocasiões de medida	7	1	1,1	0,1052	**4**
A04	14º	sistema categórico	9	1	0,8	0,0984	**6**
A04	15º	estudantes do curso	6	1	1,1	0,0902	**5**

Doc.$_{cj}$	**Pos.**	**Sintagma nominal candidato (i)**	**f**$_{ijc}$	**n**$_{ic}$	**CNP**$_{i}$	**Score**$_{ijc}$	**Relevância**$_{ijc}$ **(autor)**
A04	16º	etm de química	6	1	1,1	0,0902	**5**
A04	17º	campo elétrico	8	1	0,8	0,0875	**1**
A04	18º	rasch	28	1	0,2	0,0765	**6**
A04	19º	efeitos sobre a potência	5	1	1,1	0,0752	**1**
A04	20º	fenômeno da incandescência	5	1	1,1	0,0752	**1**
A05	1º	discurso em aulas de ciências	81	1	1,4	1,9323	**3**
A05	2º	estrutura de análise	72	1	1,1	1,3495	**3**
A05	3º	orientações discursivas	13	1	0,8	0,1772	**6**
A05	4º	pistas de contextualização	9	1	1,1	0,1687	**5**
A05	5º	quadro de narrativas	7	1	1,1	0,1312	**5**
A05	6º	quadro de apresentação das aulas	5	1	1,4	0,1193	**5**
A05	7º	salas de aula de ciências	6	2	1,4	0,1119	**5**
A05	8º	orientação discursiva	8	1	0,8	0,1091	**6**
A05	9º	teoria da atividade	7	2	1,1	0,1026	**6**
A05	10º	objetivo pragmático	7	1	0,8	0,0954	**5**
A05	11º	contraposição de ideias	5	1	1,1	0,0937	**6**
A05	12º	perspectiva do professor	5	1	1,1	0,0937	**3**
A05	13º	formação de professores de ciências	5	2	1,4	0,0933	**3**
A05	14º	estrutura analítica	10	3	0,8	0,0892	**5**
A05	15º	discurso em salas de aula de ciências	4	1	1,2	0,0818	**3**
A05	16º	pdd	23	1	0,2	0,0784	**6**
A05	17º	confirmação de um ponto de vista	3	1	1,4	0,0716	**3**

Doc.$_{cj}$	Pos.	Sintagma nominal candidato (i)	f_{ijc}	n_{ic}	CNP_i	$Score_{ijc}$	$Relevância_{ijc}$ (autor)
A05	18º	justificações recíprocas	5	1	0,8	0,0682	**6**
A05	19º	ponto mais alto	5	1	0,8	0,0682	**3**
A05	20º	ponto de vista do professor	4	3	1,4	0,0624	**3**
A06	1º	ênfase adicionada	27	1	0,8	0,2404	**0**
A06	2º	josé de alencar	13	1	1,1	0,1592	**2**
A06	3º	guimarães	124	5	0,2	0,1362	**2**
A06	4º	garganta do inferno	11	1	1,1	0,1347	**0**
A06	5º	machado de assis	12	2	1,1	0,1149	**2**
A06	6º	bernardo	66	2	0,2	0,1149	**1**
A06	7º	relatório de presidente da província	7	1	1,4	0,1091	**3**
A06	8º	noticiador	45	1	0,2	0,1002	**1**
A06	9º	ermitão de muquém	8	1	1,1	0,0980	**0**
A06	10º	duque de caxias	7	1	1,1	0,0857	**0**
A06	11º	baía de botafogo	6	1	1,1	0,0735	**0**
A06	12º	filha do fazendeiro	6	1	1,1	0,0735	**0**
A06	13º	tronco do ipê	6	1	1,1	0,0735	**1**
A06	14º	jupira	30	1	0,2	0,0668	**0**
A06	15º	ouro preto	11	3	0,8	0,0641	**1**
A06	16º	autor mineiro	7	1	0,8	0,0623	**1**
A06	17º	história de quilombolas	5	1	1,1	0,0612	**0**
A06	18º	canto épico	6	1	0,8	0,0534	**0**
A06	19º	heróides brasileiras	6	1	0,8	0,0534	**0**

Doc.$_{cj}$	Pos.	Sintagma nominal candidato (i)	f_{ijc}	n_{ic}	CNP_i	$Score_{ijc}$	$Relevância_{ijc}$ (autor)
A06	20º	personagens masculinos	6	1	0,8	0,0534	**5**
A07	1º	rede do ciclo	14	1	1,1	0,1950	**0**
A07	2º	ano do ciclo	13	3	1,1	0,1185	**4**
A07	3º	professores do ciclo	8	1	1,1	0,1114	**0**
A07	4º	gêneros primários	11	1	0,8	0,1114	**4**
A07	5º	meninas negras	11	1	0,8	0,1114	**0**
A07	6º	letramento adquiridas	10	1	0,8	0,1013	**3**
A07	7º	letramento ensinadas	10	1	0,8	0,1013	**3**
A07	8º	carta de amor	6	1	1,1	0,0836	**0**
A07	9º	escrita	45	14	0,8	0,0773	**6**
A07	10º	evento de letramento	7	2	1,1	0,0762	**6**
A07	11º	gêneros secundários	7	1	0,8	0,0709	**4**
A07	12º	escrita dos bilhetes	4	1	1,4	0,0709	**3**
A07	13º	coordenadora do projeto	5	1	1,1	0,0696	**0**
A07	14º	júlia	42	3	0,2	0,0696	**0**
A07	15º	patrick	34	2	0,2	0,0673	**0**
A07	16º	cultura escrita	8	2	0,8	0,0634	**6**
A07	17º	base alfabética	6	1	0,8	0,0608	**4**
A07	18º	tipos textuais	6	1	0,8	0,0608	**3**
A07	19º	vicente	42	4	0,2	0,0600	**0**
A07	20º	anúncio em piada	4	1	1,1	0,0557	**4**
A08	1º	índia	153	2	0,8	0,4167	**0**

Doc.$_{cj}$	Pos.	Sintagma nominal candidato (i)	f_{ijc}	n_{ic}	CNP_i	$Score_{ijc}$	$Relevância_{ijc}$ (autor)
A08	2º	cravo da índia	68	1	1,1	0,3257	**3**
A08	3º	extrato de cravo da índia	25	1	1,4	0,1524	**1**
A08	4º	juan	172	1	0,2	0,1498	**6**
A08	5º	thiago	172	1	0,2	0,1498	**6**
A08	6º	práticas epistêmicas	34	1	0,8	0,1184	**6**
A08	7º	teoria da atividade	23	2	1,1	0,0861	**5**
A08	8º	ciências por investigação	22	2	1,1	0,0824	**3**
A08	9º	atividade do grupo	17	1	1,1	0,0814	**6**
A08	10º	espécie de formiga	15	1	1,1	0,0718	**1**
A08	11º	repelente	77	1	0,2	0,0671	**3**
A08	12º	ensaios experimentais	19	1	0,8	0,0662	**3**
A08	13º	extrato	74	1	0,2	0,0644	**1**
A08	14º	ensaio experimental	18	1	0,8	0,0627	**3**
A08	15º	ana	317	12	0,2	0,0602	**4**
A08	16º	integrantes do grupo	25	5	1,1	0,0591	**3**
A08	17º	cebolinha	64	1	0,2	0,0557	**1**
A08	18º	atividade de investigação	11	1	1,1	0,0527	**6**
A08	19º	aulas de produção	11	1	1,1	0,0527	**3**
A08	20º	decisão do grupo	11	1	1,1	0,0527	**6**
A09	1º	teatro na escola	21	1	1,1	0,1540	**6**
A09	2º	arte na escola	19	1	1,1	0,1394	**6**
A09	3º	grifos meus	34	4	0,8	0,1022	**3**

Doc.$_{cj}$	Pos.	Sintagma nominal candidato (i)	f_{ijc}	n_{ic}	CNP_i	$Score_{ijc}$	$Relevância_{ijc}$ (autor)
A09	4º	artes visuais	17	2	0,8	0,0709	**2**
A09	5º	professor de arte	9	1	1,1	0,0660	**6**
A09	6º	teatro	207	12	0,2	0,0602	**4**
A09	7º	curriculares estaduais	11	1	0,8	0,0587	**4**
A09	8º	forma artística	11	1	0,8	0,0587	**3**
A09	9º	campo da arte	8	1	1,1	0,0587	**5**
A09	10º	teatro em suas aulas de arte	6	1	1,4	0,0560	**6**
A09	11º	jogo dramático	13	2	0,8	0,0542	**2**
A09	12º	grupos de teatro	9	2	1,1	0,0516	**2**
A09	13º	arte nas escolas	7	1	1,1	0,0513	**6**
A09	14º	campo do teatro na educação	5	1	1,4	0,0467	**6**
A09	15º	teatro em sala de aula	5	1	1,4	0,0467	**6**
A09	16º	trabalho com a arte na escola	5	1	1,4	0,0467	**3**
A09	17º	grupo de teatro	8	2	1,1	0,0459	**2**
A09	18º	artes cênicas	8	1	0,8	0,0427	**6**
A09	19º	área de arte	5	1	1,1	0,0367	**6**
A09	20º	escolarização do teatro	5	1	1,1	0,0367	**5**
A10	1º	banco de dados da pesquisa	47	2	1,4	0,2120	**0**
A10	2º	estado do acre	54	3	1,1	0,1601	**5**
A10	3º	qualidade da educação	23	1	1,1	0,1042	**4**
A10	4º	acre	191	3	0,2	0,1030	**6**
A10	5º	visão dos professores	16	1	1,4	0,0923	**3**

Doc.$_{cj}$	Pos.	Sintagma nominal candidato (i)	f_{ijc}	n_{ic}	CNP_i	$Score_{ijc}$	$Relevância_{ijc}$ (autor)
A10	6º	anos finais	50	5	0,8	0,0813	**0**
A10	7º	desempenho dos alunos	15	2	1,4	0,0677	**3**
A10	8º	anos iniciais	41	5	0,8	0,0667	**0**
A10	9º	etapa da pesquisa	29	6	1,1	0,0573	**0**
A10	10º	estado de educação	21	4	1,1	0,0537	**0**
A10	11º	professor em 1ºa	11	1	1,1	0,0499	**0**
A10	12º	participantes da pesquisa	13	2	1,1	0,0461	**0**
A10	13º	maioria dos docentes	14	4	1,4	0,0455	**0**
A10	14º	rio branco	19	3	0,8	0,0410	**3**
A10	15º	plano de carreira	9	1	1,1	0,0408	**6**
A10	16º	política de formação de professores	9	2	1,4	0,0406	**6**
A10	17º	total 240	12	1	0,8	0,0396	**0**
A10	18º	organização do trabalho	14	4	1,1	0,0358	**6**
A10	19º	exigências sobre o trabalho do professor	6	1	1,4	0,0346	**6**
A10	20º	jornada de trabalho	19	7	1,1	0,0334	**5**
A11	1º	escola de educação	116	2	1,1	0,2346	**0**
A11	2º	processo de escolarização da infância	32	1	1,4	0,1053	**6**
A11	3º	abordagem teórico-metodológica	36	1	0,8	0,0677	**0**
A11	4º	wanda	103	1	0,2	0,0484	**0**
A11	5º	cultura de pares	16	1	1,1	0,0414	**6**
A11	6º	oficina de artes	16	1	1,1	0,0414	**0**
A11	7º	érica	71	1	0,2	0,0334	**0**

Doc.$_{cj}$	Pos.	Sintagma nominal candidato (i)	f_{ijc}	n_{ic}	CNP_i	$Score_{ijc}$	$Relevância_{ijc}$ (autor)
A11	8º	vanessa	85	2	0,2	0,0313	**0**
A11	9º	páginas do livro	12	1	1,1	0,0310	**0**
A11	10º	sílvia	94	3	0,2	0,0289	**0**
A11	11º	balança a cabeça	15	1	0,8	0,0282	**0**
A11	12º	interações entre as crianças	10	1	1,1	0,0259	**6**
A11	13º	lúcio	68	2	0,2	0,0250	**0**
A11	14º	próxima página	13	1	0,8	0,0245	**0**
A11	15º	paula	102	5	0,2	0,0237	**0**
A11	16º	professora da turma	9	1	1,1	0,0233	**0**
A11	17º	amanda	46	1	0,2	0,0216	**0**
A11	18º	cantinho da fantasia	8	1	1,1	0,0207	**0**
A11	19º	corsaro	40	1	0,2	0,0188	**6**
A11	20º	isadora	40	1	0,2	0,0188	**0**
A12	1º	departamento de química	21	1	1,1	0,0852	**1**
A12	2º	análise das aulas	14	1	1,1	0,0568	**5**
A12	3º	aulas da professora	14	1	1,1	0,0568	**2**
A12	4º	quadro de giz	17	2	1,1	0,0540	**1**
A12	5º	aulas do professor	10	1	1,1	0,0406	**2**
A12	6º	tiago	64	2	0,2	0,0369	**0**
A12	7º	participação dos estudantes	9	2	1,4	0,0364	**4**
A12	8º	prática de sala de aula	9	2	1,4	0,0364	**6**
A12	9º	química	31	7	0,8	0,0355	**6**

Doc.$_{cj}$	Pos.	Sintagma nominal candidato (i)	f_{ijc}	n_{ic}	CNP_i	$Score_{ijc}$	$Relevância_{ijc}$ (autor)
A12	10º	classe de referentes	8	1	1,1	0,0325	**1**
A12	11º	tempo do estudante	8	1	1,1	0,0325	**2**
A12	12º	formadores de professores	12	3	1,1	0,0319	**5**
A12	13º	curso de licenciatura em química	6	1	1,4	0,0310	**4**
A12	14º	departamento de química da ufmg	6	1	1,4	0,0310	**1**
A12	15º	referente específico	10	1	0,8	0,0295	**2**
A12	16º	agenda de conteúdo	7	1	1,1	0,0284	**0**
A12	17º	aula na graduação	7	1	1,1	0,0284	**6**
A12	18º	sala de aula	87	19	1,1	0,0260	**3**
A12	19º	tipo de aula	8	2	1,1	0,0254	**3**
A12	20º	estratégias usadas por o professores	6	1	1,1	0,0244	**5**
A13	1º	dona maria	33	1	0,8	0,3196	**5**
A13	2º	escola normal	36	2	0,8	0,2726	**4**
A13	3º	cidade de sabará	19	1	1,1	0,2530	**5**
A13	4º	faculdade de educação	36	5	1,1	0,2366	**4**
A13	5º	cultura impressa e educação da mulher no século	13	1	1,4	0,2203	**5**
A13	6º	mestrado em educação	25	3	1,1	0,2179	**2**
A13	7º	escola normal de sabará	15	1	1,1	0,1998	**5**
A13	8º	faculdade de filosofia e ciências humanas	15	1	1,1	0,1998	**3**
A13	9º	ouro preto	31	3	0,8	0,1965	**1**
A13	10º	comarca do rio das velhas	11	1	1,4	0,1864	**5**
A13	11º	presença de mulheres na docência	11	1	1,4	0,1864	**5**

Doc.$_{cj}$	Pos.	Sintagma nominal candidato (i)	f_{ijc}	n_{ic}	CNP_i	$Score_{ijc}$	$Relevância_{ijc}$ (autor)
A13	12º	escolas normais	22	2	0,8	0,1666	**4**
A13	13º	belo horizonte	114	15	0,8	0,1633	**0**
A13	14º	sabará	86	2	0,2	0,1628	**4**
A13	15º	chefes de domicílio	12	1	1,1	0,1598	**2**
A13	16º	velhas	33	5	0,8	0,1578	**1**
A13	17º	museu do ouro	11	1	1,1	0,1465	**1**
A13	18º	governo dos pobres em sabará	10	1	1,2	0,1453	**1**
A13	19º	centro de história da família	8	1	1,4	0,1356	**2**
A13	20º	editora da fundação	10	1	1,1	0,1332	**0**
A14	1º	modelo científico	22	1	0,8	0,2358	**6**
A14	2º	análise de concentração	16	1	1,1	0,2358	**6**
A14	3º	primeiro ano segundo ano terceiro	16	1	1,1	0,2358	**1**
A14	4º	concatenação de influências	13	1	1,1	0,1916	**0**
A14	5º	análise de modelos	11	1	1,1	0,1621	**6**
A14	6º	estado de modelo	11	1	1,1	0,1621	**6**
A14	7º	modelos intuitivos	14	1	0,8	0,1501	**6**
A14	8º	autovalores e autovetores para o bloco de questões	8	1	1,4	0,1501	**6**
A14	9º	fator de concentração	10	1	1,1	0,1474	**6**
A14	10º	modelos mentais	13	1	0,8	0,1394	**6**
A14	11º	aplicação do fci	9	1	1,1	0,1327	**6**
A14	12º	classe de coordenação	9	1	1,1	0,1327	**6**
A14	13º	diferentes modelos	11	1	0,8	0,1179	**6**

Doc.$_{cj}$	Pos.	Sintagma nominal candidato (i)	f_{ijc}	n_{ic}	CNP_i	$Score_{ijc}$	$Relevância_{ijc}$ (autor)
A14	14º	análise de variância	8	1	1,1	0,1179	**1**
A14	15º	estudantes do primeiro ano	8	1	1,1	0,1179	**1**
A14	16º	conhecimento dos estudantes	6	1	1,4	0,1126	**6**
A14	17º	densidade para o bloco de questões	6	1	1,4	0,1126	**4**
A14	18º	distribuição das questões do fci	6	1	1,4	0,1126	**6**
A14	19º	significância da inconsistência dos modelos	7	1	1,2	0,1126	**6**
A14	20º	autovetor associado	10	1	0,8	0,1072	**6**
A15	1º	banco de dados da pesquisa	59	2	1,4	0,4370	**6**
A15	2º	estado do acre	48	3	1,1	0,2337	**6**
A15	3º	coordenadores administrativos	48	2	0,8	0,2031	**6**
A15	4º	acre	173	3	0,2	0,1532	**6**
A15	5º	gestão democrática	24	2	0,8	0,1016	**6**
A15	6º	trabalho do núcleo	13	1	1,1	0,0968	**6**
A15	7º	rio branco	27	3	0,8	0,0956	**5**
A15	8º	lei estadual	20	2	0,8	0,0846	**6**
A15	9º	gestor das escolas	10	1	1,1	0,0744	**6**
A15	10º	gestor	79	3	0,2	0,0699	**4**
A15	11º	gestor da escola	12	2	1,1	0,0698	**6**
A15	12º	coordenador administrativo	16	2	0,8	0,0677	**6**
A15	13º	governos da frente popular	9	1	1,1	0,0670	**4**
A15	14º	frente popular	12	1	0,8	0,0650	**3**
A15	15º	núcleo de direção	8	1	1,1	0,0595	**6**

Doc.$_{cj}$	Pos.	Sintagma nominal candidato (i)	f_{ijc}	n_{ic}	CNP_i	$Score_{ijc}$	$Relevância_{ijc}$ (autor)
A15	16º	profissionais da educação	14	4	1,1	0,0587	**6**
A15	17º	autonomia da escola	10	2	1,1	0,0582	**5**
A15	18º	plano de governo	7	1	1,1	0,0521	**5**
A15	19º	educação do estado do acre	7	2	1,4	0,0518	**6**
A15	20º	conceito de regulação	10	3	1,1	0,0487	**4**
A16	1º	docente em matemática à distância	19	1	1,4	0,1873	**3**
A16	2º	docente em matemática	24	1	1,1	0,1859	**5**
A16	3º	curso à distância	23	1	1,1	0,1782	**4**
A16	4º	curso de matemática	17	1	1,1	0,1317	**1**
A16	5º	educação à distância	20	2	1,1	0,1211	**5**
A16	6º	cursos à distância	14	1	1,1	0,1084	**4**
A16	7º	curso de matemática à distância	11	1	1,4	0,1084	**0**
A16	8º	formação de professores de matemática	14	2	1,4	0,1079	**5**
A16	9º	curso de licenciatura em matemática à distância	12	1	1,2	0,1014	**6**
A16	10º	curso de licenciatura em matemática	10	1	1,4	0,0986	**4**
A16	11º	licenciatura em matemática à distância	10	1	1,4	0,0986	**6**
A16	12º	lincoln	94	3	0,2	0,0866	**5**
A16	13º	excerto do memorial	11	1	1,1	0,0852	**4**
A16	14º	modalidade à distância	14	2	1,1	0,0848	**5**
A16	15º	excerto de etapa	10	1	1,1	0,0775	**4**
A16	16º	experiência da constituição	10	1	1,1	0,0775	**6**
A16	17º	modalidade de educação à distância	9	2	1,4	0,0694	**5**

Doc.$_{cj}$	Pos.	Sintagma nominal candidato (i)	f_{ijc}	n_{ic}	CNP_i	$Score_{ijc}$	$Relevância_{ijc}$ (autor)
A16	18º	formação em matemática à distância	7	1	1,4	0,0690	**6**
A16	19º	possível encontro	12	1	0,8	0,0676	**6**
A16	20º	professores de matemática	11	2	1,1	0,0666	**4**
A17	1º	herbert	327	4	0,2	0,1162	**0**
A17	2º	heliane	158	1	0,2	0,0996	**0**
A17	3º	andréia	115	1	0,2	0,0725	**0**
A17	4º	desenvolvimento profissional	65	6	0,8	0,0715	**6**
A17	5º	laboratório de informática	28	3	1,1	0,0635	**3**
A17	6º	mariano	185	5	0,2	0,0575	**0**
A17	7º	vyasa	84	1	0,2	0,0529	**0**
A17	8º	sala de informática	15	1	1,1	0,0520	**3**
A17	9º	e-group	73	1	0,2	0,0460	**6**
A17	10º	professor de história	20	4	1,1	0,0391	**6**
A17	11º	tice	62	1	0,2	0,0391	**6**
A17	12º	desenvolvimento profissional dos professores	13	3	1,4	0,0375	**6**
A17	13º	professores do grupo	15	3	1,1	0,0340	**6**
A17	14º	professores de história	12	2	1,1	0,0325	**6**
A17	15º	letrado em história	9	1	1,1	0,0312	**0**
A17	16º	tecnologias de informação e comunicação aplicadas à educação	7	1	1,4	0,0309	**6**
A17	17º	grupo virtual	12	1	0,8	0,0303	**6**
A17	18º	ambiente virtual	15	2	0,8	0,0296	**6**

Doc.$_{cj}$	Pos.	Sintagma nominal candidato (i)	f_{ijc}	n_{ic}	CNP_i	$Score_{ijc}$	$Relevância_{ijc}$ (autor)
A17	19º	grupo de trabalho	14	4	1,1	0,0274	**6**
A17	20º	desenvolvimento profissional de professores de história	6	1	1,4	0,0265	**6**
A18	1º	gdpf	147	1	0,2	0,1870	**5**
A18	2º	reuniões do gdpf	25	1	1,1	0,1749	**4**
A18	3º	reuniões do grupo	17	1	1,1	0,1189	**0**
A18	4º	teoria da ação	16	1	1,1	0,1119	**4**
A18	5º	licenciatura curta física	15	1	0,8	0,0763	**1**
A18	6º	desenvolvimento profissional do professor	13	2	1,1	0,0711	**6**
A18	7º	aprofundamento de conteúdo	9	1	1,1	0,0630	**0**
A18	8º	professores membros do gdpf	9	1	1,1	0,0630	**4**
A18	9º	vivência no gdpf	9	1	1,1	0,0630	**5**
A18	10º	jederson	46	1	0,2	0,0585	**0**
A18	11º	sessão plenária	11	1	0,8	0,0560	**0**
A18	12º	concepção do gdpf	8	1	1,1	0,0560	**4**
A18	13º	desenvolvimento do professor	8	1	1,1	0,0560	**5**
A18	14º	possibilidades de aplicação	8	1	1,1	0,0560	**0**
A18	15º	processo de conscientização	8	1	1,1	0,0560	**6**
A18	16º	qualidade das interações	8	1	1,1	0,0560	**5**
A18	17º	estudo exploratório	10	1	0,8	0,0509	**1**
A18	18º	física	50	13	0,8	0,0491	**4**
A18	19º	conscientização crítica da condição	7	1	1,1	0,0490	**5**
A18	20º	encontros do gdpf	7	1	1,1	0,0490	**1**

Doc.$_{cj}$	Pos.	Sintagma nominal candidato (i)	f_{ijc}	n_{ic}	CNP_i	$Score_{ijc}$	$Relevância_{ijc}$ (autor)
A19	1º	mulheres do bethânia	30	1	1,1	0,1946	**3**
A19	2º	grupo de mulheres	24	1	1,1	0,1557	**6**
A19	3º	história das mulheres	29	3	1,1	0,1231	**6**
A19	4º	grupos de mulheres	18	1	1,1	0,1168	**6**
A19	5º	movimento de mulheres de ipatinga	14	1	1,4	0,1156	**5**
A19	6º	associação de mulheres do bairro	12	1	1,4	0,0991	**4**
A19	7º	bethânia	72	1	0,2	0,0849	**1**
A19	8º	mulheres de ipatinga	13	1	1,1	0,0843	**4**
A19	9º	clube de mães	12	1	1,1	0,0779	**4**
A19	10º	associadas	25	5	0,8	0,0582	**0**
A19	11º	clubes de mães	8	1	1,1	0,0519	**4**
A19	12º	município de ipatinga	8	1	1,1	0,0519	**4**
A19	13º	prefeitura municipal de ipatinga	8	1	1,1	0,0519	**0**
A19	14º	participantes dos grupos	6	1	1,4	0,0495	**1**
A19	15º	trabalhos manuais	16	3	0,8	0,0494	**3**
A19	16º	casa própria	13	2	0,8	0,0480	**0**
A19	17º	ação social	10	1	0,8	0,0472	**5**
A19	18º	assistentes sociais	9	1	0,8	0,0425	**0**
A19	19º	integrantes dos grupos	5	1	1,4	0,0413	**2**
A19	20º	mulheres da associação do bethânia	5	1	1,4	0,0413	**3**
A20	1º	depoimento gravado em vídeo	143	1	1,1	0,8650	**0**
A20	2º	sessão de rede	100	1	1,1	0,6049	**0**

Doc.$_{cj}$	Pos.	Sintagma nominal candidato (i)	f_{ijc}	n_{ic}	CNP_i	$Score_{ijc}$	$Relevância_{ijc}$ (autor)
A20	3º	turno da manhã	63	3	1,1	0,2493	**0**
A20	4º	história da áfrica	25	1	1,1	0,1512	**4**
A20	5º	professores de história	25	2	1,1	0,1182	**3**
A20	6º	educação das relações étnico-raciais	15	1	1,1	0,0907	**6**
A20	7º	sessões de redes	15	1	1,1	0,0907	**0**
A20	8º	história e cultura africana e afro-brasileira	18	1	0,8	0,0792	**6**
A20	9º	município de contagem	13	1	1,1	0,0786	**1**
A20	10º	diversos professores	17	1	0,8	0,0748	**0**
A20	11º	questão racial	17	1	0,8	0,0748	**4**
A20	12º	curriculares nacionais para a educação das relações étnico-raciais	9	1	1,4	0,0693	**1**
A20	13º	conhecimentos históricos	17	2	0,8	0,0585	**3**
A20	14º	discriminação racial	13	1	0,8	0,0572	**2**
A20	15º	inúmeros outros	13	1	0,8	0,0572	**0**
A20	16º	históricos escolares	16	2	0,8	0,0550	**0**
A20	17º	fins dos anos	7	1	1,4	0,0539	**0**
A20	18º	texto das diretrizes	8	1	1,1	0,0484	**0**
A20	19º	interior das escolas	12	3	1,1	0,0475	**0**
A20	20º	promulgação da lei	7	1	1,1	0,0423	**0**
A21	1º	projetos de trabalho	18	1	1,1	0,1872	**6**
A21	2º	villas boas	24	1	0,8	0,1815	**3**
A21	3º	excerto do portfólio da estudante	10	1	1,4	0,1323	**3**

Doc.$_{cj}$	Pos.	Sintagma nominal candidato (i)	f_{ijc}	n_{ic}	CNP_i	$Score_{ijc}$	$Relevância_{ijc}$ (autor)
A21	4º	portfólio	94	3	0,2	0,1163	**6**
A21	5º	oportunidades formativas	12	1	0,8	0,0908	**6**
A21	6º	professor legal	11	1	0,8	0,0832	**3**
A21	7º	projeto de trabalho	8	1	1,1	0,0832	**6**
A21	8º	apresentação de projeto de trabalho	5	1	1,4	0,0662	**1**
A21	9º	portfólios	33	1	0,2	0,0624	**6**
A21	10º	social cognitive	7	1	0,8	0,0529	**6**
A21	11º	excerto do portfólio de evandro	4	1	1,4	0,0529	**3**
A21	12º	fotografia	49	4	0,2	0,0522	**6**
A21	13º	alunos por a autora	5	1	1,1	0,0520	**3**
A21	14º	ano de referência	5	1	1,1	0,0520	**3**
A21	15º	cenas das aulas	5	1	1,1	0,0520	**6**
A21	16º	concepções de avaliação	5	1	1,1	0,0520	**6**
A21	17º	encontro casual entre a desobediência e escrita	5	1	1,1	0,0520	**3**
A21	18º	excerto do texto escrito e apresentado	5	1	1,1	0,0520	**3**
A21	19º	leitão de almeida	5	1	1,1	0,0520	**1**
A21	20º	uso do portfólio	5	1	1,1	0,0520	**6**
A22	1º	membros da casa	29	1	1,1	0,1983	**2**
A22	2º	umbanda	154	1	0,2	0,1915	**5**
A22	3º	terreiro	129	1	0,2	0,1604	**3**
A22	4º	comunidade de prática	24	2	1,1	0,1283	**6**
A22	5º	umbandista	87	1	0,2	0,1082	**5**

Doc.$_{cj}$	Pos.	Sintagma nominal candidato (i)	f_{ijc}	n_{ic}	CNP_i	$Score_{ijc}$	$Relevância_{ijc}$ (autor)
A22	6º	terreiro de umbanda	15	1	1,1	0,1026	**4**
A22	7º	conversa gravada	18	1	0,8	0,0895	**0**
A22	8º	umbandistas	71	1	0,2	0,0883	**5**
A22	9º	ingold	63	1	0,2	0,0783	**6**
A22	10º	ogã	60	1	0,2	0,0746	**2**
A22	11º	médium	53	1	0,2	0,0659	**2**
A22	12º	gravada	12	1	0,8	0,0597	**0**
A22	13º	sessão semanal	12	1	0,8	0,0597	**2**
A22	14º	wenger	72	3	0,2	0,0586	**6**
A22	15º	dona	17	3	0,8	0,0553	**0**
A22	16º	mãe-pequena jnt	11	1	0,8	0,0547	**1**
A22	17º	prática de umbanda	8	1	1,1	0,0547	**3**
A22	18º	estratégias de aprendizagem	10	2	1,1	0,0535	**6**
A22	19º	prática religiosa	10	1	0,8	0,0497	**3**
A22	20º	casa de culto	7	1	1,1	0,0479	**2**
A23	1º	1995-2008	144	1	0,2	0,2160	**1**
A23	2º	superior no brasil	29	3	1,1	0,1566	**1**
A23	3º	graduação em odontologia	15	1	1,1	0,1238	**5**
A23	4º	curso de odontologia	13	1	1,1	0,1073	**5**
A23	5º	período estudado	16	1	0,8	0,0960	**2**
A23	6º	setor privado	19	3	0,8	0,0746	**5**
A23	7º	expansão da educação	9	1	1,1	0,0743	**4**

Doc.$_{cj}$	Pos.	Sintagma nominal candidato (i)	f_{ijc}	n_{ic}	CNP_i	$Score_{ijc}$	$Relevância_{ijc}$ (autor)
A23	8º	graduação em odontologia no brasil	7	1	1,4	0,0735	**6**
A23	9º	formação em odontologia	8	1	1,1	0,0660	**6**
A23	10º	tab	43	1	0,2	0,0645	**0**
A23	11º	comissões de especialistas	7	1	1,1	0,0578	**2**
A23	12º	conselho nacional de saúde	7	1	1,1	0,0578	**2**
A23	13º	cursos de odontologia	7	1	1,1	0,0578	**3**
A23	14º	expansão e democratização da educação	7	1	1,1	0,0578	**6**
A23	15º	superior do governo	7	1	1,1	0,0578	**1**
A23	16º	cursos de graduação em odontologia	5	1	1,4	0,0525	**6**
A23	17º	projeto de lei	8	2	1,1	0,0516	**1**
A23	18º	cursos de graduação	13	6	1,1	0,0468	**1**
A23	19º	regiões do brasil	7	2	1,1	0,0452	**1**
A23	20º	setor público	11	3	0,8	0,0432	**5**
A24	1º	episódio do currículo	30	1	1,1	0,1745	**0**
A24	2º	currículo do orkut	27	1	1,1	0,1571	**6**
A24	3º	orkut	261	6	0,2	0,1204	**6**
A24	4º	episódio do currículo do orkut	14	1	1,4	0,1036	**0**
A24	5º	ciborgue	85	1	0,2	0,0899	**6**
A24	6º	tecnologia da zuação	13	1	1,1	0,0756	**4**
A24	7º	anjos do orkut	12	1	1,1	0,0698	**0**
A24	8º	turma de ano	10	1	1,1	0,0582	**0**
A24	9º	tecnologia da liberdade	9	1	1,1	0,0524	**4**

Doc.$_{cj}$	Pos.	Sintagma nominal candidato (i)	f_{ijc}	n_{ic}	CNP_i	$Score_{ijc}$	$Relevância_{ijc}$ (autor)
A24	10º	louro	27	6	0,8	0,0498	**0**
A24	11º	relações de poder	34	11	1,1	0,0486	**6**
A24	12º	garotas	57	2	0,2	0,0471	**0**
A24	13º	discursos analisados	11	1	0,8	0,0465	**4**
A24	14º	escolar e currículo do orkut	8	1	1,1	0,0465	**0**
A24	15º	processo de produção das subjetividades	6	1	1,4	0,0444	**6**
A24	16º	processo de produção de subjetividades	6	1	1,4	0,0444	**6**
A24	17º	coltec	51	2	0,2	0,0422	**3**
A24	18º	comunidades do orkut	7	1	1,1	0,0407	**4**
A24	19º	produção das subjetividades	7	1	1,1	0,0407	**6**
A24	20º	juvenil	56	3	0,2	0,0388	**6**
B01	1º	tipo resultado estratégico	8	1	0,8	0,2752	**5**
B01	2º	entrevistados	28	1	0,2	0,2408	**6**
B01	3º	bezerro	6	1	0,8	0,2064	**6**
B01	4º	representações sociais	6	1	0,8	0,2064	**6**
B01	5º	maioria das vezes	4	1	1,1	0,1892	**0**
B01	6º	produção de leite	7	4	1,1	0,1656	**5**
B01	7º	carrapato	19	1	0,2	0,1634	**6**
B01	8º	gente	18	1	0,2	0,1548	**6**
B01	9º	grau de escolaridade	3	1	1,1	0,1419	**5**
B01	10º	uso de epi	3	1	1,1	0,1419	**5**
B01	11º	uso do epi	3	1	1,1	0,1419	**5**

Doc.$_{cj}$	Pos.	Sintagma nominal candidato (i)	f_{ijc}	n_{ic}	CNP_i	$Score_{ijc}$	$Relevância_{ijc}$ (autor)
B01	12º	extensão rural	4	1	0,8	0,1376	**5**
B01	13º	funcionários	4	1	0,8	0,1376	**5**
B01	14º	novas tecnologias	4	1	0,8	0,1376	**5**
B01	15º	parasitos	21	2	0,2	0,1355	**6**
B01	16º	microplus	15	1	0,2	0,1290	**4**
B01	17º	ee1	14	1	0,2	0,1204	**6**
B01	18º	dias após a inoculação das larvas	2	1	1,4	0,1204	**1**
B01	19º	eclodibilidade e a viabilidade das larvas no ambiente	2	1	1,4	0,1204	**5**
B01	20º	escala de produção de leite	2	1	1,4	0,1204	**3**
B02	1º	estádio do desenvolvimento sexual	20	1	1,1	0,2976	**4**
B02	2º	concentração de leptina	17	1	1,1	0,2530	**5**
B02	3º	dias à puberdade	17	1	1,1	0,2530	**4**
B02	4º	concentração de insulina	16	1	1,1	0,2381	**5**
B02	5º	afinidade à heparina	16	2	1,1	0,1786	**1**
B02	6º	andrológica por pontos	16	2	1,1	0,1786	**3**
B02	7º	protéicos com afinidade à heparina	8	1	1,4	0,1515	**2**
B02	8º	espermática	89	3	0,2	0,1454	**3**
B02	9º	animais reg	12	1	0,8	0,1299	**4**
B02	10º	touros da raça	11	2	1,1	0,1228	**0**
B02	11º	pico com afinidade	8	1	1,1	0,1191	**0**
B02	12º	journal animal	14	2	0,8	0,1136	**0**
B02	13º	bovine seminal	13	2	0,8	0,1055	**5**

Doc.$_{cj}$	Pos.	Sintagma nominal candidato (i)	f_{ijc}	n_{ic}	CNP_i	$Score_{ijc}$	$Relevância_{ijc}$ (autor)
B02	14º	estádio de desenvolvimento sexual	7	1	1,1	0,1042	**4**
B02	15º	estádios de desenvolvimento sexual	7	1	1,1	0,1042	**4**
B02	16º	desenvolvimento sexual	12	2	0,8	0,0974	**4**
B02	17º	insulin	34	1	0,2	0,0920	**4**
B02	18º	idade à puberdade	8	2	1,1	0,0893	**4**
B02	19º	meses de idade	14	5	1,1	0,0874	**0**
B02	20º	animais prec	10	2	0,8	0,0812	**4**
B03	1º	perfringens	100	1	0,2	0,2408	**4**
B03	2º	difficile	73	1	0,2	0,1758	**4**
B03	3º	sete dias de vida	11	1	1,1	0,1457	**4**
B03	4º	clínico de diarreia	10	1	1,1	0,1325	**3**
B03	5º	detecção das toxinas	8	1	1,1	0,1060	**6**
B03	6º	mecanismo de ação	8	1	1,1	0,1060	**1**
B03	7º	beta de clostridium	7	1	1,1	0,0927	**4**
B03	8º	clostridium	33	1	0,2	0,0795	**6**
B03	9º	concentração inibitória mínima	7	1	0,8	0,0674	**6**
B03	10º	50µl de mem e 50µl de células	4	1	1,4	0,0674	**0**
B03	11º	médio do título de antitoxina	4	1	1,4	0,0674	**0**
B03	12º	alfa de clostridium	5	1	1,1	0,0662	**4**
B03	13º	difficile em leitões	5	1	1,1	0,0662	**4**
B03	14º	presença do gene	5	1	1,1	0,0662	**1**
B03	15º	titulação da antitoxina	5	1	1,1	0,0662	**4**

Doc.$_{cj}$	Pos.	Sintagma nominal candidato (i)	f_{ijc}	n_{ic}	CNP_i	$Score_{ijc}$	$Relevância_{ijc}$ (autor)
B03	16º	experimental polivalente	6	1	0,8	0,0578	**3**
B03	17º	edema de mesocólon	4	1	1,1	0,0530	**6**
B03	18º	titulação de antitoxina	4	1	1,1	0,0530	**4**
B03	19º	título de antitoxina	4	1	1,1	0,0530	**4**
B03	20º	100µl de mem e 50µl de células	3	1	1,4	0,0506	**0**
B04	1º	intracellularis	118	1	0,2	0,2408	**4**
B04	2º	cultura pura	16	1	0,8	0,1306	**3**
B04	3º	pilosicoli	48	1	0,2	0,0980	**4**
B04	4º	hyodysenteriae	37	1	0,2	0,0755	**4**
B04	5º	marcador de pares de base	5	1	1,4	0,0714	**0**
B04	6º	sorotipo	28	1	0,2	0,0571	**2**
B04	7º	dias após a inoculação	5	1	1,1	0,0561	**4**
B04	8º	enterica	35	2	0,2	0,0536	**6**
B04	9º	mbh	26	1	0,2	0,0531	**1**
B04	10º	typhimurium	24	1	0,2	0,0490	**4**
B04	11º	ihq	23	1	0,2	0,0469	**4**
B04	12º	jacobson	23	1	0,2	0,0469	**0**
B04	13º	salmonella	30	2	0,2	0,0459	**4**
B04	14º	área metropolitana de belo horizonte	4	1	1,1	0,0449	**3**
B04	15º	camundongos da linhagem	4	1	1,1	0,0449	**6**
B04	16º	fímbria de adesão	4	1	1,1	0,0449	**2**
B04	17º	bactérias por grama de fezes	3	1	1,4	0,0429	**2**

Doc.$_{cj}$	Pos.	Sintagma nominal candidato (i)	f_{ijc}	n_{ic}	CNP_i	$Score_{ijc}$	$Relevância_{ijc}$ (autor)
B04	18º	causadores de diarreia	5	2	1,1	0,0421	**6**
B04	19º	marcação positiva	5	1	0,8	0,0408	**3**
B04	20º	terminação	26	2	0,2	0,0398	**6**
B05	1º	animais precoces	16	2	0,8	0,2359	**6**
B05	2º	meses de idade	18	5	1,1	0,2041	**4**
B05	3º	journal animal	13	2	0,8	0,1917	**0**
B05	4º	dias após a puberdade	9	2	1,1	0,1825	**6**
B05	5º	animais não-precoces	9	1	0,8	0,1769	**6**
B05	6º	espermáticos maiores	9	1	0,8	0,1769	**5**
B05	7º	congresso brasileiro de reprodução animal	6	1	1,1	0,1622	**0**
B05	8º	bovine seminal	10	2	0,8	0,1474	**5**
B05	9º	espermática	49	3	0,2	0,1454	**5**
B05	10º	belo horizonte	17	5	0,8	0,1402	**2**
B05	11º	glândulas sexuais acessórias	9	2	0,8	0,1327	**5**
B05	12º	maturidade sexual	11	3	0,8	0,1306	**6**
B05	13º	bulls	35	2	0,2	0,1290	**6**
B05	14º	idades em relação à puberdade	5	2	1,4	0,1290	**6**
B05	15º	modelo de regressão	6	2	1,1	0,1216	**3**
B05	16º	espermática progressiva	6	1	0,8	0,1180	**5**
B05	17º	animal reproduction	8	2	0,8	0,1180	**6**
B05	18º	linha vermelha	8	2	0,8	0,1180	**0**
B05	19º	10% de motilidade	4	1	1,1	0,1081	**4**

Doc.$_{cj}$	Pos.	Sintagma nominal candidato (i)	f_{ijc}	n_{ic}	CNP_i	$Score_{ijc}$	$Relevância_{ijc}$ (autor)
B05	20º	relação à idade	4	1	1,1	0,1081	**4**
B06	1º	traqueal	127	1	0,2	0,2408	**1**
B06	2º	portadores de colapso	22	1	1,1	0,2294	**0**
B06	3º	traqueal dorsal	18	1	0,8	0,1365	**0**
B06	4º	hialina do anel	9	1	1,1	0,0939	**2**
B06	5º	radiográfica compatível com colapso	9	1	1,1	0,0939	**3**
B06	6º	colapso de traquéia	7	1	1,1	0,0730	**6**
B06	7º	alargamento da membrana	6	1	1,1	0,0626	**2**
B06	8º	porcentagem de animais com imagem	4	1	1,4	0,0531	**0**
B06	9º	diminuição do lúmen	5	1	1,1	0,0521	**3**
B06	10º	traquéia	22	1	0,2	0,0417	**6**
B06	11º	área de substituição	4	1	1,1	0,0417	**0**
B06	12º	cães com colapso	4	1	1,1	0,0417	**1**
B06	13º	coloração de safranina	4	1	1,1	0,0417	**3**
B06	14º	deficiência de gags	4	1	1,1	0,0417	**3**
B06	15º	portador de colapso	4	1	1,1	0,0417	**0**
B06	16º	região da transição	4	1	1,1	0,0417	**0**
B06	17º	traqueal em diferentes faixas etárias	4	1	1,1	0,0417	**0**
B06	18º	redor dos condrócitos	3	1	1,4	0,0398	**0**
B06	19º	colapso	20	1	0,2	0,0379	**0**
B06	20º	radiograficamente positivos	5	1	0,8	0,0379	**0**
B07	1º	índices de mamite	21	1	1,1	0,3091	**4**

Doc$_{.cj}$	**Pos.**	**Sintagma nominal candidato (i)**	**f**$_{ijc}$	**n**$_{ic}$	**CNP**$_{i}$	**Score**$_{ijc}$	**Relevância**$_{ijc}$ **(autor)**
B07	2º	sul do estado	20	1	1,1	0,2943	**0**
B07	3º	aureus	90	1	0,2	0,2408	**0**
B07	4º	dezembro de 2006	16	2	1,1	0,1766	**0**
B07	5º	período de março de 2004	9	1	1,4	0,1686	**0**
B07	6º	leiteira da região	10	1	1,1	0,1472	**0**
B07	7º	período de janeiro de 2004	7	1	1,4	0,1311	**0**
B07	8º	mamite bovina	12	1	0,8	0,1284	**6**
B07	9º	agalactiae	45	1	0,2	0,1204	**0**
B07	10º	leiteiros do sul	7	1	1,1	0,1030	**0**
B07	11º	alterações na qualidade do leite	5	1	1,4	0,0937	**6**
B07	12º	subclínica da mamite	6	1	1,1	0,0883	**2**
B07	13º	mamite	42	2	0,2	0,0843	**5**
B07	14º	santos e fonseca	30	1	0,2	0,0803	**0**
B07	15º	staphylococcus	28	1	0,2	0,0749	**3**
B07	16º	aureus envolvidos na etiologia da mamite bovina	4	1	1,4	0,0749	**0**
B07	17º	índices de resistência	5	1	1,1	0,0736	**4**
B07	18º	leiteiras da região	5	1	1,1	0,0736	**0**
B07	19º	ponto de vista econômico	5	1	1,1	0,0736	**2**
B07	20º	precoce da mamite	5	1	1,1	0,0736	**0**
B08	1º	caev	33	1	0,2	0,2408	**6**
B08	2º	soroconversão tardia	8	1	0,8	0,2335	**6**
B08	3º	presença do caev	5	1	1,1	0,2007	**0**

Doc.$_{cj}$	Pos.	Sintagma nominal candidato (i)	f_{ijc}	n_{ic}	CNP_i	$Score_{ijc}$	$Relevância_{ijc}$ (autor)
B08	4º	pró-viral do caev	5	1	1,1	0,2007	**2**
B08	5º	idga	23	1	0,2	0,1678	**6**
B08	6º	vírus da imunodeficiência	4	1	1,1	0,1605	**0**
B08	7º	presença do caev no sêmen	3	1	1,4	0,1533	**6**
B08	8º	células do sistema	3	1	1,1	0,1204	**0**
B08	9º	mononucleares do sangue periférico	3	1	1,1	0,1204	**0**
B08	10º	imunodifusão em gel de agar	2	1	1,4	0,1022	**6**
B08	11º	infecção dos macrófagos	2	1	1,4	0,1022	**0**
B08	12º	regiões dos genes	2	1	1,4	0,1022	**0**
B08	13º	variação na detecção do caev	2	1	1,4	0,1022	**6**
B08	14º	la concha-bermejillo	3	1	0,8	0,0876	**0**
B08	15º	iniciadores externos	4	2	0,8	0,0876	**0**
B08	16º	amostra de campo	2	1	1,1	0,0803	**0**
B08	17º	bandas de 393	2	1	1,1	0,0803	**0**
B08	18º	caev no sêmen	2	1	1,1	0,0803	**6**
B08	19º	co-infectados com brucella	2	1	1,1	0,0803	**0**
B08	20º	erradicação da cae	2	1	1,1	0,0803	**3**
B09	1º	melitensis	80	1	0,2	0,2408	**3**
B09	2º	trato digestivo	19	1	0,8	0,2288	**5**
B09	3º	infecção por brucella	11	1	1,1	0,1821	**5**
B09	4º	brucella	60	2	0,2	0,1355	**6**
B09	5º	através do trato digestivo	7	1	1,1	0,1159	**4**

Doc.$_{cj}$	Pos.	Sintagma nominal candidato (i)	f_{ijc}	n_{ic}	CNP_i	$Score_{ijc}$	$Relevância_{ijc}$ (autor)
B09	6º	estabelecimento da infecção	7	1	1,1	0,1159	**2**
B09	7º	virulenta 16m	7	1	0,8	0,0843	**3**
B09	8º	infecção por brucella em camundongos	4	1	1,4	0,0843	**6**
B09	9º	infectados com brucella	5	1	1,1	0,0828	**2**
B09	10º	abortus	26	1	0,2	0,0783	**1**
B09	11º	ure1	24	1	0,2	0,0722	**2**
B09	12º	brucelose humana	6	1	0,8	0,0722	**4**
B09	13º	curso de infecção	4	1	1,1	0,0662	**2**
B09	14º	maturação de células	4	1	1,1	0,0662	**0**
B09	15º	requerimento da urease	4	1	1,1	0,0662	**3**
B09	16º	amostra virulenta	5	1	0,8	0,0602	**3**
B09	17º	mutantes	19	1	0,2	0,0572	**3**
B09	18º	16m	18	1	0,2	0,0542	**1**
B09	19º	10% de bile bovina	3	1	1,1	0,0497	**1**
B09	20º	16m e clones com resistência	3	1	1,1	0,0497	**1**
B10	1º	probabilidade de significância	32	1	1,1	0,3686	**3**
B10	2º	análise comparativa entre os anos de colheita	26	1	1,2	0,3267	**5**
B10	3º	arsênio	115	1	0,2	0,2408	**5**
B10	4º	cádmio	106	1	0,2	0,2220	**5**
B10	5º	ano da colheita medidas descritivas mínimo máximo mediana média	17	1	1,1	0,1958	**2**
B10	6º	ano da colheita	15	1	1,1	0,1728	**4**

Doc.$_{cj}$	**Pos.**	**Sintagma nominal candidato (i)**	**f**$_{ijc}$	**n**$_{ic}$	**CNP**$_{i}$	**Score**$_{ijc}$	**Relevância**$_{ijc}$ **(autor)**
B10	7º	máximo mediana média desvio	20	1	0,8	0,1675	**1**
B10	8º	nível de concentração	12	1	1,1	0,1382	**4**
B10	9º	efsa	55	1	0,2	0,1152	**1**
B10	10º	matéria seca	16	2	0,8	0,1005	**1**
B10	11º	análise comparativa	12	1	0,8	0,1005	**4**
B10	12º	fígado	89	4	0,2	0,0932	**4**
B10	13º	contaminação por arsênio	8	1	1,1	0,0921	**5**
B10	14º	apresentados os dados médios de contaminação	6	1	1,4	0,0880	**1**
B10	15º	anos de 2002	7	1	1,1	0,0806	**1**
B10	16º	avaliação das diferenças	7	1	1,1	0,0806	**5**
B10	17º	contaminação por cádmio	7	1	1,1	0,0806	**5**
B10	18º	rins	49	2	0,2	0,0770	**4**
B10	19º	tecido medidas descritivas mínimo máximo mediana média desvio	9	1	0,8	0,0754	**0**
B10	20º	metais estudados durante os anos de colheita	6	1	1,2	0,0754	**4**
B11	1º	estação chuvosa	88	2	0,8	0,3654	**4**
B11	2º	estação seca	81	2	0,8	0,3363	**4**
B11	3º	dias de lactação	40	1	1,1	0,3045	**4**
B11	4º	dias em relação	36	1	1,1	0,2740	**0**
B11	5º	escore da condição corporal	25	1	1,1	0,1903	**5**
B11	6º	base genética	29	1	0,8	0,1605	**4**
B11	7º	vacas de base	21	1	1,1	0,1599	**0**

Doc.$_{cj}$	Pos.	Sintagma nominal candidato (i)	f_{ijc}	n_{ic}	CNP_i	$Score_{ijc}$	$Relevância_{ijc}$ (autor)
B11	8º	folicular	102	1	0,2	0,1412	**4**
B11	9º	ovulação	102	1	0,2	0,1412	**4**
B11	10º	suplementadas no pré%parto	17	1	1,1	0,1294	**1**
B11	11º	pós%parto	88	1	0,2	0,1218	**4**
B11	12º	início da lactação	13	1	1,1	0,0990	**3**
B11	13º	produção de leite	25	4	1,1	0,0952	**4**
B11	14º	observou%se	68	1	0,2	0,0941	**0**
B11	15º	zebu média	17	1	0,8	0,0941	**0**
B11	16º	parto	112	3	0,2	0,0936	**4**
B11	17º	vacas	132	4	0,2	0,0913	**3**
B11	18º	vacas mestiças	14	1	0,8	0,0775	**4**
B11	19º	plasmáticas de colesterol	10	1	1,1	0,0761	**1**
B11	20º	plasmáticas de insulina	10	1	1,1	0,0761	**1**
B12	1º	veneno de tityus	52	1	1,1	0,4075	**6**
B12	2º	serrulatus	169	1	0,2	0,2408	**6**
B12	3º	diferentes tempos	35	1	0,8	0,1995	**4**
B12	4º	fasciolatus	130	1	0,2	0,1852	**6**
B12	5º	estatisticamente entre os grupos	17	1	1,4	0,1696	**2**
B12	6º	estatisticamente entre os tempos	17	1	1,4	0,1696	**2**
B12	7º	tityus	93	1	0,2	0,1325	**6**
B12	8º	camundongos inoculados	35	3	0,8	0,1204	**4**
B12	9º	letras maiúsculas	17	1	0,8	0,0969	**1**

Doc.$_{cj}$	Pos.	Sintagma nominal candidato (i)	f_{ijc}	n_{ic}	CNP_i	$Score_{ijc}$	$Relevância_{ijc}$ (autor)
B12	10º	letras minúsculas	17	1	0,8	0,0969	**1**
B12	11º	gráfico valores	16	1	0,8	0,0912	**3**
B12	12º	análise de variância	17	3	1,1	0,0804	**1**
B12	13º	canais de sódio	9	1	1,1	0,0705	**6**
B12	14º	veneno	61	2	0,2	0,0652	**4**
B12	15º	inoculação do veneno	8	1	1,1	0,0627	**6**
B12	16º	veneno de escorpião	7	1	1,1	0,0549	**5**
B12	17º	veneno de escorpiões	7	1	1,1	0,0549	**5**
B12	18º	concentração de hemoglobina	6	1	1,1	0,0470	**3**
B12	19º	possani	32	1	0,2	0,0456	**6**
B12	20º	9µg do veneno	5	1	1,1	0,0392	**5**
B13	1º	virus	72	1	0,2	0,2408	**6**
B13	2º	chicken	68	1	0,2	0,2274	**6**
B13	3º	cav	48	1	0,2	0,1605	**6**
B13	4º	quantificadas por leitura em espectrofotômetro	6	1	1,4	0,1405	**0**
B13	5º	anemia	63	3	0,2	0,1272	**6**
B13	6º	todd	31	1	0,2	0,1037	**0**
B13	7º	comercial avipro	7	1	0,8	0,0937	**0**
B13	8º	lohman animal	7	1	0,8	0,0937	**0**
B13	9º	vez no japão em 1979	4	1	1,4	0,0937	**0**
B13	10º	avian	25	1	0,2	0,0836	**6**
B13	11º	reação de sequenciamento	6	2	1,1	0,0828	**0**

Doc.$_{cj}$	Pos.	Sintagma nominal candidato (i)	f_{ijc}	n_{ic}	CNP_i	$Score_{ijc}$	$Relevância_{ijc}$ (autor)
B13	12º	virol	24	1	0,2	0,0803	**0**
B13	13º	visualização dos resultados das amplificações	4	1	1,2	0,0803	**0**
B13	14º	vp2	22	1	0,2	0,0736	**5**
B13	15º	ciclo inicial de desnaturação	4	1	1,1	0,0736	**0**
B13	16º	condições de amplificação	4	1	1,1	0,0736	**0**
B13	17º	corante de amostra	4	1	1,1	0,0736	**0**
B13	18º	criações comerciais em quase todo o mundo	4	1	1,1	0,0736	**0**
B13	19º	reação de amplificação	4	1	1,1	0,0736	**0**
B13	20º	reação de nested-pcr	4	1	1,1	0,0736	**0**
B14	1º	departamento de santa	17	1	1,1	0,2298	**0**
B14	2º	cruz de la sierra	13	1	1,1	0,1757	**0**
B14	3º	cruz	98	4	0,2	0,1204	**0**
B14	4º	aftosa no departamento de santa	7	1	1,4	0,1204	**4**
B14	5º	área de estudo	7	1	1,1	0,0946	**0**
B14	6º	bolívia	37	1	0,2	0,0909	**0**
B14	7º	municípios do departamento de santa	5	1	1,4	0,0860	**0**
B14	8º	forma de produção	6	1	1,1	0,0811	**6**
B14	9º	aftosa	38	2	0,2	0,0700	**4**
B14	10º	pecuária	7	1	0,8	0,0688	**0**
B14	11º	finalidade de movimentação	5	1	1,1	0,0676	**4**
B14	12º	grau de centralidade	5	1	1,1	0,0676	**4**
B14	13º	produção pecuária	6	1	0,8	0,0590	**4**

Doc.$_{cj}$	Pos.	Sintagma nominal candidato (i)	f_{ijc}	n_{ic}	CNP_i	$Score_{ijc}$	$Relevância_{ijc}$ (autor)
B14	14º	valores em negrito	4	1	1,1	0,0541	**0**
B14	15º	bolívia no ano de 2006	3	1	1,4	0,0516	**0**
B14	16º	bolívia no período de 2004-2006	3	1	1,4	0,0516	**0**
B14	17º	cociente proporcional da pecuária de estabelecimentos pequenos	3	1	1,4	0,0516	**0**
B14	18º	freqüência de propriedades afetadas por a febre	3	1	1,4	0,0516	**0**
B14	19º	engorda	20	1	0,2	0,0491	**3**
B14	20º	ano de 2006	5	3	1,1	0,0408	**0**
B15	1º	dunn dos grupos vacinados no dia	39	1	1,2	0,4696	**2**
B15	2º	dias após a vacinação	34	1	1,1	0,3753	**0**
B15	3º	aglutininas contra a sorovariedade	33	1	1,1	0,3642	**6**
B15	4º	dias após prim	33	1	1,1	0,3642	**0**
B15	5º	período da 420	33	1	1,1	0,3642	**2**
B15	6º	elisa com a amostra	32	1	1,1	0,3532	**4**
B15	7º	comparação de médias	40	2	1,1	0,3311	**3**
B15	8º	igg determinados	33	1	0,8	0,2649	**1**
B15	9º	vacinação média geométrica	33	1	0,8	0,2649	**3**
B15	10º	dia após vacinação média aritmética	22	1	1,1	0,2428	**2**
B15	11º	médio dos titulos	15	1	1,4	0,2107	**3**
B15	12º	240***	55	7	1,1	0,1810	**0**
B15	13º	elisa para hardjo	16	1	1,1	0,1766	**5**
B15	14º	microaglutinação com a amostra	15	1	1,1	0,1656	**2**

Doc.$_{cj}$	Pos.	Sintagma nominal candidato (i)	f_{ijc}	n_{ic}	CNP_i	$Score_{ijc}$	$Relevância_{ijc}$ (autor)
B15	15º	negativo 30**	18	1	0,8	0,1445	**0**
B15	16º	revacinação	71	1	0,2	0,1425	**2**
B15	17º	reforço	70	1	0,2	0,1405	**2**
B15	18º	graf	69	1	0,2	0,1385	**0**
B15	19º	hardjo	64	1	0,2	0,1284	**6**
B15	20º	dias após vacinação média aritmética	11	1	1,1	0,1214	**0**
B16	1º	indução da reação	18	1	1,1	0,4675	**3**
B16	2º	iodeto de propídio	17	1	1,1	0,4415	**3**
B16	3º	seminal eqüino	21	1	0,8	0,3967	**6**
B16	4º	inibidor de serino	12	1	1,1	0,3117	**5**
B16	5º	inibidores de serino	12	1	1,1	0,3117	**5**
B16	6º	cromatografia de exclusão	15	2	1,1	0,2922	**3**
B16	7º	ionóforo de cálcio	10	1	1,1	0,2597	**3**
B16	8º	estrutura da cromatina	9	1	1,1	0,2337	**6**
B16	9º	aa vermelho	12	1	0,8	0,2267	**5**
B16	10º	plasmática íntegra	11	1	0,8	0,2078	**1**
B16	11º	plasmática do espermatozóide	10	2	1,1	0,1948	**1**
B16	12º	inibidor purificado	10	1	0,8	0,1889	**1**
B16	13º	espermatozóides com membrana	7	1	1,1	0,1818	**4**
B16	14º	alto teor	9	1	0,8	0,1700	**5**
B16	15º	dna alto	9	1	0,8	0,1700	**3**
B16	16º	população principal	9	1	0,8	0,1700	**1**

Doc.$_{cj}$	Pos.	Sintagma nominal candidato (i)	f_{ijc}	n_{ic}	CNP_i	$Score_{ijc}$	$Relevância_{ijc}$ (autor)
B16	17º	eqüino	44	2	0,2	0,1558	**5**
B16	18º	fl3	33	1	0,2	0,1558	**3**
B16	19º	avaliação do sêmen	6	1	1,1	0,1558	**6**
B16	20º	azul de tripan	6	1	1,1	0,1558	**3**
C01	1º	calvino	306	4	0,2	0,1024	**5**
C01	2º	obras de borges e calvino	14	1	1,1	0,0561	**6**
C01	3º	grifos do autor	12	1	1,1	0,0481	**0**
C01	4º	noite de inverno	14	2	1,1	0,0409	**1**
C01	5º	memória do mundo	10	1	1,1	0,0400	**1**
C01	6º	borges	218	7	0,2	0,0383	**5**
C01	7º	pensamento complexo	13	1	0,8	0,0379	**6**
C01	8º	biblioteca de babel	9	1	1,1	0,0360	**5**
C01	9º	grifos meus	22	4	0,8	0,0294	**0**
C01	10º	cidades e os símbolos	7	1	1,1	0,0280	**1**
C01	11º	lisa block de behar	7	1	1,1	0,0280	**1**
C01	12º	memória de shakespeare	9	2	1,1	0,0263	**1**
C01	13º	arquivo da literatura	6	1	1,1	0,0240	**6**
C01	14º	coleção de areia	6	1	1,1	0,0240	**1**
C01	15º	coleção de livros	6	1	1,1	0,0240	**5**
C01	16º	rosa dos ventos	4	1	1,4	0,0204	**1**
C01	17º	exercícios de memória	5	1	1,1	0,0200	**4**
C01	18º	literaturas de jorge	5	1	1,1	0,0200	**1**

Doc.$_{cj}$	**Pos.**	**Sintagma nominal candidato (i)**	**f**$_{ijc}$	**n**$_{ic}$	**CNP**$_{i}$	**Score**$_{ijc}$	**Relevância**$_{ijc}$ **(autor)**
C01	19º	objeto de reflexão	5	1	1,1	0,0200	**1**
C01	20º	borges e italo	27	1	0,2	0,0197	**6**
C02	1º	candido	486	6	0,2	0,0672	**6**
C02	2º	educação por a noite	24	1	1,1	0,0605	**2**
C02	3º	observador literário	23	1	0,8	0,0422	**3**
C02	4º	teresina	74	1	0,2	0,0339	**6**
C02	5º	albatroz e o chinês	18	1	0,8	0,0330	**3**
C02	6º	junho de 1993	13	1	1,1	0,0328	**2**
C02	7º	mundos de um humanista	13	1	1,1	0,0328	**2**
C02	8º	textos de intervenção	11	1	1,1	0,0277	**3**
C02	9º	literatura pessoal	14	1	0,8	0,0257	**6**
C02	10º	memorialismo de antonio	9	1	1,1	0,0227	**6**
C02	11º	brigada ligeira	12	1	0,8	0,0220	**3**
C02	12º	memorialismo	45	1	0,2	0,0206	**6**
C02	13º	poços de caldas	8	1	1,1	0,0202	**3**
C02	14º	esquema de machado de assis	6	1	1,4	0,0193	**1**
C02	15º	sala de aula	10	2	1,1	0,0184	**3**
C02	16º	funcionário da monarquia	7	1	1,1	0,0176	**3**
C02	17º	parceiros do rio bonito	7	1	1,1	0,0176	**2**
C02	18º	machado de assis	14	4	1,1	0,0162	**2**
C02	19º	mário de andrade	17	5	1,1	0,0160	**4**
C02	20º	oswald de andrade	17	5	1,1	0,0160	**2**

Doc.$_{cj}$	**Pos.**	**Sintagma nominal candidato (i)**	**f**$_{ijc}$	**n**$_{ic}$	**CNP**$_{i}$	**Score**$_{ijc}$	**Relevância**$_{ijc}$ **(autor)**
C03	1º	edições da narrativa oral no brasil	19	1	1,4	0,1200	**6**
C03	2º	contos populares brasileiros	33	1	0,8	0,1191	**6**
C03	3º	ciência do folk-lore	23	1	1,1	0,1141	**2**
C03	4º	mitos africanos no brasil	22	1	1,1	0,1091	**4**
C03	5º	contos populares do brasil	21	1	1,1	0,1042	**6**
C03	6º	faculdade de letras da ufmg	16	1	1,4	0,1010	**3**
C03	7º	folclore no brasil	20	1	1,1	0,0992	**4**
C03	8º	joão da silva	18	1	1,1	0,0893	**2**
C03	9º	luís da câmara	18	1	1,1	0,0893	**2**
C03	10º	souza carneiro	21	1	0,8	0,0758	**3**
C03	11º	contribuição do folk-lore	14	1	1,1	0,0695	**2**
C03	12º	vale do jequitinhonha	14	1	1,1	0,0695	**3**
C03	13º	vocabulário afro-brasileiro	18	1	0,8	0,0649	**6**
C03	14º	histórias de pai	13	1	1,1	0,0645	**4**
C03	15º	contos tradicionais do brasil	12	1	1,1	0,0595	**6**
C03	16º	lendas e fábulas do brasil	12	1	1,1	0,0595	**6**
C03	17º	acervo do projeto	11	1	1,1	0,0546	**2**
C03	18º	brazileiro para a bibliotheca	11	1	1,1	0,0546	**0**
C03	19º	revista do arquivo municipal	11	1	1,1	0,0546	**3**
C03	20º	narrativas orais no vale do jequitinhonha	8	1	1,4	0,0505	**5**
C04	1º	teatral artur	32	1	0,8	0,0745	**3**
C04	2º	álbuns	106	1	0,2	0,0617	**6**

Doc.$_{cj}$	**Pos.**	**Sintagma nominal candidato (i)**	**f**$_{ijc}$	**n**$_{ic}$	**CNP**$_{i}$	**Score**$_{ijc}$	**Relevância**$_{ijc}$ **(autor)**
C04	3º	amador	112	2	0,2	0,0475	**2**
C04	4º	álbum	81	1	0,2	0,0471	**6**
C04	5º	bispo do rosário	10	1	1,1	0,0320	**4**
C04	6º	manoel de souza	10	1	1,1	0,0320	**1**
C04	7º	são-joanense	55	1	0,2	0,0320	**2**
C04	8º	del-rei	51	1	0,2	0,0297	**0**
C04	9º	leitores dos álbuns	7	1	1,4	0,0285	**6**
C04	10º	montagem dos álbuns	7	1	1,4	0,0285	**6**
C04	11º	páginas dos álbuns	7	1	1,4	0,0285	**5**
C04	12º	apresentações cênicas	11	1	0,8	0,0256	**2**
C04	13º	cartazes cênicos	11	1	0,8	0,0256	**2**
C04	14º	álbuns de antonio	8	1	1,1	0,0256	**0**
C04	15º	guerra	383	10	0,2	0,0228	**0**
C04	16º	biblioteca do clube	7	1	1,1	0,0224	**2**
C04	17º	objetos de antonio	7	1	1,1	0,0224	**2**
C04	18º	pequena história de teatro	7	1	1,1	0,0224	**3**
C04	19º	biblioteca do artur	6	1	1,1	0,0192	**3**
C04	20º	ferreira da rocha	6	1	1,1	0,0192	**0**
C05	1º	poemas concretos	61	1	0,8	0,2589	**6**
C05	2º	haroldo de campos	76	5	1,1	0,1652	**6**
C05	3º	poema concreto	35	1	0,8	0,1485	**6**
C05	4º	pignatari	134	3	0,2	0,0813	**6**

Doc.$_{cj}$	Pos.	Sintagma nominal candidato (i)	f_{ijc}	n_{ic}	CNP_i	$Score_{ijc}$	$Relevância_{ijc}$ (autor)
C05	5º	poesia concreta	23	2	0,8	0,0712	**5**
C05	6º	lacan	117	3	0,2	0,0710	**6**
C05	7º	obra de arte aberta	8	1	1,1	0,0467	**6**
C05	8º	artes visuais	10	1	0,8	0,0424	**4**
C05	9º	cidade dos signos	5	1	1,4	0,0371	**2**
C05	10º	vocal	46	2	0,2	0,0356	**6**
C05	11º	poetas do grupo	6	1	1,1	0,0350	**6**
C05	12º	verbivocovisual	39	2	0,2	0,0302	**6**
C05	13º	não-relação endereçada	7	1	0,8	0,0297	**6**
C05	14º	cansada cornucópia entre festões de rosas murchas	4	1	1,4	0,0297	**1**
C05	15º	décio	60	4	0,2	0,0293	**6**
C05	16º	poetas concretos	9	2	0,8	0,0279	**5**
C05	17º	aguilar	25	1	0,2	0,0265	**2**
C05	18º	noigandres	34	2	0,2	0,0263	**6**
C05	19º	troc	24	1	0,2	0,0255	**4**
C05	20º	anos de 1970	4	1	1,1	0,0233	**4**
C06	1º	buarque	518	4	0,2	0,0982	**0**
C06	2º	sérgio	540	5	0,2	0,0830	**0**
C06	3º	1996a	192	1	0,2	0,0792	**0**
C06	4º	buarque de holanda	84	5	1,1	0,0710	**4**
C06	5º	holanda	279	3	0,2	0,0658	**0**
C06	6º	guilherme de almeida	17	1	1,1	0,0386	**1**

Doc.$_{cj}$	Pos.	Sintagma nominal candidato (i)	f_{ijc}	n_{ic}	CNP_i	$Score_{ijc}$	$Relevância_{ijc}$ (autor)
C06	7º	poesia de manuel	17	1	1,1	0,0386	**0**
C06	8º	originalidade literária	20	1	0,8	0,0330	**6**
C06	9º	tristão de athayde	14	1	1,1	0,0318	**5**
C06	10º	lado oposto e outros	17	1	0,8	0,0281	**0**
C06	11º	raízes do brasil	16	2	1,1	0,0265	**5**
C06	12º	ronald de carvalho	11	1	1,1	0,0250	**1**
C06	13º	revista do brasil	14	2	1,1	0,0232	**0**
C06	14º	ensaio de 1926	10	1	1,1	0,0227	**6**
C06	15º	pensamento de sérgio	10	1	1,1	0,0227	**0**
C06	16º	cigarra	13	1	0,8	0,0215	**0**
C06	17º	instinto de nacionalidade	9	1	1,1	0,0204	**4**
C06	18º	alceu amoroso	12	1	0,8	0,0198	**0**
C06	19º	1996b	44	1	0,2	0,0182	**0**
C06	20º	dezembro de 1948	8	1	1,1	0,0182	**0**
C07	1º	leminski	291	1	0,2	0,2228	**6**
C07	2º	acat	97	1	0,2	0,0743	**0**
C07	3º	poesia marginal	23	1	0,8	0,0704	**2**
C07	4º	poesia concreta	19	2	0,8	0,0425	**2**
C07	5º	emd	54	1	0,2	0,0413	**0**
C07	6º	eac	45	1	0,2	0,0345	**0**
C07	7º	intelectual	28	5	0,8	0,0319	**5**
C07	8º	anseios teóricos	10	1	0,8	0,0306	**1**

Doc.$_{cj}$	**Pos.**	**Sintagma nominal candidato (i)**	**f**$_{ijc}$	**n**$_{ic}$	**CNP**$_{i}$	**Score**$_{ijc}$	**Relevância**$_{ijc}$ **(autor)**
C07	9º	último acesso	10	1	0,8	0,0306	**0**
C07	10º	bonvicino	38	1	0,2	0,0291	**0**
C07	11º	poesia dos anos	5	1	1,4	0,0268	**1**
C07	12º	pcl	29	1	0,2	0,0222	**0**
C07	13º	régis	29	1	0,2	0,0222	**0**
C07	14º	livro de ensaios	7	2	1,1	0,0215	**5**
C07	15º	imprensa alternativa	7	1	0,8	0,0214	**3**
C07	16º	meados dos anos	4	1	1,4	0,0214	**0**
C07	17º	correio de notícias	5	1	1,1	0,0211	**0**
C07	18º	paixão da linguagem	5	1	1,1	0,0211	**1**
C07	19º	panorama de um pensamento	5	1	1,1	0,0211	**5**
C07	20º	campo literário	9	2	0,8	0,0201	**3**
C08	1º	hilda	240	1	0,2	0,2228	**6**
C08	2º	hilst	176	1	0,2	0,1634	**6**
C08	3º	casa do sol	26	1	1,1	0,1327	**5**
C08	4º	cadernos de literatura brasileira	23	1	1,1	0,1174	**5**
C08	5º	prazer do texto	25	2	1,1	0,0931	**5**
C08	6º	rumor da língua	15	1	1,1	0,0766	**4**
C08	7º	odes mínimas	18	1	0,8	0,0668	**5**
C08	8º	barthes por roland	11	1	1,1	0,0562	**4**
C08	9º	olhos de cão	11	1	1,1	0,0562	**5**
C08	10º	testamento para greco	10	1	1,1	0,0511	**5**

Doc.$_{cj}$	Pos.	Sintagma nominal candidato (i)	f_{ijc}	n_{ic}	CNP_i	$Score_{ijc}$	$Relevância_{ijc}$ (autor)
C08	11º	vermelho da vida	10	1	1,1	0,0511	**6**
C08	12º	e-mail de josé	9	1	1,1	0,0460	**4**
C08	13º	nota do organizador	8	1	1,1	0,0408	**0**
C08	14º	obra de hilda	8	1	1,1	0,0408	**6**
C08	15º	poesia de hilda	8	1	1,1	0,0408	**6**
C08	16º	biografema	44	1	0,2	0,0408	**6**
C08	17º	desafio biográfico	10	1	0,8	0,0371	**1**
C08	18º	gênero biográfico	10	1	0,8	0,0371	**3**
C08	19º	história do olho	7	1	1,1	0,0357	**1**
C08	20º	vida escrita	9	1	0,8	0,0334	**5**
C09	1º	manoel de barros	28	1	1,1	0,1796	**5**
C09	2º	riachinho sirimim	14	1	0,8	0,0653	**5**
C09	3º	infância da escrita	9	1	1,1	0,0577	**6**
C09	4º	língua maior	16	2	0,8	0,0545	**3**
C09	5º	sirimim	42	1	0,2	0,0490	**5**
C09	6º	idem	81	4	0,2	0,0434	**0**
C09	7º	ideia de infância	6	1	1,1	0,0385	**6**
C09	8º	margens da alegria	6	1	1,1	0,0385	**5**
C09	9º	literatura menor	14	3	0,8	0,0373	**5**
C09	10º	própria língua	16	4	0,8	0,0343	**3**
C09	11º	ariès	28	1	0,2	0,0327	**2**
C09	12º	língua menor	9	2	0,8	0,0306	**4**

Doc.$_{cj}$	Pos.	Sintagma nominal candidato (i)	f_{ijc}	n_{ic}	CNP_i	$Score_{ijc}$	$Relevância_{ijc}$ (autor)
C09	13º	queirós	26	1	0,2	0,0303	**2**
C09	14º	escrita de rosa	6	2	1,1	0,0281	**5**
C09	15º	arte literária	6	1	0,8	0,0280	**5**
C09	16º	escrita rosiana	8	2	0,8	0,0272	**6**
C09	17º	conceito de devir	4	1	1,1	0,0257	**4**
C09	18º	conceito de infância	4	1	1,1	0,0257	**6**
C09	19º	manoel de barros e bartolomeu	4	1	1,1	0,0257	**5**
C09	20º	zona de vizinhança	4	1	1,1	0,0257	**4**
C10	1º	conhecimento do inferno	34	1	1,1	0,1138	**6**
C10	2º	memória de elefante	34	1	1,1	0,1138	**6**
C10	3º	antunes	366	4	0,2	0,1024	**6**
C10	4º	antuniana	164	1	0,2	0,0998	**6**
C10	5º	morte de carlos	27	1	1,1	0,0904	**0**
C10	6º	cus de judas	25	1	1,1	0,0837	**6**
C10	7º	paixões da alma	24	1	1,1	0,0804	**0**
C10	8º	ordem natural das coisas	29	2	1,1	0,0709	**6**
C10	9º	sombra no mar	21	1	1,1	0,0703	**0**
C10	10º	dicionário da obra de antónio	12	1	1,4	0,0511	**3**
C10	11º	romances de antónio	12	1	1,1	0,0402	**6**
C10	12º	textualidade	73	2	0,2	0,0324	**6**
C10	13º	cartas da guerra	9	1	1,1	0,0301	**6**
C10	14º	antuniano	46	1	0,2	0,0280	**6**

Doc$_{.cj}$	**Pos.**	**Sintagma nominal candidato (i)**	**f**$_{ijc}$	**n**$_{ic}$	**CNP**$_{i}$	**Score**$_{ijc}$	**Relevância**$_{ijc}$ **(autor)**
C10	15º	presente da enunciação	11	2	1,1	0,0269	**1**
C10	16º	arquipélago da insónia	8	1	1,1	0,0268	**6**
C10	17º	bico da areia	8	1	1,1	0,0268	**0**
C10	18º	lobo	95	4	0,2	0,0266	**6**
C10	19º	fragilidade dos laços humanos	6	1	1,4	0,0256	**6**
C10	20º	sistema dos objetos	6	1	1,4	0,0256	**1**
C11	1º	maxakali	248	1	0,2	0,2228	**5**
C11	2º	yãmîy	153	1	0,2	0,1374	**6**
C11	3º	tikmû'ûn	91	1	0,2	0,0817	**1**
C11	4º	maxakalis	83	1	0,2	0,0746	**5**
C11	5º	rituais	13	1	0,8	0,0467	**4**
C11	6º	koxuk	42	1	0,2	0,0377	**6**
C11	7º	yãmîyxop	41	1	0,2	0,0368	**6**
C11	8º	ritual	17	3	0,8	0,0349	**4**
C11	9º	casa de religião	7	1	1,1	0,0346	**3**
C11	10º	idem	72	4	0,2	0,0297	**0**
C11	11º	livro de cantos rituais	6	1	1,1	0,0296	**3**
C11	12º	inmõxã	32	1	0,2	0,0287	**2**
C11	13º	escrita alfabética	8	1	0,8	0,0287	**3**
C11	14º	yõg	29	1	0,2	0,0261	**0**
C11	15º	comida	7	1	0,8	0,0252	**1**
C11	16º	casa dos cantos	4	1	1,4	0,0252	**3**

Doc.$_{cj}$	**Pos.**	**Sintagma nominal candidato (i)**	**f**$_{ijc}$	**n**$_{ic}$	**CNP**$_{i}$	**Score**$_{ijc}$	**Relevância**$_{ijc}$ **(autor)**
C11	17º	pau de religião	5	1	1,1	0,0247	**4**
C11	18º	terreiro de religião	5	1	1,1	0,0247	**3**
C11	19º	tihik	25	1	0,2	0,0225	**1**
C11	20º	hãm	24	1	0,2	0,0216	**1**
C12	1º	perec	328	3	0,2	0,1274	**6**
C12	2º	vida modo	43	1	0,8	0,1168	**3**
C12	3º	oulipo	108	2	0,2	0,0535	**6**
C12	4º	biblioteca de babelr	14	1	1,1	0,0523	**3**
C12	5º	borges	302	7	0,2	0,0495	**6**
C12	6º	matemáticos	59	1	0,2	0,0401	**5**
C12	7º	dans	56	1	0,2	0,0380	**0**
C12	8º	milliards de poèmes	10	1	1,1	0,0374	**3**
C12	9º	la disparition	13	1	0,8	0,0353	**3**
C12	10º	números naturais	13	1	0,8	0,0353	**3**
C12	11º	autor do quixoter	8	1	1,1	0,0299	**3**
C12	12º	obra de perec	8	1	1,1	0,0299	**5**
C12	13º	contrainte	42	1	0,2	0,0285	**6**
C12	14º	queneau	53	2	0,2	0,0263	**5**
C12	15º	qui	82	4	0,2	0,0256	**0**
C12	16º	contraintes	35	1	0,2	0,0238	**6**
C12	17º	est	60	3	0,2	0,0233	**0**
C12	18º	roubaud	34	1	0,2	0,0231	**4**

Doc.$_{cj}$	Pos.	Sintagma nominal candidato (i)	f_{ijc}	n_{ic}	CNP_i	$Score_{ijc}$	$Relevância_{ijc}$ (autor)
C12	19º	analìtico de john	6	1	1,1	0,0224	**0**
C12	20º	jardim de veredas que se bifurcamr	6	1	1,1	0,0224	**3**
C13	1º	riobaldo	92	3	0,2	0,1001	**4**
C13	2º	ricoeur	99	4	0,2	0,0866	**5**
C13	3º	aporias do tempo	7	1	1,1	0,0733	**6**
C13	4º	grande sertão	25	6	0,8	0,0574	**5**
C13	5º	obra de luandino	5	1	1,1	0,0524	**4**
C13	6º	makulusu	25	1	0,2	0,0476	**4**
C13	7º	diabo na rua	4	1	1,1	0,0419	**4**
C13	8º	mais-velho	5	1	0,8	0,0381	**2**
C13	9º	ponto de fuga	3	1	1,1	0,0314	**0**
C13	10º	diadorim	15	1	0,2	0,0286	**4**
C13	11º	mero tropo de ornamentação do discurso	2	1	1,4	0,0267	**1**
C13	12º	vendedor de vinho de palma	2	1	1,4	0,0267	**0**
C13	13º	laban	13	1	0,2	0,0248	**2**
C13	14º	luandino	17	2	0,2	0,0236	**5**
C13	15º	futuro	30	10	0,8	0,0234	**4**
C13	16º	vieira	40	6	0,2	0,0230	**3**
C13	17º	narrativa de ficção	3	2	1,1	0,0229	**4**
C13	18º	teoria do caos	3	2	1,1	0,0229	**2**
C13	19º	beardsley	12	1	0,2	0,0229	**2**
C13	20º	mero ornamento	3	1	0,8	0,0229	**1**

Doc.$_{cj}$	Pos.	Sintagma nominal candidato (i)	f_{ijc}	n_{ic}	CNP_i	$Score_{ijc}$	$Relevância_{ijc}$ (autor)
D01	1º	influência na estrutura e propriedades superficiais	83	1	1,1	0,9295	**6**
D01	2º	influência dos tratamentos de superfície na resistência	82	1	0,8	0,6679	**5**
D01	3º	série no	67	1	0,8	0,5457	**0**
D01	4º	volume desgastado total	54	1	0,8	0,4398	**5**
D01	5º	influência dos tratamentos de superfície na resistência à corrosão	35	1	0,8	0,2851	**6**
D01	6º	microabrasivo	102	1	0,2	0,2077	**5**
D01	7º	volume desgastado	22	1	0,8	0,1792	**5**
D01	8º	volume desgastado no recobrimento	16	1	1,1	0,1792	**5**
D01	9º	volume desgastado no substrato	16	1	1,1	0,1792	**5**
D01	10º	substrato de aço	21	2	1,1	0,1696	**5**
D01	11º	diâmetro externo da calota	15	1	1,1	0,1680	**3**
D01	12º	diâmetro interno da calota	15	1	1,1	0,1680	**3**
D01	13º	ubc	74	1	0,2	0,1507	**6**
D01	14º	análise de regressão	12	1	1,1	0,1344	**3**
D01	15º	diâmetro da calota	12	1	1,1	0,1344	**3**
D01	16º	calota	64	1	0,2	0,1303	**3**
D01	17º	base na perfilometria de contato	9	1	1,4	0,1283	**3**
D01	18º	rugosidade da superfície desgastada após 1350rev	9	1	1,4	0,1283	**3**
D01	19º	rugosidade na seção transversal central da calota	9	1	1,4	0,1283	**3**
D01	20º	topografia invertida da superfície desgastada após 1350rev	9	1	1,4	0,1283	**3**

Doc.$_{cj}$	Pos.	Sintagma nominal candidato (i)	f_{ijc}	n_{ic}	CNP_i	$Score_{ijc}$	$Relevância_{ijc}$ (autor)
D02	1º	concreto	14	1	0,8	0,3555	**2**
D02	2º	rochas ornamentais	13	1	0,8	0,3301	**4**
D02	3º	tomé das letras	9	1	1,1	0,3142	**1**
D02	4º	quartzito de grau de sanidade	7	1	1,4	0,3111	**4**
D02	5º	quartzito como agregado	12	1	0,8	0,3047	**6**
D02	6º	retida acum	10	1	0,8	0,2539	**0**
D02	7º	comprimento de quadro	7	1	1,1	0,2444	**0**
D02	8º	quartzito	34	1	0,2	0,2158	**5**
D02	9º	extração de quartzito	6	1	1,1	0,2095	**6**
D02	10º	los angeles	8	1	0,8	0,2031	**4**
D02	11º	cimento	31	1	0,2	0,1968	**0**
D02	12º	seca	10	2	0,8	0,1831	**0**
D02	13º	médio com comprimento de quadro igual	4	1	1,4	0,1777	**0**
D02	14º	extração de areia	5	1	1,1	0,1746	**1**
D02	15º	abrasão los	6	1	0,8	0,1524	**0**
D02	16º	la serna	6	1	0,8	0,1524	**0**
D02	17º	produtos cerâmicos	6	1	0,8	0,1524	**1**
D02	18º	brita	23	1	0,2	0,1460	**3**
D02	19º	maioria dos valores	3	1	1,4	0,1333	**0**
D02	20º	quartzito branco de grau de sanidade	3	1	1,4	0,1333	**4**
D03	1º	tempo de austêmpera	16	1	1,1	0,3453	**2**
D03	2º	estrutura do aço	15	1	1,1	0,3238	**1**

Doc.$_{cj}$	Pos.	Sintagma nominal candidato (i)	f_{ijc}	n_{ic}	CNP_i	$Score_{ijc}$	$Relevância_{ijc}$ (autor)
D03	3º	temperatura de austêmpera	13	1	1,1	0,2806	**2**
D03	4º	blocos de austenita	8	1	1,1	0,1727	**3**
D03	5º	austemperado	43	1	0,2	0,1687	**3**
D03	6º	corpos de prova	22	5	1,1	0,1673	**0**
D03	7º	bhadeshia	37	1	0,2	0,1452	**1**
D03	8º	mateo	34	1	0,2	0,1334	**0**
D03	9º	placas de ferrita	6	1	1,1	0,1295	**3**
D03	10º	200°c	32	1	0,2	0,1256	**0**
D03	11º	300°c	32	1	0,2	0,1256	**0**
D03	12º	espessura das placas	5	1	1,1	0,1079	**1**
D03	13º	placas de bainita	5	1	1,1	0,1079	**3**
D03	14º	produção do aço	5	1	1,1	0,1079	**5**
D03	15º	quantidade de austenita	5	1	1,1	0,1079	**1**
D03	16º	austêmpera	25	1	0,2	0,0981	**5**
D03	17º	resistência à fadiga	8	3	1,1	0,0963	**5**
D03	18º	bainítica	22	1	0,2	0,0863	**4**
D03	19º	5ºc até 750ºc	4	1	1,1	0,0863	**0**
D03	20º	ciclos térmicos de austêmpera	4	1	1,1	0,0863	**4**
D04	1º	#REF!	1	0	#N/D	#DIV/0!	**0**
D04	2º	cecília	178	1	0,2	0,2158	**0**
D04	3º	janaina	160	1	0,2	0,1940	**0**
D04	4º	villanova	153	1	0,2	0,1855	**6**

Doc.$_{cj}$	Pos.	Sintagma nominal candidato (i)	f_{ijc}	n_{ic}	CNP_i	$Score_{ijc}$	$Relevância_{ijc}$ (autor)
D04	5º	cloridrato de propranolol	15	1	1,1	0,1000	**3**
D04	6º	oliveira	178	4	0,2	0,0954	**0**
D04	7º	sulfato de sódio	11	1	1,1	0,0734	**0**
D04	8º	polimerização em emulsão	10	1	1,1	0,0667	**6**
D04	9º	tamanho das partículas	13	2	1,1	0,0625	**5**
D04	10º	compressão direta	12	1	0,8	0,0582	**6**
D04	11º	nanofibras de celulose	8	1	1,1	0,0534	**6**
D04	12º	ácido acrílico	10	1	0,8	0,0485	**3**
D04	13º	água purificada	10	1	0,8	0,0485	**0**
D04	14º	acrilato de etila	7	1	1,1	0,0467	**3**
D04	15º	gentilmente doado por a pharma	7	1	1,1	0,0467	**0**
D04	16º	metacrilato de butila	7	1	1,1	0,0467	**3**
D04	17º	metacrilato de glicidila	7	1	1,1	0,0467	**3**
D04	18º	metacrilato de metila	7	1	1,1	0,0467	**3**
D04	19º	rampa de aquecimento de 10º c por minuto	6	1	1,2	0,0437	**0**
D04	20º	distribuição do tamanho das partículas	5	1	1,4	0,0424	**6**
D05	1º	índice de oxidação	19	1	1,1	0,3366	**6**
D05	2º	altura da banda	12	1	1,1	0,2126	**5**
D05	3º	peuapm	58	1	0,2	0,1868	**6**
D05	4º	nùmero de onda	10	1	1,1	0,1772	**3**
D05	5º	imagem de mev do peuapm oxidado	7	1	1,4	0,1579	**3**
D05	6º	espectro de ftir	8	1	1,1	0,1417	**6**

Doc$_{.cj}$	**Pos.**	**Sintagma nominal candidato (i)**	**f**$_{ijc}$	**n**$_{ic}$	**CNP**$_{i}$	**Score**$_{ijc}$	**Relevância**$_{ijc}$ **(autor)**
D05	7º	tempo de exposição	10	2	1,1	0,1278	**5**
D05	8º	fluxograma do caminho	7	1	1,1	0,1240	**4**
D05	9º	peróxido de hidrogênio	11	3	1,1	0,1087	**2**
D05	10º	oxidação do peuapm	6	1	1,1	0,1063	**6**
D05	11º	oxidativa do peuapm	6	1	1,1	0,1063	**6**
D05	12º	peróxido de benzoíla	9	3	1,1	0,0890	**0**
D05	13º	próteses de joelho	5	1	1,1	0,0886	**5**
D05	14º	resposta de macrófagos	5	1	1,1	0,0886	**3**
D05	15º	lasmat	22	1	0,2	0,0709	**2**
D05	16º	gráfico proporção do grupamento	4	1	1,1	0,0709	**6**
D05	17º	inflamatória de macrófagos	4	1	1,1	0,0709	**5**
D05	18º	cristalinidade	21	1	0,2	0,0677	**2**
D05	19º	prótese	21	1	0,2	0,0677	**2**
D05	20º	espectro de ftir do peuapm	3	1	1,4	0,0677	**6**
D06	1º	fenômeno de delayed	63	1	1,1	0,3895	**6**
D06	2º	reembutidos dos aços	25	1	1,4	0,1967	**0**
D06	3º	inoxidáveis	192	2	0,2	0,1556	**6**
D06	4º	fração volumétrica de martensita	24	1	1,1	0,1484	**6**
D06	5º	304a	117	1	0,2	0,1315	**6**
D06	6º	embutimento do delayed	21	1	1,1	0,1298	**0**
D06	7º	cracking	112	1	0,2	0,1259	**6**
D06	8º	304h	110	1	0,2	0,1237	**6**

Doc.$_{cj}$	Pos.	Sintagma nominal candidato (i)	f_{ijc}	n_{ic}	CNP_i	$Score_{ijc}$	$Relevância_{ijc}$ (autor)
D06	9º	razão de embutimento	18	1	1,1	0,1113	**6**
D06	10º	paredes dos copos	13	1	1,4	0,1023	**6**
D06	11º	304n	85	1	0,2	0,0956	**6**
D06	12º	304b	78	1	0,2	0,0877	**6**
D06	13º	razões de embutimento	14	1	1,1	0,0866	**6**
D06	14º	reembutido do aço	14	1	1,1	0,0866	**0**
D06	15º	austeníticos	118	3	0,2	0,0740	**6**
D06	16º	fração volumétrica de martensita induzida por deformação	9	1	1,4	0,0708	**6**
D06	17º	aços	165	5	0,2	0,0653	**6**
D06	18º	classe de aços	10	1	1,1	0,0618	**6**
D06	19º	quantidade de martensita	10	1	1,1	0,0618	**6**
D06	20º	embutimento	50	1	0,2	0,0562	**6**
D07	1º	superfícies modificadas	64	1	0,8	0,5756	**6**
D07	2º	sistemas recobertos	33	2	0,8	0,2140	**6**
D07	3º	tempo de incubação	15	1	1,1	0,1855	**4**
D07	4º	tempo de nitretação	13	2	1,1	0,1159	**6**
D07	5º	linear para o sistema	9	1	1,1	0,1113	**0**
D07	6º	n0f0	48	1	0,2	0,1079	**1**
D07	7º	materiais da ee-ufmg	8	1	1,1	0,0989	**3**
D07	8º	processo de nitretação	11	2	1,1	0,0981	**5**
D07	9º	duplex	60	2	0,2	0,0973	**5**

Doc$_{.cj}$	**Pos.**	**Sintagma nominal candidato (i)**	**f**$_{ijc}$	**n**$_{ic}$	**CNP**$_{i}$	**Score**$_{ijc}$	**Relevância**$_{ijc}$ **(autor)**
D07	10º	n0f1	42	1	0,2	0,0944	**1**
D07	11º	n2f2	41	1	0,2	0,0922	**1**
D07	12º	n2f1	40	1	0,2	0,0899	**1**
D07	13º	dados experimentais	10	1	0,8	0,0899	**1**
D07	14º	n0f2	39	1	0,2	0,0877	**1**
D07	15º	n4f0	38	1	0,2	0,0854	**1**
D07	16º	abnt	82	4	0,2	0,0815	**0**
D07	17º	perfis de rugosidade tridimensional para os sistemas	6	1	1,2	0,0809	**5**
D07	18º	n2f0	33	1	0,2	0,0742	**1**
D07	19º	nitretados	45	2	0,2	0,0730	**4**
D07	20º	profundidade de penetração	8	2	1,1	0,0713	**5**
D08	1º	pva	141	1	0,2	0,2158	**5**
D08	2º	qui	105	1	0,2	0,1607	**6**
D08	3º	genipin	71	1	0,2	0,1087	**6**
D08	4º	adesão de células	8	1	1,1	0,0674	**5**
D08	5º	grau de intumescimento	8	1	1,1	0,0674	**4**
D08	6º	grau de desacetilação	6	1	1,1	0,0505	**4**
D08	7º	quitosana	43	2	0,2	0,0475	**6**
D08	8º	quitosana pura	7	1	0,8	0,0429	**6**
D08	9º	ligações de hidrogênio	7	2	1,1	0,0425	**5**
D08	10º	ampliação de 1500x	5	1	1,1	0,0421	**3**
D08	11º	intensidade da absorbância	5	1	1,1	0,0421	**4**

Doc.$_{cj}$	Pos.	Sintagma nominal candidato (i)	f_{ijc}	n_{ic}	CNP_i	$Score_{ijc}$	$Relevância_{ijc}$ (autor)
D08	12º	reticulante	24	1	0,2	0,0367	**6**
D08	13º	ângulo de contato	6	2	1,1	0,0364	**2**
D08	14º	ampliação de 500x	4	1	1,1	0,0337	**2**
D08	15º	concentração de quitosana	4	1	1,1	0,0337	**4**
D08	16º	grau de expansão	4	1	1,1	0,0337	**4**
D08	17º	registro por microscopia óptica de mtt	3	1	1,4	0,0321	**5**
D08	18º	terminações nervosas	5	1	0,8	0,0306	**3**
D08	19º	harris	17	1	0,2	0,0260	**3**
D08	20º	acetato de vinila	3	1	1,1	0,0253	**4**
D09	1º	instrumentos	247	1	0,2	0,2158	**3**
D09	2º	protaper do grupo	44	1	1,1	0,2115	**0**
D09	3º	diferença mínima significativa	27	1	0,8	0,0944	**0**
D09	4º	instrumento	105	1	0,2	0,0918	**3**
D09	5º	endodônticos de niti	17	1	1,1	0,0817	**4**
D09	6º	protaper	89	1	0,2	0,0778	**4**
D09	7º	padrão dos instrumentos	12	1	1,4	0,0734	**4**
D09	8º	interior do canal	13	1	1,1	0,0625	**1**
D09	9º	instrumentos de niti acionados	12	1	1,1	0,0577	**4**
D09	10º	análise estatística dos valores	9	1	1,4	0,0551	**0**
D09	11º	formatação dos canais	9	1	1,4	0,0551	**2**
D09	12º	protaper dos grupos	9	1	1,4	0,0551	**0**
D09	13º	instrumentos do grupo	11	1	1,1	0,0529	**0**

Doc.$_{cj}$	Pos.	Sintagma nominal candidato (i)	f_{ijc}	n_{ic}	CNP_i	$Score_{ijc}$	$Relevância_{ijc}$ (autor)
D09	14º	número de ciclos	11	1	1,1	0,0529	**6**
D09	15º	torque máximo	14	1	0,8	0,0489	**6**
D09	16º	instrumentos de niti	10	1	1,1	0,0481	**6**
D09	17º	diâmetro do instrumento	9	1	1,1	0,0433	**4**
D09	18º	instrumentos de finalização	9	1	1,1	0,0433	**4**
D09	19º	instrumentos de formatação	9	1	1,1	0,0433	**4**
D09	20º	linear entre torque máximo	9	1	1,1	0,0433	**0**
D10	1º	área por molécula	16	1	1,1	0,2435	**1**
D10	2º	banda soret	19	1	0,8	0,2103	**2**
D10	3º	compressão da barreira em uma velocidade	8	1	1,4	0,1550	**1**
D10	4º	minutos para a evaporação dos solventes	8	1	1,2	0,1328	**0**
D10	5º	fig	78	3	0,2	0,1204	**0**
D10	6º	temperatura da subfase	7	1	1,1	0,1065	**0**
D10	7º	comprimento de onda	19	5	1,1	0,1019	**2**
D10	8º	tpp	49	2	0,2	0,0978	**0**
D10	9º	banda	19	4	0,8	0,0930	**1**
D10	10º	pressão superficial	7	1	0,8	0,0775	**2**
D10	11º	estudo da variação da concentração	4	1	1,4	0,0775	**1**
D10	12º	variação da quantidade de moléculas	4	1	1,4	0,0775	**0**
D10	13º	estabilização da frequência	5	1	1,1	0,0761	**0**
D10	14º	massa de no2	5	1	1,1	0,0761	**1**
D10	15º	solução de n hexano	5	1	1,1	0,0761	**0**

Doc.$_{cj}$	Pos.	Sintagma nominal candidato (i)	f_{ijc}	n_{ic}	CNP_i	$Score_{ijc}$	$Relevância_{ijc}$ (autor)
D10	16º	tubo de ensaio	5	1	1,1	0,0761	**0**
D10	17º	mmol	26	1	0,2	0,0719	**0**
D10	18º	molécula	70	5	0,2	0,0682	**0**
D10	19º	bandas	11	3	0,8	0,0679	**1**
D10	20º	no2	24	1	0,2	0,0664	**6**
D11	1º	jusante do córrego	11	1	1,1	0,2418	**0**
D11	2º	francisco	54	1	0,2	0,2158	**0**
D11	3º	psf1	50	1	0,2	0,1998	**0**
D11	4º	efeito adverso à biota	9	1	1,1	0,1978	**2**
D11	5º	volatilizáveis por acidificação	9	1	1,1	0,1978	**0**
D11	6º	psf4	46	1	0,2	0,1839	**0**
D11	7º	ponto de referência	8	1	1,1	0,1759	**0**
D11	8º	duplicata	43	1	0,2	0,1719	**0**
D11	9º	jusante do lançamento de efluentes	6	1	1,4	0,1679	**0**
D11	10º	toxicidade dos sedimentos	6	1	1,4	0,1679	**6**
D11	11º	psf6	39	1	0,2	0,1559	**0**
D11	12º	amostra de referência	6	1	1,1	0,1319	**0**
D11	13º	ensaios de ecotoxicidade	6	1	1,1	0,1319	**3**
D11	14º	rio	32	1	0,2	0,1279	**1**
D11	15º	efeito adverso improvável	8	1	0,8	0,1279	**0**
D11	16º	efeito adverso provável	8	1	0,8	0,1279	**0**
D11	17º	intersticial bruta	8	1	0,8	0,1279	**0**

Doc$_{.cj}$	**Pos.**	**Sintagma nominal candidato (i)**	**f**$_{ijc}$	**n**$_{ic}$	**CNP**$_{i}$	**Score**$_{ijc}$	**Relevância**$_{ijc}$ **(autor)**
D11	18º	massa bruta	8	1	0,8	0,1279	**0**
D11	19º	retiro velho	8	1	0,8	0,1279	**0**
D11	20º	conama	28	1	0,2	0,1119	**0**
D12	1º	rio grande do sul	23	1	1,1	0,3175	**2**
D12	2º	ametista	86	1	0,2	0,2158	**6**
D12	3º	centros de cor da ametista	11	1	1,4	0,1932	**6**
D12	4º	felício dos santos	10	1	1,4	0,1757	**2**
D12	5º	número de onda	34	5	1,1	0,1653	**2**
D12	6º	formação de centros	11	1	1,1	0,1518	**5**
D12	7º	posições dos picos de absorção sugeridos	10	1	1,2	0,1506	**2**
D12	8º	ametista natural	14	1	0,8	0,1405	**6**
D12	9º	teores das impurezas	10	1	1,1	0,1380	**2**
D12	10º	prasiolita	52	1	0,2	0,1305	**5**
D12	11º	espectro de absorção	13	2	1,1	0,1294	**3**
D12	12º	comprimento de onda	26	5	1,1	0,1264	**2**
D12	13º	exposição à radiação	9	1	1,1	0,1242	**4**
D12	14º	minutos de exposição à radiação	7	1	1,4	0,1230	**2**
D12	15º	amostra de ametista	8	1	1,1	0,1104	**6**
D12	16º	brejinho das ametistas	8	1	1,1	0,1104	**3**
D12	17º	irradiada	44	1	0,2	0,1104	**5**
D12	18º	ametista sintética	11	1	0,8	0,1104	**3**
D12	19º	coloração amarela	11	1	0,8	0,1104	**2**

Doc.$_{cj}$	Pos.	Sintagma nominal candidato (i)	f_{ijc}	n_{ic}	CNP_i	$Score_{ijc}$	$Relevância_{ijc}$ (autor)
D12	20º	paramagnética eletrônica	11	1	0,8	0,1104	**2**
E01	1º	nanotubos de titanato	63	1	1,1	0,5331	**6**
E01	2º	síntese e purificação dos nanotubos de carbono	40	1	1,2	0,3692	**6**
E01	3º	decoração de nanotubos de carbono	33	1	1,4	0,3554	**6**
E01	4º	reações de oxidação com 2o2 promovidas por ntc e tints	29	1	1,2	0,2677	**6**
E01	5º	remoção de compostos sulfurados do petróleo	23	1	1,4	0,2477	**6**
E01	6º	método do sal	28	1	1,1	0,2369	**0**
E01	7º	comportamento térmico dos nanotubos de carbono	24	1	1,2	0,2215	**6**
E01	8º	nanotubos de carbono	22	1	1,1	0,1862	**6**
E01	9º	estudo do comportamento térmico dos nanotubos de carbono	24	1	0,8	0,1477	**6**
E01	10º	minutos de reação	13	1	1,1	0,1100	**0**
E01	11º	decorados	59	1	0,2	0,0908	**0**
E01	12º	nanotubos de carbono de paredes múltiplas	8	1	1,4	0,0862	**6**
E01	13º	nanotubos de carbono de paredes simples	7	1	1,4	0,0754	**6**
E01	14º	fig	61	2	0,2	0,0656	**0**
E01	15º	nanotubos de titanato de hidrogênio	6	1	1,4	0,0646	**6**
E01	16º	decorados com ouro	7	1	1,1	0,0592	**0**
E01	17º	método do polieletrólito	7	1	1,1	0,0592	**4**
E01	18º	ordem com relação	7	1	1,1	0,0592	**0**
E01	19º	ntcpm	38	1	0,2	0,0585	**4**
E01	20º	polieletrólito	37	1	0,2	0,0569	**2**

Doc.$_{cj}$	Pos.	Sintagma nominal candidato (i)	f_{ijc}	n_{ic}	CNP_i	$Score_{ijc}$	$Relevância_{ijc}$ (autor)
E02	1º	diesel	210	1	0,2	0,2000	**6**
E02	2º	previsão de propriedades físico-químicas do óleo	19	1	1,4	0,1267	**6**
E02	3º	índice de cetano	24	1	1,1	0,1257	**6**
E02	4º	previsão do teor de biodiesel no óleo	21	1	1,2	0,1200	**5**
E02	5º	determinação da origem e tipo do óleo	18	1	1,4	0,1200	**6**
E02	6º	ponto de fulgor	22	1	1,1	0,1152	**6**
E02	7º	previsão da massa específica	22	1	1,1	0,1152	**6**
E02	8º	método proposto	30	1	0,8	0,1143	**1**
E02	9º	cinemática do óleo	20	1	1,1	0,1048	**2**
E02	10º	diesel relacionadas à flamabilidade	18	1	1,1	0,0943	**4**
E02	11º	teor de biodiesel	18	1	1,1	0,0943	**6**
E02	12º	massa específica	19	1	0,8	0,0724	**6**
E02	13º	variável latente	18	1	0,8	0,0686	**5**
E02	14º	valores de rmsep	13	1	1,1	0,0681	**5**
E02	15º	volume recuperado	17	1	0,8	0,0648	**3**
E02	16º	variância explicada	14	1	0,8	0,0533	**5**
E02	17º	conjunto de validação	10	1	1,1	0,0524	**6**
E02	18º	número de cetano	10	1	1,1	0,0524	**3**
E02	19º	precisão do método proposto	10	1	1,1	0,0524	**3**
E02	20º	tipo do óleo	10	1	1,1	0,0524	**3**
E03	1º	#REF!	3	0	#N/D	#DIV/0!	**0**
E03	2º	caracterização dos compostos sintetizados	61	1	1,4	0,5770	**6**

Doc.$_{cj}$	Pos.	Sintagma nominal candidato (i)	f_{ijc}	n_{ic}	CNP_i	$Score_{ijc}$	$Relevância_{ijc}$ (autor)
E03	3º	acetato de etila	37	1	1,1	0,2750	**2**
E03	4º	espectro de rmn	66	3	1,1	0,2565	**6**
E03	5º	espectro na região	22	1	1,1	0,1635	**4**
E03	6º	dados de rmn	18	1	1,1	0,1338	**5**
E03	7º	anexo	130	2	0,2	0,1228	**4**
E03	8º	alifático	82	1	0,2	0,1108	**6**
E03	9º	aspecto físico	19	1	0,8	0,1027	**6**
E03	10º	evolução da reação	13	1	1,1	0,0966	**6**
E03	11º	síntese do malonato	11	1	1,1	0,0818	**2**
E03	12º	término da reação	11	1	1,1	0,0818	**3**
E03	13º	cdcl3	86	2	0,2	0,0812	**5**
E03	14º	mhz	142	4	0,2	0,0764	**0**
E03	15º	seção expandida do subespectro	10	1	1,1	0,0743	**1**
E03	16º	éster	54	1	0,2	0,0730	**1**
E03	17º	derivado fulerênico	13	1	0,8	0,0703	**4**
E03	18º	síntese do derivado fulerênico	9	1	1,1	0,0669	**2**
E03	19º	fase orgânica	23	3	0,8	0,0650	**6**
E03	20º	pressão reduzida	17	2	0,8	0,0642	**2**
E04	1º	teor de nióbio	12	1	1,1	0,3070	**6**
E04	2º	estrutura da goethita	8	1	1,1	0,2047	**3**
E04	3º	nióbio	43	1	0,2	0,2000	**6**
E04	4º	área específica	9	1	0,8	0,1674	**3**

Doc.$_{cj}$	Pos.	Sintagma nominal candidato (i)	f_{ijc}	n_{ic}	CNP_i	$Score_{ijc}$	$Relevância_{ijc}$ (autor)
E04	5º	gt-nb11	29	1	0,2	0,1349	**3**
E04	6º	goethita	28	1	0,2	0,1302	**3**
E04	7º	gt-nb4	26	1	0,2	0,1209	**3**
E04	8º	gtpura	25	1	0,2	0,1163	**3**
E04	9º	gt-nb1	23	1	0,2	0,1070	**3**
E04	10º	processos avançados de oxidação	4	1	1,1	0,1023	**6**
E04	11º	goethita pura	5	1	0,8	0,0930	**3**
E04	12º	fenton heterogêneo	7	2	0,8	0,0910	**6**
E04	13º	min de reação	5	2	1,1	0,0894	**0**
E04	14º	mössbauer	26	2	0,2	0,0845	**5**
E04	15º	decomposição de 2o2	3	1	1,1	0,0767	**0**
E04	16º	espectroscopia de energia dispersiva	3	1	1,1	0,0767	**3**
E04	17º	largura de linha	3	1	1,1	0,0767	**0**
E04	18º	padrão de difração	3	1	1,1	0,0767	**5**
E04	19º	presente na amostra	3	1	1,1	0,0767	**0**
E04	20º	teores de nióbio	3	1	1,1	0,0767	**6**
E05	1º	metodologia experimental	25	1	0,8	0,4255	**1**
E05	2º	dióxido de titânio	8	2	1,1	0,1309	**6**
E05	3º	contato com a superfície do material	4	1	1,4	0,1191	**1**
E05	4º	ramo da curva de carregamento	4	1	1,4	0,1191	**2**
E05	5º	razão entre a tensão aplicada e a deformação elástica do material	4	1	1,4	0,1191	**2**

Doc.$_{cj}$	Pos.	Sintagma nominal candidato (i)	f_{ijc}	n_{ic}	CNP_i	$Score_{ijc}$	$Relevância_{ijc}$ (autor)
E05	6º	função da profundidade de deformação provocada no material	4	1	1,2	0,1021	**0**
E05	7º	combinação de deformação elástica	4	1	1,1	0,0936	**0**
E05	8º	módulo de young	4	1	1,1	0,0936	**3**
E05	9º	processo de carregamento	4	1	1,1	0,0936	**0**
E05	10º	umidade relativa do ar	4	1	1,1	0,0936	**1**
E05	11º	ponta	5	1	0,8	0,0851	**2**
E05	12º	sol-gel	19	1	0,2	0,0809	**6**
E05	13º	silver	18	1	0,2	0,0766	**5**
E05	14º	dois tipos de deformação	3	1	1,1	0,0702	**0**
E05	15º	isopropóxido de titânio	3	1	1,1	0,0702	**3**
E05	16º	nanocompósitos formados por prata	3	1	1,1	0,0702	**6**
E05	17º	nanopartículas de prata	3	1	1,1	0,0702	**5**
E05	18º	preparação das soluções	3	1	1,1	0,0702	**1**
E05	19º	processo de deposição	3	1	1,1	0,0702	**3**
E05	20º	substratos de vidro	3	1	1,1	0,0702	**4**
E06	1º	abundância relativa	22	1	0,8	0,1266	**3**
E06	2º	alíquotas retiradas após sucessivos tempos de exposição	8	1	1,4	0,0806	**3**
E06	3º	etinilestradiol	54	1	0,2	0,0777	**6**
E06	4º	min de reação	13	2	1,1	0,0719	**4**
E06	5º	fenton heterogêneo	17	2	0,8	0,0684	**6**

Doc.$_{cj}$	Pos.	Sintagma nominal candidato (i)	f_{ijc}	n_{ic}	CNP_i	$Score_{ijc}$	$Relevância_{ijc}$ (autor)
E06	6º	fólico	46	1	0,2	0,0662	**6**
E06	7º	função do tempo	8	1	1,1	0,0633	**3**
E06	8º	clofíbrico	44	1	0,2	0,0633	**6**
E06	9º	solução inicial	11	1	0,8	0,0633	**4**
E06	10º	carbamazepina	40	1	0,2	0,0576	**6**
E06	11º	degradação da carbamazepina	6	1	1,1	0,0475	**6**
E06	12º	degradação do etinilestradiol	6	1	1,1	0,0475	**6**
E06	13º	presença de nacl	6	1	1,1	0,0475	**6**
E06	14º	ozônio	28	1	0,2	0,0403	**5**
E06	15º	infusão direta	7	1	0,8	0,0403	**5**
E06	16º	coletadas em os tempos	4	1	1,4	0,0403	**0**
E06	17º	frações dos ânions	4	1	1,4	0,0403	**4**
E06	18º	função do tempo de exposição	4	1	1,4	0,0403	**5**
E06	19º	degradação do hormônio	5	1	1,1	0,0396	**6**
E06	20º	solução de etinilestradiol	5	1	1,1	0,0396	**6**
E07	1º	gálio	80	1	0,2	0,1151	**4**
E07	2º	chem	139	3	0,2	0,1046	**0**
E07	3º	espectros de rmn	23	3	1,1	0,0952	**1**
E07	4º	lassbio-1064	62	1	0,2	0,0892	**1**
E07	5º	hidrazonas	55	1	0,2	0,0791	**4**
E07	6º	ponto de fusão	10	1	1,1	0,0791	**1**
E07	7º	hidrazona	50	1	0,2	0,0719	**4**

Doc.$_{cj}$	Pos.	Sintagma nominal candidato (i)	f_{ijc}	n_{ic}	CNP_i	$Score_{ijc}$	$Relevância_{ijc}$ (autor)
E07	8º	padrão entre parênteses	9	1	1,1	0,0712	**0**
E07	9º	isômero	116	4	0,2	0,0664	**0**
E07	10º	células de glioblastoma	8	1	1,1	0,0633	**4**
E07	11º	hidrazonas derivadas de 2-acetilpiridina	8	1	1,1	0,0633	**5**
E07	12º	med	62	2	0,2	0,0624	**0**
E07	13º	2ac4oclph	41	1	0,2	0,0590	**1**
E07	14º	hidrazonas derivadas de 2-acetilpiridina e 2-benzoilpiridina	7	1	1,1	0,0554	**5**
E07	15º	influência de algumas	7	1	1,1	0,0554	**0**
E07	16º	avaliação da atividade	10	2	1,1	0,0553	**6**
E07	17º	zinco	54	2	0,2	0,0543	**4**
E07	18º	antimônio	53	2	0,2	0,0533	**4**
E07	19º	2ac4ofph	37	1	0,2	0,0532	**1**
E07	20º	h2bz4ono2	37	1	0,2	0,0532	**1**
E08	1º	tppo	352	1	0,2	0,2000	**0**
E08	2º	·tfnm0	94	1	0,2	0,0534	**0**
E08	3º	precursores isolados	22	1	0,8	0,0500	**0**
E08	4º	·actl0	85	1	0,2	0,0483	**0**
E08	5º	tfnm	76	1	0,2	0,0432	**0**
E08	6º	espectroscopia de vida média de pósitrons	10	1	1,4	0,0398	**6**
E08	7º	actl	68	1	0,2	0,0386	**0**
E08	8º	volume livre	14	1	0,8	0,0318	**3**

Doc.$_{cj}$	Pos.	Sintagma nominal candidato (i)	f_{ijc}	n_{ic}	CNP_i	$Score_{ijc}$	$Relevância_{ijc}$ (autor)
E08	9º	·tfnmx	52	1	0,2	0,0295	**0**
E08	10º	molecular na região	13	2	1,1	0,0284	**0**
E08	11º	…	49	1	0,2	0,0278	**0**
E08	12º	existentes em os precursores isolados	7	1	1,4	0,0278	**0**
E08	13º	curva de dtg	8	1	1,1	0,0250	**3**
E08	14º	faixa de composição	8	1	1,1	0,0250	**0**
E08	15º	através da temperatura no ponto	6	1	1,4	0,0239	**0**
E08	16º	oxigênio conjugado	10	1	0,8	0,0227	**0**
E08	17º	tppo isolado	10	1	0,8	0,0227	**0**
E08	18º	faixa de composição compreendida	7	1	1,1	0,0219	**0**
E08	19º	temperatura no ponto	7	1	1,1	0,0219	**0**
E08	20º	complexos supramoleculares	9	1	0,8	0,0205	**6**
E09	1º	solos	60	1	0,2	0,2000	**4**
E09	2º	ardley	56	1	0,2	0,1867	**6**
E09	3º	antarctica	52	1	0,2	0,1733	**6**
E09	4º	fildes	52	1	0,2	0,1733	**6**
E09	5º	ilha	50	1	0,2	0,1667	**0**
E09	6º	soils	50	1	0,2	0,1667	**4**
E09	7º	ornitogênicos	49	1	0,2	0,1633	**4**
E09	8º	antártica marítima	12	1	0,8	0,1600	**6**
E09	9º	ornitogênica	47	1	0,2	0,1567	**4**
E09	10º	faixa de concentração	8	1	1,1	0,1467	**0**

Doc.$_{cj}$	Pos.	Sintagma nominal candidato (i)	f_{ijc}	n_{ic}	CNP_i	$Score_{ijc}$	$Relevância_{ijc}$ (autor)
E09	11º	antarctic	41	1	0,2	0,1367	**6**
E09	12º	península	37	1	0,2	0,1233	**0**
E09	13º	rei	33	1	0,2	0,1100	**0**
E09	14º	george	46	2	0,2	0,1072	**0**
E09	15º	from	28	1	0,2	0,0933	**0**
E09	16º	terra nova	7	1	0,8	0,0933	**0**
E09	17º	autovalores da pc3 em função	4	1	1,4	0,0933	**0**
E09	18º	agitadora com velocidade	5	1	1,1	0,0917	**0**
E09	19º	horas em uma mesa	5	1	1,1	0,0917	**0**
E09	20º	antártica	26	1	0,2	0,0867	**6**
E10	1º	#REF!	3	0	#N/D	#DIV/0!	**0**
E10	2º	avaliação da atividade	29	2	1,1	0,1715	**4**
E10	3º	2ac4	99	1	0,2	0,1523	**0**
E10	4º	antimônio	130	2	0,2	0,1398	**5**
E10	5º	bismuto	129	2	0,2	0,1387	**5**
E10	6º	tiossemicarbazonas	102	2	0,2	0,1097	**6**
E10	7º	três experimentos independentes feitos em triplicata	12	1	1,1	0,1015	**0**
E10	8º	principais bandas em os espectros	9	1	1,4	0,0969	**2**
E10	9º	h2bz4m	61	1	0,2	0,0938	**1**
E10	10º	avaliada segundo método descrito na seção	8	1	1,4	0,0862	**1**
E10	11º	bandas em os espectros	8	1	1,4	0,0862	**1**
E10	12º	tiossemicarbazona	78	2	0,2	0,0839	**6**

Doc.$_{cj}$	Pos.	Sintagma nominal candidato (i)	f_{ijc}	n_{ic}	CNP_i	$Score_{ijc}$	$Relevância_{ijc}$ (autor)
E10	13º	jurkat	52	1	0,2	0,0800	**6**
E10	14º	atribuições e deslocamentos químicos dos sinais de rmn	12	2	1,2	0,0774	**0**
E10	15º	procedimento descrito na seção	9	1	1,1	0,0762	**1**
E10	16º	valores teóricos em parêntesis	9	1	1,1	0,0762	**0**
E10	17º	valores de cim	12	2	1,1	0,0710	**4**
E10	18º	linhagens de célula	8	1	1,1	0,0677	**5**
E10	19º	linhagens de células	8	1	1,1	0,0677	**5**
E10	20º	mapas de contorno	8	1	1,1	0,0677	**1**
F01	1º	#REF!	12	0	#N/D	#DIV/0!	**0**
F01	2º	espectrometria de massa	12	2	1,1	0,3974	**5**
F01	3º	glândula de veneno	6	1	1,1	0,2980	**4**
F01	4º	modo	9	2	0,8	0,2167	**0**
F01	5º	barata	6	1	0,8	0,2167	**1**
F01	6º	escoubas	20	1	0,2	0,1806	**0**
F01	7º	armadilha de íons	3	1	1,1	0,1490	**2**
F01	8º	cid de alta energia	3	1	1,1	0,1490	**2**
F01	9º	cid de baixa energia	3	1	1,1	0,1490	**2**
F01	10º	determinação da massa	3	1	1,1	0,1490	**2**
F01	11º	efeitos da toxina	3	1	1,1	0,1490	**4**
F01	12º	veneno de acanthoscurria	3	1	1,1	0,1490	**6**
F01	13º	periplaneta americana	4	1	0,8	0,1445	**1**
F01	14º	frações de interesse obtidas no passo	2	1	1,4	0,1264	**0**

Doc.$_{cj}$	Pos.	Sintagma nominal candidato (i)	f_{ijc}	n_{ic}	CNP_i	$Score_{ijc}$	$Relevância_{ijc}$ (autor)
F01	15º	número de grupos contidos no fragmento	2	1	1,4	0,1264	**0**
F01	16º	que a massa dos íons	2	1	1,4	0,1264	**0**
F01	17º	seqüenciamento n-terminal por degradação de edman	2	1	1,4	0,1264	**4**
F01	18º	universidade estadual de feira de santana	2	1	1,4	0,1264	**0**
F01	19º	trtx	13	1	0,2	0,1174	**3**
F01	20º	íons	19	2	0,2	0,1144	**0**
F02	1º	isoformas de tripsina	23	1	1,1	0,2355	**6**
F02	2º	fase móvel constituída de tampão	15	1	1,1	0,1536	**1**
F02	3º	ß-tripsina	69	1	0,2	0,1285	**6**
F02	4º	tripsinogênio	67	1	0,2	0,1248	**6**
F02	5º	mmol	97	2	0,2	0,1204	**0**
F02	6º	estacionária se-sephadex	16	1	0,8	0,1192	**3**
F02	7º	fase móvel	16	1	0,8	0,1192	**3**
F02	8º	tempo de retenção	11	1	1,1	0,1127	**4**
F02	9º	hcal	56	1	0,2	0,1043	**2**
F02	10º	mol	83	2	0,2	0,1030	**0**
F02	11º	amidásica relativa à isoforma	9	1	1,1	0,0922	**4**
F02	12º	dados de massa	9	1	1,1	0,0922	**0**
F02	13º	medida da atividade	9	1	1,1	0,0922	**3**
F02	14º	a-tripsina	48	1	0,2	0,0894	**6**
F02	15º	resultados e discussão	90	3	0,2	0,0790	**0**
F02	16º	atividade relativa	10	1	0,8	0,0745	**0**

Doc.$_{cj}$	Pos.	Sintagma nominal candidato (i)	f_{ijc}	n_{ic}	CNP_i	$Score_{ijc}$	$Relevância_{ijc}$ (autor)
F02	17º	seguintes	10	1	0,8	0,0745	**0**
F02	18º	faixa de temperatura	7	1	1,1	0,0717	**1**
F02	19º	variação da capacidade	7	1	1,1	0,0717	**0**
F02	20º	atividade demonstrada	9	1	0,8	0,0670	**0**
F03	1º	#REF!	1	0	#N/D	#DIV/0!	**0**
F03	2º	abortus	130	1	0,2	0,1806	**2**
F03	3º	bmmøs de camundongos	13	1	1,1	0,0993	**5**
F03	4º	receptor de interferon do tipo	9	1	1,4	0,0875	**5**
F03	5º	indução de interferon do tipo	6	1	1,4	0,0584	**4**
F03	6º	ifn-aßr	41	1	0,2	0,0570	**6**
F03	7º	provenientes de camundongos	7	1	1,1	0,0535	**0**
F03	8º	semanas após a infecção	7	1	1,1	0,0535	**1**
F03	9º	celular programada	9	1	0,8	0,0500	**0**
F03	10º	imune inata à infecção por brucella	5	1	1,4	0,0486	**3**
F03	11º	indução de interferon do tipo i	5	1	1,4	0,0486	**6**
F03	12º	índice de morte	6	1	1,1	0,0458	**0**
F03	13º	dna purificado	12	2	0,8	0,0445	**4**
F03	14º	129sv	31	1	0,2	0,0431	**2**
F03	15º	myd88	29	1	0,2	0,0403	**5**
F03	16º	trif	29	1	0,2	0,0403	**5**
F03	17º	adaptadora	28	1	0,2	0,0389	**0**
F03	18º	expressão de interferon do tipo	4	1	1,4	0,0389	**5**

Doc.$_{cj}$	Pos.	Sintagma nominal candidato (i)	f_{ijc}	n_{ic}	CNP_i	$Score_{ijc}$	$Relevância_{ijc}$ (autor)
F03	19º	sistema de interferon do tipo i	4	1	1,4	0,0389	**6**
F03	20º	receptores do tipo	5	1	1,1	0,0382	**0**
F04	1º	10µg de ova	28	1	1,1	0,2991	**2**
F04	2º	meses de idade	28	1	1,1	0,2991	**6**
F04	3º	diferença estatística entre os grupos	22	1	1,4	0,2991	**2**
F04	4º	tolerância oral	28	1	0,8	0,2175	**6**
F04	5º	ova	93	1	0,2	0,1806	**3**
F04	6º	animais grupo	23	1	0,8	0,1787	**6**
F04	7º	20mg de ova	16	1	1,1	0,1709	**2**
F04	8º	dias após o tratamento oral	14	1	1,1	0,1495	**6**
F04	9º	3mg	64	1	0,2	0,1243	**1**
F04	10º	animais	46	4	0,8	0,1191	**6**
F04	11º	manutenção da tolerância oral	11	1	1,1	0,1175	**6**
F04	12º	gavagem	56	1	0,2	0,1088	**6**
F04	13º	2ml de solução salina	10	1	1,1	0,1068	**1**
F04	14º	cultura de células	9	1	1,1	0,0961	**6**
F04	15º	elisa e os resultados expressos como média aritmética	9	1	1,1	0,0961	**4**
F04	16º	número de animais grupo	9	1	1,1	0,0961	**3**
F04	17º	absorbância obtida com os soros totais	7	1	1,4	0,0952	**1**
F04	18º	padrão da porcentagem de células	7	1	1,4	0,0952	**3**
F04	19º	imunização primária	12	1	0,8	0,0932	**6**
F04	20º	média aritmética desvio	12	1	0,8	0,0932	**3**

Doc.$_{cj}$	Pos.	Sintagma nominal candidato (i)	f_{ijc}	n_{ic}	CNP_i	$Score_{ijc}$	$Relevância_{ijc}$ (autor)
F05	1º	nocautes para tbrad51	19	1	1,1	0,1226	**4**
F05	2º	cruzi	154	2	0,2	0,1204	**6**
F05	3º	rad51	88	1	0,2	0,1032	**6**
F05	4º	gene de resistência	14	1	1,1	0,0903	**0**
F05	5º	heminocautes de tcrad51	13	1	1,1	0,0839	**5**
F05	6º	brucei	100	2	0,2	0,0782	**6**
F05	7º	média dos valores obtidos	9	1	1,4	0,0739	**0**
F05	8º	parasitos	83	2	0,2	0,0649	**5**
F05	9º	tcrad51	83	2	0,2	0,0649	**5**
F05	10º	peróxido de hidrogênio	15	2	1,1	0,0645	**4**
F05	11º	nocautes para rad51	9	1	1,1	0,0581	**4**
F05	12º	valores mostrados	12	1	0,8	0,0563	**0**
F05	13º	parasitos selvagens	17	2	0,8	0,0532	**3**
F05	14º	padrão das triplicatas	8	1	1,1	0,0516	**0**
F05	15º	média percentual de triplicatas das células tratadas em relação	7	1	1,2	0,0493	**0**
F05	16º	tbrad51	40	1	0,2	0,0469	**5**
F05	17º	brucei selvagem	9	1	0,8	0,0422	**1**
F05	18º	média de um experimento realizado em triplicata	5	1	1,4	0,0410	**0**
F05	19º	número de células	9	2	1,1	0,0387	**0**
F05	20º	nocaute de tbrad51	6	1	1,1	0,0387	**4**
F06	1º	expressão da lgmn	11	1	1,1	0,1301	**6**

Doc.$_{cj}$	Pos.	Sintagma nominal candidato (i)	f_{ijc}	n_{ic}	CNP_i	$Score_{ijc}$	$Relevância_{ijc}$ (autor)
F06	2º	expressão de lgmn	11	1	1,1	0,1301	**6**
F06	3º	núcleo da célula	10	1	1,1	0,1183	**5**
F06	4º	fatores de transcrição	9	1	1,1	0,1064	**2**
F06	5º	tamponamento do nuclear	8	1	1,1	0,0946	**6**
F06	6º	regeneração hepática	10	1	0,8	0,0860	**2**
F06	7º	fatores de crescimento	7	1	1,1	0,0828	**2**
F06	8º	média desvio	9	1	0,8	0,0774	**5**
F06	9º	padrão de três experimentos individuais	6	1	1,1	0,0710	**5**
F06	10º	silenciamento da lgmn	6	1	1,1	0,0710	**6**
F06	11º	lgmn	30	1	0,2	0,0645	**6**
F06	12º	análises de western	8	2	1,1	0,0631	**6**
F06	13º	skhep1	29	1	0,2	0,0624	**5**
F06	14º	pfu animal	7	1	0,8	0,0602	**0**
F06	15º	silenciamento dos insp	4	1	1,4	0,0602	**0**
F06	16º	silenciamento dos insp3	4	1	1,4	0,0602	**0**
F06	17º	processo de regeneração hepática	5	1	1,1	0,0591	**2**
F06	18º	progressão do ciclo	5	1	1,1	0,0591	**6**
F06	19º	insp	27	1	0,2	0,0581	**4**
F06	20º	hepatectomia parcial	6	1	0,8	0,0516	**2**
F07	1º	inibidores de fviii	13	1	1,1	0,1819	**6**
F07	2º	haa-fviii	71	1	0,2	0,1806	**3**
F07	3º	portadores de hemofilia	10	1	1,1	0,1399	**2**

Doc.$_{cj}$	Pos.	Sintagma nominal candidato (i)	f_{ijc}	n_{ic}	CNP_i	$Score_{ijc}$	$Relevância_{ijc}$ (autor)
F07	4º	fviii	39	1	0,2	0,0992	**5**
F07	5º	desenvolvimento de inibidores	7	1	1,1	0,0979	**5**
F07	6º	anti-fviii	32	1	0,2	0,0814	**6**
F07	7º	desenvolvimento de inibidores de fviii	4	1	1,4	0,0712	**6**
F07	8º	inibidores do fviii	5	1	1,1	0,0700	**6**
F07	9º	peptídeo cíclico	6	1	0,8	0,0611	**2**
F07	10º	segmento	35	2	0,2	0,0594	**0**
F07	11º	imune contra o fviii	4	1	1,1	0,0560	**2**
F07	12º	predominância de resposta	4	1	1,1	0,0560	**0**
F07	13º	promotora do gene	4	1	1,1	0,0560	**0**
F07	14º	síntese de anticorpos	4	1	1,1	0,0560	**2**
F07	15º	comparações com os grupos	3	1	1,4	0,0534	**0**
F07	16º	ligação dos anticorpos	3	1	1,4	0,0534	**4**
F07	17º	barras cinza-claro	5	1	0,8	0,0509	**0**
F07	18º	barras cinza-escuro	5	1	0,8	0,0509	**0**
F07	19º	inibidores	19	1	0,2	0,0483	**5**
F07	20º	pacientes	28	2	0,2	0,0475	**1**
F08	1º	página	198	1	0,2	0,1806	**0**
F08	2º	cruzi	196	2	0,2	0,1192	**3**
F08	3º	frequência de mutantes	16	1	1,1	0,0803	**5**
F08	4º	artigo em preparação	11	1	1,1	0,0552	**0**
F08	5º	reparo por excisão de bases	8	1	1,4	0,0511	**6**

Doc.$_{cj}$	Pos.	Sintagma nominal candidato (i)	f_{ijc}	n_{ic}	CNP_i	$Score_{ijc}$	$Relevância_{ijc}$ (autor)
F08	6º	clonagem em pmal	10	1	1,1	0,0502	**2**
F08	7º	reparo de dna	14	2	1,1	0,0468	**6**
F08	8º	reparo de erros de pareamento	7	1	1,4	0,0447	**6**
F08	9º	crimp4	43	1	0,2	0,0392	**4**
F08	10º	corados com brometo de etídeo	6	1	1,4	0,0383	**0**
F08	11º	diferentes construções	10	1	0,8	0,0365	**1**
F08	12º	genômico de brener	7	1	1,1	0,0351	**0**
F08	13º	tratamento com 2o2	7	1	1,1	0,0351	**4**
F08	14º	mansoni	38	1	0,2	0,0347	**3**
F08	15º	oxidativo	56	2	0,2	0,0341	**5**
F08	16º	c2g vazio	9	1	0,8	0,0328	**0**
F08	17º	danos no dna	9	2	1,1	0,0301	**6**
F08	18º	clonado no vetor	6	1	1,1	0,0301	**0**
F08	19º	cloreto de cádmio	6	1	1,1	0,0301	**3**
F08	20º	média de três experimentos independentes	6	1	1,1	0,0301	**0**
G01	1º	ciências cognitivas	43	2	0,8	0,1849	**6**
G01	2º	estudo em os artigos científicos publicados	12	1	1,4	0,1355	**3**
G01	3º	redes cognitivas na ciência da informação brasileira	12	1	1,4	0,1355	**5**
G01	4º	análise do artigo	14	1	1,1	0,1242	**6**
G01	5º	grifos do pesquisador	13	1	1,1	0,1153	**4**
G01	6º	migração conceitual	16	1	0,8	0,1032	**6**
G01	7º	organização virtual do conhecimento no ciberespaço	9	1	1,4	0,1016	**4**

Doc.$_{cj}$	Pos.	Sintagma nominal candidato (i)	f_{ijc}	n_{ic}	CNP_i	$Score_{ijc}$	$Relevância_{ijc}$ (autor)
G01	8º	art1 na recuperação da informação	8	1	1,4	0,0903	**5**
G01	9º	análise do interdiscurso	9	1	1,1	0,0798	**6**
G01	10º	poder cognitivo das redes neurais artificiais	9	1	1,1	0,0798	**5**
G01	11º	conceito central	17	2	0,8	0,0731	**5**
G01	12º	carga cognitiva	11	1	0,8	0,0710	**5**
G01	13º	patologia metodológica	11	1	0,8	0,0710	**5**
G01	14º	conceito de carga cognitiva	7	1	1,1	0,0621	**5**
G01	15º	conceito de rizoma	7	1	1,1	0,0621	**5**
G01	16º	artificiais modelo	9	1	0,8	0,0581	**2**
G01	17º	sri por o autores dos artigos	6	1	1,2	0,0581	**4**
G01	18º	maingueneau	35	1	0,2	0,0564	**6**
G01	19º	conceito entre os dois artigos	5	1	1,4	0,0564	**5**
G01	20º	conceito de imagem mental	6	1	1,1	0,0532	**5**
G02	1º	saúde da família	49	1	1,1	0,1894	**2**
G02	2º	siab	257	1	0,2	0,1806	**6**
G02	3º	relações de poder	43	1	1,1	0,1662	**4**
G02	4º	ministério da saúde	35	1	1,1	0,1353	**0**
G02	5º	fluxo informacional do siab	29	1	1,1	0,1121	**3**
G02	6º	dados do siab	25	1	1,1	0,0966	**3**
G02	7º	branco	25	1	0,8	0,0703	**0**
G02	8º	categoria de análise	22	2	1,1	0,0567	**1**
G02	9º	sistemas de informação em saúde	11	1	1,4	0,0541	**5**

Doc.$_{cj}$	Pos.	Sintagma nominal candidato (i)	f_{ijc}	n_{ic}	CNP_i	$Score_{ijc}$	$Relevância_{ijc}$ (autor)
G02	10º	profissionais de saúde	13	1	1,1	0,0502	**1**
G02	11º	informação em saúde	12	1	1,1	0,0464	**4**
G02	12º	enfermeiro	16	1	0,8	0,0450	**1**
G02	13º	sistema de informação da atenção básica	9	1	1,4	0,0443	**6**
G02	14º	sistema único de saúde	11	1	1,1	0,0425	**1**
G02	15º	profissionais	64	5	0,8	0,0407	**0**
G02	16º	unidade de saúde da família	8	1	1,4	0,0394	**2**
G02	17º	informações em saúde	10	1	1,1	0,0387	**4**
G02	18º	rede de olhares	8	1	1,1	0,0309	**2**
G02	19º	consequências das relações de poder	5	1	1,4	0,0246	**2**
G02	20º	estratégia de saúde da família	5	1	1,4	0,0246	**2**
G03	1º	cultura do sorgo	12	1	1,1	0,2020	**5**
G03	2º	método analítico-sintético	14	1	0,8	0,1714	**5**
G03	3º	sorgo	45	1	0,2	0,1378	**5**
G03	4º	representação do conhecimento	12	2	1,1	0,1347	**6**
G03	5º	teoria do conceito	8	1	1,1	0,1347	**5**
G03	6º	atividades de pesquisa	6	1	1,1	0,1010	**1**
G03	7º	empresa brasileira de pesquisa agropecuária	6	1	1,1	0,1010	**1**
G03	8º	ciência da computação	23	5	1,1	0,0875	**1**
G03	9º	modelagem para representação do conhecimento	4	1	1,4	0,0857	**6**
G03	10º	mapa conceitual da classe	5	1	1,1	0,0842	**4**
G03	11º	plano das idéias	5	1	1,1	0,0842	**1**

Doc.$_{cj}$	Pos.	Sintagma nominal candidato (i)	f_{ijc}	n_{ic}	CNP_i	$Score_{ijc}$	$Relevância_{ijc}$ (autor)
G03	12º	garantia literária	10	2	0,8	0,0816	**5**
G03	13º	embrapa	24	1	0,2	0,0735	**1**
G03	14º	estrutura semântica	9	2	0,8	0,0735	**2**
G03	15º	alimentação animal	6	1	0,8	0,0735	**1**
G03	16º	análise de assunto	9	3	1,1	0,0715	**4**
G03	17º	árvore do conhecimento	4	1	1,1	0,0673	**1**
G03	18º	classificatória do sorgo	4	1	1,1	0,0673	**2**
G03	19º	divisão do texto	4	1	1,1	0,0673	**3**
G03	20º	nome em citações bibliográficas	4	1	1,1	0,0673	**0**
G04	1º	canais do youtube	37	1	1,1	0,2723	**6**
G04	2º	qualidade da informação	30	1	1,1	0,2208	**6**
G04	3º	youtube	135	1	0,2	0,1806	**6**
G04	4º	qualidade da informação e produsage	17	1	1,1	0,1251	**6**
G04	5º	blog	76	1	0,2	0,1017	**6**
G04	6º	blogueiro	71	1	0,2	0,0950	**5**
G04	7º	blogs	86	2	0,2	0,0767	**6**
G04	8º	blogueiros	52	1	0,2	0,0696	**5**
G04	9º	canal	52	1	0,2	0,0696	**0**
G04	10º	software livre	13	1	0,8	0,0696	**4**
G04	11º	canal do youtube	9	1	1,1	0,0662	**6**
G04	12º	vídeos	72	2	0,2	0,0642	**3**
G04	13º	sites de redes sociais	8	1	1,1	0,0589	**3**

Doc.$_{cj}$	Pos.	Sintagma nominal candidato (i)	f_{ijc}	n_{ic}	CNP_i	$Score_{ijc}$	$Relevância_{ijc}$ (autor)
G04	14º	tecnologias digitais	10	1	0,8	0,0535	**4**
G04	15º	blogs e canais do youtube	7	1	1,1	0,0515	**6**
G04	16º	percepção de qualidade	7	1	1,1	0,0515	**5**
G04	17º	vídeo	38	1	0,2	0,0508	**3**
G04	18º	objeto dinâmico	9	1	0,8	0,0482	**4**
G04	19º	produsers	34	1	0,2	0,0455	**6**
G04	20º	pesquisa de mestrado	6	1	1,1	0,0442	**0**
G05	1º	busca e uso da informação	20	1	1,1	0,1352	**6**
G05	2º	decisão estratégica em empresas	17	1	1,1	0,1149	**6**
G05	3º	integrativo	80	1	0,2	0,0983	**1**
G05	4º	decisão estratégica	18	1	0,8	0,0885	**6**
G05	5º	fontes pessoais	15	1	0,8	0,0737	**2**
G05	6º	fontes internas	14	1	0,8	0,0688	**2**
G05	7º	„modelo para identificação das necessidades	8	1	1,4	0,0688	**6**
G05	8º	„modelo	52	1	0,2	0,0639	**6**
G05	9º	contato direto	13	1	0,8	0,0639	**0**
G05	10º	demandante da informação	9	1	1,1	0,0608	**3**
G05	11º	autor com base em choo	7	1	1,4	0,0602	**3**
G05	12º	„modelo geral	11	1	0,8	0,0541	**4**
G05	13º	comportamento de uso da informação	6	1	1,4	0,0516	**6**
G05	14º	„modelo da cadeia	7	1	1,1	0,0473	**4**
G05	15º	comportamento informacional para decisões estratégicas	7	1	1,1	0,0473	**6**

Doc.$_{cj}$	**Pos.**	**Sintagma nominal candidato (i)**	**f**$_{ijc}$	**n**$_{ic}$	**CNP**$_{i}$	**Score**$_{ijc}$	**Relevância**$_{ijc}$ **(autor)**
G05	16º	uso da informação	14	3	1,1	0,0446	**6**
G05	17º	informações obtidas	9	1	0,8	0,0442	**0**
G05	18º	sócio diretor	9	1	0,8	0,0442	**0**
G05	19º	autor com base em dervin	5	1	1,4	0,0430	**3**
G05	20º	conflito sobre os objetivos	5	1	1,4	0,0430	**0**
G06	1º	segundo os autores	105	3	1,4	0,2846	**0**
G06	2º	recuperação de informação	35	2	1,1	0,1054	**5**
G06	3º	área de pln	22	1	1,1	0,0993	**4**
G06	4º	tratamento de ambiguidade	22	1	1,1	0,0993	**3**
G06	5º	strube de lima	15	1	1,1	0,0677	**0**
G06	6º	linguagem natural	20	1	0,8	0,0657	**6**
G06	7º	lexical	67	1	0,2	0,0550	**2**
G06	8º	processamento de linguagem natural	18	2	1,1	0,0542	**6**
G06	9º	gramática	23	2	0,8	0,0504	**3**
G06	10º	corpus	60	1	0,2	0,0493	**3**
G06	11º	tradução automática	14	1	0,8	0,0460	**4**
G06	12º	ainda segundo os autores	8	1	1,4	0,0460	**0**
G06	13º	experimentos práticos	13	1	0,8	0,0427	**4**
G06	14º	respondedores automáticos	11	1	0,8	0,0361	**4**
G06	15º	capítulo de revisão	8	1	1,1	0,0361	**4**
G06	16º	moraes e strube de lima	8	1	1,1	0,0361	**0**
G06	17º	tamanho da amostra	8	1	1,1	0,0361	**0**

Doc$_{.cj}$	**Pos.**	**Sintagma nominal candidato (i)**	**f**$_{ijc}$	**n**$_{ic}$	**CNP**$_{i}$	**Score**$_{ijc}$	**Relevância**$_{ijc}$ **(autor)**
G06	18º	sentenças	42	1	0,2	0,0345	**0**
G06	19º	capítulos de revisão do arist	6	1	1,4	0,0345	**5**
G06	20º	humana til	10	1	0,8	0,0328	**0**
G07	1º	gestão de documentos	27	1	1,1	0,3119	**3**
G07	2º	autora no protégé	25	1	1,1	0,2888	**0**
G07	3º	categoria específica	33	1	0,8	0,2772	**3**
G07	4º	análise de domínio	21	1	1,1	0,2426	**6**
G07	5º	categorias específicas	23	1	0,8	0,1932	**3**
G07	6º	dirks	86	1	0,2	0,1806	**6**
G07	7º	subcategorias e entidades dos formulários	12	1	1,4	0,1764	**3**
G07	8º	fonte de informação	30	3	1,1	0,1635	**1**
G07	9º	ato normativo	19	1	0,8	0,1596	**0**
G07	10º	data do evento	11	1	1,1	0,1271	**0**
G07	11º	documentos de arquivo	11	1	1,1	0,1271	**1**
G07	12º	instrumentos de apoio	11	1	1,1	0,1271	**0**
G07	13º	passos da metodologia	11	1	1,1	0,1271	**2**
G07	14º	categoria fundamental	15	1	0,8	0,1260	**3**
G07	15º	ato legal	14	1	0,8	0,1176	**0**
G07	16º	arquivo nacional da austrália	10	1	1,1	0,1155	**2**
G07	17º	ciclo de vida	14	2	1,1	0,1078	**2**
G07	18º	elemento de identificação	9	1	1,1	0,1040	**0**
G07	19º	fluxo da transação	9	1	1,1	0,1040	**1**

Doc.$_{cj}$	Pos.	Sintagma nominal candidato (i)	f_{ijc}	n_{ic}	CNP_i	$Score_{ijc}$	$Relevância_{ijc}$ (autor)
G07	20º	operações da organização	9	1	1,1	0,1040	**0**
G08	1º	rita do sapucaí	59	1	1,1	0,5861	**2**
G08	2º	gonzález de gómez	66	4	1,1	0,2185	**4**
G08	3º	conceito de regime de informação	13	1	1,4	0,1644	**6**
G08	4º	regime de informação	16	1	1,1	0,1589	**6**
G08	5º	política de informação	15	1	1,1	0,1490	**6**
G08	6º	inatel	72	1	0,2	0,1300	**2**
G08	7º	mercado de destino da produção	10	1	1,4	0,1264	**2**
G08	8º	tradução da autora	14	2	1,1	0,0927	**0**
G08	9º	atores locais	11	1	0,8	0,0795	**6**
G08	10º	compartilhamento de informação	8	1	1,1	0,0795	**5**
G08	11º	foto da autora	8	1	1,1	0,0795	**0**
G08	12º	presentes no território	8	1	1,1	0,0795	**5**
G08	13º	teoria do regime	8	1	1,1	0,0795	**4**
G08	14º	conti	42	1	0,2	0,0759	**4**
G08	15º	santa	42	1	0,2	0,0759	**0**
G08	16º	interação entre os atores	6	1	1,4	0,0759	**6**
G08	17º	arranjos produtivos locais	10	1	0,8	0,0722	**6**
G08	18º	serviços de informação	9	2	1,1	0,0596	**4**
G08	19º	gonzález de gómez e canongia	6	1	1,1	0,0596	**1**
G08	20º	instituto nacional de telecomunicações	6	1	1,1	0,0596	**2**
H01	1º	sala de emergência	75	1	1,1	0,5909	**6**

Doc.$_{cj}$	Pos.	Sintagma nominal candidato (i)	f_{ijc}	n_{ic}	CNP_i	$Score_{ijc}$	$Relevância_{ijc}$ (autor)
H01	2º	profissionais de saúde	58	3	1,1	0,1990	**6**
H01	3º	profissionais de saúde da sala de emergência	15	1	1,2	0,1289	**3**
H01	4º	pessoa doente	20	1	0,8	0,1146	**0**
H01	5º	maffesoli	72	1	0,2	0,1031	**2**
H01	6º	julho de 2009	20	2	1,1	0,1014	**0**
H01	7º	humano	27	2	0,8	0,0996	**6**
H01	8º	humanização da assistência	12	1	1,1	0,0945	**6**
H01	9º	humanização	43	1	0,2	0,0616	**4**
H01	10º	profissional de saúde	12	2	1,1	0,0609	**6**
H01	11º	espaço-tempo da sala de emergência	5	1	1,4	0,0501	**6**
H01	12º	profissionais da sala de emergência	5	1	1,4	0,0501	**6**
H01	13º	pronto-socorro	34	1	0,2	0,0487	**6**
H01	14º	atendimento de urgência e emergência	6	1	1,1	0,0473	**6**
H01	15º	pessoas doentes	8	1	0,8	0,0458	**6**
H01	16º	vida quotidiana	8	1	0,8	0,0458	**6**
H01	17º	área da saúde	9	2	1,1	0,0456	**6**
H01	18º	enfermagem	30	1	0,2	0,0430	**4**
H01	19º	janeiro de 2010	8	2	1,1	0,0406	**0**
H01	20º	enfermeira	7	1	0,8	0,0401	**4**
H02	1º	questionário de competências específicas em medicina	10	1	1,4	0,1878	**6**
H02	2º	médico da ufmg	11	1	1,1	0,1623	**6**
H02	3º	centrada no paciente	10	1	1,1	0,1476	**4**

Doc.$_{cj}$	Pos.	Sintagma nominal candidato (i)	f_{ijc}	n_{ic}	CNP_i	$Score_{ijc}$	$Relevância_{ijc}$ (autor)
H02	4º	realização de procedimentos	10	1	1,1	0,1476	**6**
H02	5º	revista brasileira de educação	10	1	1,1	0,1476	**6**
H02	6º	estudantes	48	1	0,2	0,1288	**6**
H02	7º	faculdade de medicina da ufmg	10	2	1,4	0,1209	**6**
H02	8º	osce	45	1	0,2	0,1207	**6**
H02	9º	alunos	43	1	0,2	0,1154	**6**
H02	10º	realização do exame físico	7	1	1,1	0,1033	**6**
H02	11º	solução de problemas	7	1	1,1	0,1033	**5**
H02	12º	estudante	35	1	0,2	0,0939	**6**
H02	13º	adequação no trato com o paciente	5	1	1,4	0,0939	**4**
H02	14º	centradas no paciente	6	1	1,1	0,0885	**4**
H02	15º	comissão permanente de avaliação	6	1	1,1	0,0885	**6**
H02	16º	internato de pediatria	6	1	1,1	0,0885	**6**
H02	17º	revisão da literatura	34	5	1,1	0,0867	**6**
H02	18º	aluno	30	1	0,2	0,0805	**6**
H02	19º	conteúdo teórico	7	1	0,8	0,0751	**5**
H02	20º	avaliação em serviço	5	1	1,1	0,0738	**6**
H03	1º	centro de saúde	17	1	1,1	0,1816	**4**
H03	2º	sanitário centro-sul	22	1	0,8	0,1710	**2**
H03	3º	telessaúde	87	1	0,2	0,1690	**6**
H03	4º	recursos da telessaúde	11	1	1,1	0,1175	**6**
H03	5º	sanitário centro-sul de belo horizonte	11	1	1,1	0,1175	**3**

Doc.$_{cj}$	Pos.	Sintagma nominal candidato (i)	f_{ijc}	n_{ic}	CNP_i	$Score_{ijc}$	$Relevância_{ijc}$ (autor)
H03	6º	instrumento de suporte assistencial e educação permanente	10	1	1,1	0,1069	**3**
H03	7º	gerais brasil	12	1	0,8	0,0933	**3**
H03	8º	periódico line	12	1	0,8	0,0933	**0**
H03	9º	telemedicina	45	1	0,2	0,0874	**6**
H03	10º	uso da telessaúde	7	1	1,1	0,0748	**6**
H03	11º	ix congresso brasileiro de informática em saúde	5	1	1,4	0,0680	**0**
H03	12º	visão geral do estado da arte	5	1	1,4	0,0680	**0**
H03	13º	gerais universidade da fundação mineira de educação e cultura	4	1	1,4	0,0544	**0**
H03	14º	prática da telessaúde	5	1	1,1	0,0534	**6**
H03	15º	saúde de belo horizonte	5	1	1,1	0,0534	**6**
H03	16º	servicio de telesalud	5	1	1,1	0,0534	**6**
H03	17º	tecnologias de informação e comunicação	5	1	1,1	0,0534	**6**
H03	18º	tipo de atendimento	5	1	1,1	0,0534	**5**
H03	19º	telemedicine	27	1	0,2	0,0525	**6**
H03	20º	belo horizonte	23	4	0,8	0,0514	**4**
H04	1º	lacan	208	1	0,2	0,1347	**6**
H04	2º	tradução nossa	38	1	0,8	0,0984	**0**
H04	3º	imagem do corpo	22	1	1,1	0,0784	**6**
H04	4º	estádio do espelho	21	1	1,1	0,0748	**6**
H04	5º	corpo	261	3	0,2	0,0736	**6**

Doc.$_{cj}$	Pos.	Sintagma nominal candidato (i)	f_{ijc}	n_{ic}	CNP_i	$Score_{ijc}$	$Relevância_{ijc}$ (autor)
H04	6º	espelho	111	1	0,2	0,0719	**6**
H04	7º	foto alterada	21	1	0,8	0,0544	**0**
H04	8º	significante	77	1	0,2	0,0499	**5**
H04	9º	imagem	115	2	0,2	0,0479	**6**
H04	10º	imagem no espelho	11	1	1,1	0,0392	**6**
H04	11º	dismórfico corporal	15	1	0,8	0,0389	**6**
H04	12º	imagem alterada	14	1	0,8	0,0363	**0**
H04	13º	modo	32	3	0,8	0,0361	**0**
H04	14º	objeto	84	2	0,2	0,0350	**5**
H04	15º	sintoma	52	1	0,2	0,0337	**5**
H04	16º	dismorfofobia ligada	13	1	0,8	0,0337	**6**
H04	17º	orkut	49	1	0,2	0,0317	**1**
H04	18º	comida	12	1	0,8	0,0311	**1**
H04	19º	pai	106	3	0,2	0,0299	**5**
H04	20º	dismorfofobia	43	1	0,2	0,0278	**6**
H05	1º	primeiro momento da avaliação e o segundo momento da avaliação	14	1	1,2	0,2958	**6**
H05	2º	estados de saúde	13	1	1,1	0,2518	**6**
H05	3º	avaliação da qvrs	10	1	1,1	0,1937	**6**
H05	4º	global correspondente	12	1	0,8	0,1690	**6**
H05	5º	escores negativos	11	1	0,8	0,1549	**6**
H05	6º	estatisticamente significativa	11	1	0,8	0,1549	**6**

Doc.$_{cj}$	Pos.	Sintagma nominal candidato (i)	f_{ijc}	n_{ic}	CNP_i	$Score_{ijc}$	$Relevância_{ijc}$ (autor)
H05	7º	avaliação por terceiros	8	1	1,1	0,1549	**6**
H05	8º	dois momentos do estudo	8	1	1,1	0,1549	**6**
H05	9º	acometimento dos atributos	6	1	1,4	0,1479	**6**
H05	10º	hui2	41	1	0,2	0,1444	**6**
H05	11º	hui3	38	1	0,2	0,1338	**6**
H05	12º	saúde perfeita	9	1	0,8	0,1268	**6**
H05	13º	dor total de pacientes	6	1	1,1	0,1162	**6**
H05	14º	total de pacientes	6	1	1,1	0,1162	**6**
H05	15º	observada diferença	7	1	0,8	0,0986	**6**
H05	16º	escores globais de qvrs segundo o hui2	4	1	1,4	0,0986	**6**
H05	17º	escores globais de qvrs segundo o hui3	4	1	1,4	0,0986	**6**
H05	18º	estado de saúde	5	1	1,1	0,0968	**6**
H05	19º	atributo acometido	6	1	0,8	0,0845	**6**
H05	20º	c e d escore global	6	1	0,8	0,0845	**6**
H06	1º	novo cruzeiro	70	1	0,8	0,3506	**3**
H06	2º	deficiência de vitamina	30	1	1,1	0,2066	**6**
H06	3º	vitamin	135	1	0,2	0,1690	**1**
H06	4º	rio de janeiro	39	3	1,1	0,1169	**0**
H06	5º	badaró	85	1	0,2	0,1064	**0**
H06	6º	retinol	85	1	0,2	0,1064	**2**
H06	7º	leite por a vida	15	1	1,1	0,1033	**2**
H06	8º	saúde pública	32	2	0,8	0,1032	**5**

Doc.$_{cj}$	Pos.	Sintagma nominal candidato (i)	f_{ijc}	n_{ic}	CNP_i	$Score_{ijc}$	$Relevância_{ijc}$ (autor)
H06	9º	nutr	82	1	0,2	0,1027	**0**
H06	10º	prevalência de deficiência de vitamina	10	1	1,4	0,0876	**6**
H06	11º	fatores de risco	12	1	1,1	0,0826	**6**
H06	12º	prevalência de dva	12	1	1,1	0,0826	**5**
H06	13º	resposta de fase	12	1	1,1	0,0826	**2**
H06	14º	francisco	63	1	0,2	0,0789	**0**
H06	15º	ministério da saúde	26	3	1,1	0,0780	**0**
H06	16º	deficiency	62	1	0,2	0,0776	**2**
H06	17º	composição de alimentos	11	1	1,1	0,0757	**0**
H06	18º	suplementação de vitamina	11	1	1,1	0,0757	**2**
H06	19º	badaró e novo cruzeiro	14	1	0,8	0,0701	**0**
H06	20º	anos em francisco	9	1	1,1	0,0620	**0**
H07	1º	vje	69	1	0,2	0,1690	**6**
H07	2º	progressão do fio	8	1	1,1	0,1078	**5**
H07	3º	seldinger	41	1	0,2	0,1004	**5**
H07	4º	cvcp	38	1	0,2	0,0931	**6**
H07	5º	fio	34	1	0,2	0,0833	**0**
H07	6º	número de casos	6	1	1,1	0,0808	**0**
H07	7º	agulha metálica	7	1	0,8	0,0686	**0**
H07	8º	jugular externa	7	1	0,8	0,0686	**6**
H07	9º	junção da vji com a vsc	4	1	1,4	0,0686	**3**
H07	10º	através do fio	5	1	1,1	0,0674	**6**

Doc$_{.cj}$	**Pos.**	**Sintagma nominal candidato (i)**	**f**$_{ijc}$	**n**$_{ic}$	**CNP**$_{i}$	**Score**$_{ijc}$	**Relevância**$_{ijc}$ **(autor)**
H07	11º	introdução do fio	5	1	1,1	0,0674	**5**
H07	12º	punção da vje	5	1	1,1	0,0674	**5**
H07	13º	realização do cvcp	5	1	1,1	0,0674	**1**
H07	14º	posição periférica	6	1	0,8	0,0588	**2**
H07	15º	diâmetro do cateter	4	1	1,1	0,0539	**3**
H07	16º	introdução do cateter	4	1	1,1	0,0539	**0**
H07	17º	média de idade	4	1	1,1	0,0539	**0**
H07	18º	punção das veias	4	1	1,1	0,0539	**1**
H07	19º	blitt	21	1	0,2	0,0514	**0**
H07	20º	cateter	20	1	0,2	0,0490	**5**

Fonte: Elaborado pelo autor.

APÊNDICE J - ATRIBUIÇÃO DE VALOR DE RELEVÂNCIA EM DEZ PARTES DE CADA TESE DO *CORPUS*

Corpus	DOC	10%	20%	30%	40%	50%	60%	70%	80%	90%	100%	Total
A	**A01**	0,854	1,456	1,129	0,870	0,469	0,514	0,304	0,305	0,704	2,145	**8,75**
A	**A02**	1,097	1,247	0,591	1,052	1,850	0,867	1,097	1,272	1,425	1,251	**11,75**
A	**A03**	1,100	0,935	0,805	1,984	1,599	0,506	1,982	1,508	1,527	1,804	**13,75**
A	**A04**	0,778	0,477	1,208	2,163	2,451	1,300	3,265	2,704	0,551	1,353	**16,25**
A	**A05**	0,491	1,507	1,978	1,621	0,669	2,367	2,660	1,180	1,397	2,131	**16,00**
A	**A06**	0,024	0,205	0,559	0,100	0,129	0,458	0,812	0,346	0,326	0,542	**3,50**
A	**A07**	0,467	1,588	0,337	0,569	0,089	0,947	0,354	1,319	0,612	1,217	**7,50**
A	**A08**	1,067	1,215	0,451	0,833	0,814	1,492	1,240	1,782	1,345	1,760	**12,00**
A	**A09**	1,624	0,991	1,257	2,150	1,190	2,423	1,298	1,375	0,436	1,755	**14,50**
A	**A10**	1,098	0,961	0,890	0,000	0,395	0,960	0,889	1,076	1,611	1,120	**9,00**
A	**A11**	1,469	0,538	0,113	0,088	0,225	0,225	0,000	0,163	0,725	0,456	**4,00**
A	**A12**	0,870	0,878	0,804	0,327	1,214	1,436	0,557	1,041	0,927	1,946	**10,00**
A	**A13**	1,537	0,532	0,419	1,047	1,237	0,684	0,920	1,412	1,994	0,469	**10,25**
A	**A14**	1,270	1,288	1,230	0,857	0,605	1,285	1,767	3,155	4,169	0,625	**16,25**
A	**A15**	2,631	1,059	2,115	0,299	0,594	1,290	1,965	1,683	2,678	3,686	**18,00**
A	**A16**	1,638	1,431	1,190	0,852	0,991	1,928	1,830	0,984	1,007	2,898	**14,75**
A	**A17**	2,964	1,512	0,995	1,053	1,292	0,572	0,423	1,295	1,011	1,883	**13,00**
A	**A18**	0,616	2,335	0,754	0,796	0,909	0,744	0,246	0,940	1,183	1,729	**10,25**
A	**A19**	1,276	1,379	1,318	1,758	1,763	0,381	0,476	0,147	0,562	0,690	**9,75**
A	**A20**	0,905	0,259	0,655	0,514	0,365	0,138	0,524	0,457	0,345	0,590	**4,75**
A	**A21**	1,110	0,697	0,615	3,306	1,325	1,314	1,406	1,889	1,122	1,717	**14,50**
A	**A22**	0,735	1,062	1,085	0,713	0,911	1,332	0,655	0,775	2,010	1,221	**10,50**
A	**A23**	0,721	1,194	1,366	0,656	0,727	0,771	1,498	2,034	1,294	0,989	**11,25**
A	**A24**	1,458	1,494	0,654	1,227	0,951	0,715	0,957	1,068	0,573	1,404	**10,50**
B	**B01**	1,241	1,482	2,000	0,476	1,205	1,181	2,643	2,180	1,519	3,322	**17,25**
B	**B02**	0,447	0,264	0,596	0,310	1,784	0,336	0,458	2,918	0,815	1,072	**9,00**
B	**B03**	0,740	0,440	0,322	0,078	0,946	1,066	1,269	2,447	0,975	1,716	**10,00**
B	**B04**	1,141	0,317	0,574	1,146	0,210	1,653	1,376	1,080	0,940	1,563	**10,00**

Corpus	DOC	10%	20%	30%	40%	50%	60%	70%	80%	90%	100%	Total
B	**B05**	0,706	0,961	0,286	1,474	1,483	2,689	3,103	0,598	0,077	2,874	**14,25**
B	**B06**	0,608	0,656	0,251	0,290	0,696	0,430	0,260	0,532	0,664	0,613	**5,00**
B	**B07**	0,321	0,520	0,095	0,620	0,354	0,720	0,525	0,240	0,498	1,106	**5,00**
B	**B08**	1,538	0,121	0,152	1,051	0,415	0,654	0,204	0,074	1,452	2,090	**7,75**
B	**B09**	1,480	0,325	0,686	0,993	0,866	0,191	1,132	1,162	0,968	1,722	**9,53**
B	**B10**	0,298	0,376	0,114	0,157	0,325	0,179	1,333	2,701	3,598	1,965	**11,05**
B	**B11**	0,571	0,366	0,648	0,832	0,686	0,797	1,326	0,922	0,642	0,960	**7,75**
B	**B12**	0,555	1,702	0,778	0,234	1,607	2,913	2,685	1,312	0,914	1,184	**13,88**
B	**B13**	0,479	0,466	1,500	0,486	0,392	0,324	0,485	0,802	0,746	0,320	**6,00**
B	**B14**	0,207	0,445	0,399	0,092	0,013	0,318	0,900	0,317	0,821	0,487	**4,00**
B	**B15**	0,141	0,148	0,141	0,058	0,129	0,162	1,834	2,239	2,096	0,229	**7,17**
B	**B16**	0,479	0,357	0,250	0,660	0,729	0,586	1,412	2,050	4,837	1,639	**13,00**
C	**C01**	1,055	1,681	0,408	1,727	0,606	0,598	0,922	1,393	1,046	1,313	**10,75**
C	**C02**	1,113	0,886	1,295	0,917	0,989	0,916	1,435	1,164	1,212	0,822	**10,75**
C	**C03**	0,789	1,835	0,134	0,500	0,518	0,757	0,733	1,453	4,382	0,646	**11,75**
C	**C04**	1,334	0,404	0,244	1,349	0,630	0,668	0,716	0,624	1,034	1,747	**8,75**
C	**C05**	2,893	1,834	1,597	0,672	1,280	0,829	1,505	1,584	1,139	2,919	**16,25**
C	**C06**	0,689	0,430	0,091	0,595	1,355	0,669	0,182	0,932	0,238	0,319	**5,50**
C	**C07**	1,039	0,328	0,264	0,262	0,551	1,060	0,487	0,577	0,773	0,909	**6,25**
C	**C08**	1,479	0,858	1,479	0,710	2,868	2,179	1,044	1,863	1,680	1,341	**15,50**
C	**C09**	1,674	0,810	1,490	0,446	0,525	1,312	3,581	1,313	1,879	1,972	**15,00**
C	**C10**	1,869	2,619	2,609	1,212	1,664	0,716	0,685	0,544	1,023	1,060	**14,00**
C	**C11**	0,602	0,889	0,747	1,873	1,788	1,580	0,970	0,667	0,532	0,602	**10,25**
C	**C12**	1,838	2,008	1,114	1,483	1,175	1,140	0,477	1,146	0,720	0,917	**12,02**
C	**C13**	1,487	0,540	0,931	0,912	0,406	1,431	0,566	0,697	1,063	1,466	**9,50**
D	**D01**	0,606	0,283	0,227	0,830	0,657	1,713	5,890	2,838	0,751	0,704	**14,50**
D	**D02**	0,583	0,154	0,490	0,996	0,394	0,321	0,321	0,863	1,578	0,799	**6,50**
D	**D03**	0,580	0,490	0,595	0,926	0,931	0,722	0,248	0,867	0,991	1,150	**7,50**
D	**D04**	1,044	2,020	0,414	0,426	1,404	0,797	1,156	1,467	0,118	0,155	**9,00**
D	**D05**	0,955	1,300	0,372	0,045	0,535	1,999	1,767	3,302	1,836	1,889	**14,00**

Corpus	DOC	10%	20%	30%	40%	50%	60%	70%	80%	90%	100%	Total
D	**D06**	1,471	0,586	1,223	2,569	0,663	0,952	1,061	3,566	2,753	2,168	**17,01**
D	**D07**	0,526	0,433	0,893	0,000	1,020	1,292	1,421	1,851	1,171	0,000	**8,61**
D	**D08**	0,389	0,755	0,745	0,371	0,788	0,843	1,273	2,211	2,107	4,518	**14,00**
D	**D09**	0,879	0,435	0,000	1,105	0,763	1,025	0,591	1,148	0,000	1,346	**7,29**
D	**D10**	0,023	0,179	0,260	0,098	0,078	0,440	0,306	0,399	0,972	0,494	**3,25**
D	**D11**	0,273	0,186	0,091	0,341	0,083	0,286	0,059	0,305	0,036	0,340	**2,00**
D	**D12**	0,225	0,524	0,721	0,205	0,087	0,514	1,400	2,245	3,223	2,105	**11,25**
E	**E01**	2,033	1,098	0,179	1,040	1,092	2,447	1,345	1,391	0,565	1,059	**12,25**
E	**E02**	0,160	0,766	0,105	0,700	1,079	1,845	2,720	3,894	2,178	2,596	**16,04**
E	**E03**	0,000	0,038	0,155	0,376	1,171	1,747	1,729	1,329	2,500	2,471	**11,52**
E	**E04**	0,683	1,313	0,429	0,484	2,338	2,106	1,200	0,649	1,533	0,767	**11,50**
E	**E05**	1,553	0,477	0,125	0,556	0,333	1,391	2,403	0,040	1,550	0,322	**8,75**
E	**E06**	0,523	0,167	0,000	1,089	0,695	1,536	4,337	2,997	2,635	2,020	**16,00**
E	**E07**	0,103	0,192	0,331	0,791	1,089	1,555	1,439	0,532	0,580	0,905	**7,52**
E	**E08**	0,544	0,425	0,511	0,111	0,361	0,343	0,188	0,174	0,343	0,000	**3,00**
E	**E09**	1,538	0,783	0,750	0,779	0,263	0,480	1,078	0,633	0,601	1,095	**8,00**
E	**E10**	0,354	0,588	0,338	0,829	1,185	1,190	1,340	1,344	0,743	1,588	**9,50**
F	**F01**	0,000	0,667	0,500	0,500	0,167	0,455	0,542	0,460	0,199	2,011	**5,50**
F	**F02**	0,179	0,650	0,233	0,422	1,157	0,980	1,145	0,596	0,955	0,977	**7,29**
F	**F03**	0,012	0,639	0,239	0,889	0,889	0,996	1,614	1,626	1,938	1,408	**10,25**
F	**F04**	0,624	0,242	0,296	1,518	3,178	1,908	2,180	1,882	0,662	0,512	**13,00**
F	**F05**	0,306	0,189	0,457	0,246	0,291	0,582	2,323	3,335	1,396	0,755	**9,88**
F	**F06**	0,313	0,504	0,617	0,437	0,206	0,387	3,986	3,391	1,249	1,751	**12,84**
F	**F07**	0,475	1,440	0,617	0,570	0,381	0,461	1,054	1,001	1,452	1,800	**9,25**
F	**F08**	1,690	0,460	0,898	0,998	1,008	0,646	0,274	1,587	0,632	0,808	**9,00**
G	**G01**	1,326	0,525	0,209	1,015	1,261	3,209	2,115	2,293	1,831	3,465	**17,25**
G	**G02**	1,673	0,070	0,629	0,767	0,515	0,976	0,857	1,145	0,492	0,626	**7,75**
G	**G03**	0,952	1,910	0,637	1,230	0,951	0,250	1,074	0,969	1,454	1,322	**10,75**
G	**G04**	0,250	0,543	1,074	0,969	2,013	1,666	1,756	1,539	2,913	2,277	**15,00**
G	**G05**	1,308	0,276	1,001	1,595	1,034	0,554	2,756	0,202	1,659	0,865	**11,25**

Corpus	DOC	10%	20%	30%	40%	50%	60%	70%	80%	90%	100%	Total
G	**G06**	1,514	1,809	1,264	0,242	0,244	0,142	0,327	0,300	0,560	1,849	**8,25**
G	**G07**	0,499	0,000	0,571	0,114	0,313	0,561	0,497	0,525	1,019	1,900	**6,00**
G	**G08**	0,691	0,899	2,519	0,644	0,574	0,454	1,691	0,260	0,430	3,087	**11,25**
H	**H01**	1,956	0,614	0,988	0,913	1,145	1,709	0,997	1,416	1,667	2,842	**14,25**
H	**H02**	0,916	1,137	0,750	2,139	0,414	1,756	2,594	1,796	4,685	2,312	**18,50**
H	**H03**	1,605	0,806	1,211	1,059	1,495	2,368	0,676	1,108	1,408	1,013	**12,75**
H	**H04**	1,053	0,525	2,624	2,569	0,833	1,347	1,270	1,329	1,069	1,895	**14,51**
H	**H05**	0,000	0,425	0,677	0,692	0,677	6,446	8,364	1,257	1,001	0,604	**20,14**
H	**H06**	0,442	0,483	0,504	0,174	0,194	0,308	1,604	0,856	1,290	1,211	**7,07**
H	**H07**	0,641	1,402	0,587	1,140	2,206	0,798	0,539	0,911	1,736	0,790	**10,75**

Fonte: Elaborado pelo autor.

APÊNDICE L - MÉDIA DA ATRIBUIÇÃO DE VALOR DE RELEVÂNCIA PARA OS SINTAGMAS NOMINAIS NAS PARTES ESTRUTURAIS DE CADA TESE DO *CORPUS*

Corpus	DOC	Introdução*	Desenvolvimento*	Conclusão*	Total*
A	**A01**	0,590	0,405	1,210	**2,204**
A	**A02**	0,829	0,949	1,973	**3,751**
A	**A03**	1,374	1,487	2,497	**5,358**
A	**A04**	1,590	2,152	3,442	**7,184**
A	**A05**	2,123	5,547	4,360	**12,030**
A	**A06**	0,087	0,291	0,372	**0,749**
A	**A07**	0,437	1,167	0,642	**2,246**
A	**A08**	1,906	0,352	0,767	**3,025**
A	**A09**	1,697	1,117	1,877	**4,691**
A	**A10**	0,479	0,348	0,665	**1,492**
A	**A11**	0,724	0,131	1,081	**1,936**
A	**A12**	0,543	0,556	1,297	**2,396**
A	**A13**	0,689	0,673	0,897	**2,258**
A	**A14**	1,879	2,931	0,571	**5,381**
A	**A15**	1,321	0,741	1,322	**3,384**
A	**A16**	1,362	0,827	1,475	**3,665**
A	**A17**	3,016	0,604	0,849	**4,470**
A	**A18**	1,214	0,769	1,678	**3,662**
A	**A19**	1,264	0,593	0,634	**2,492**
A	**A20**	0,418	0,195	0,215	**0,829**
A	**A21**	0,683	0,821	1,442	**2,945**
A	**A22**	0,896	0,633	1,022	**2,551**
A	**A23**	2,487	1,313	1,048	**4,847**
A	**A24**	1,008	0,633	1,052	**2,694**
B	**B01**	3,425	3,775	7,752	**14,952**
B	**B02**	0,379	1,295	1,454	**3,128**
B	**B03**	0,598	2,177	4,621	**7,396**
B	**B04**	6,500	1,646	0,339	**8,485**

Corpus	DOC	Introdução*	Desenvolvimento*	Conclusão*	Total*
B	**B05**	0,735	2,763	0,718	**4,217**
B	**B06**	2,514	1,125	4,489	**8,128**
B	**B07**	1,261	0,647	2,198	**4,106**
B	**B08**	8,845	2,648	23,548	**35,041**
B	**B09**	6,460	2,292	5,870	**14,622**
B	**B10**	0,343	1,636	0,248	**2,227**
B	**B11**	1,378	0,747	2,351	**4,476**
B	**B12**	0,978	1,838	1,915	**4,731**
B	**B13**	0,473	1,465	0,000	**1,939**
B	**B14**	1,819	0,959	0,000	**2,779**
B	**B15**	0,601	1,552	1,054	**3,207**
B	**B16**	1,756	0,707	4,028	**6,490**
C	**C01**	0,675	0,586	1,034	**2,295**
C	**C02**	0,902	0,764	0,578	**2,244**
C	**C03**	0,067	0,662	0,352	**1,081**
C	**C04**	0,526	0,551	1,124	**2,201**
C	**C05**	1,867	0,852	2,616	**5,335**
C	**C06**	0,175	0,299	0,043	**0,516**
C	**C07**	0,814	0,240	0,935	**1,989**
C	**C08**	1,432	1,240	1,717	**4,389**
C	**C09**	1,152	1,235	3,674	**6,061**
C	**C10**	1,271	0,936	0,576	**2,783**
C	**C11**	0,826	0,776	0,315	**1,918**
C	**C12**	0,987	0,525	1,170	**2,682**
C	**C13**	0,971	0,924	1,406	**3,301**
D	**D01**	0,302	1,538	0,495	**2,335**
D	**D02**	1,602	1,205	2,124	**4,930**
D	**D03**	1,860	1,412	2,192	**5,464**
D	**D04**	0,847	0,999	0,074	**1,921**
D	**D05**	3,031	2,557	5,597	**11,185**
D	**D06**	2,038	1,664	3,473	**7,175**

Corpus	DOC	Introdução*	Desenvolvimento*	Conclusão*	Total*
D	**D07**	0,648	2,000	1,120	**3,768**
D	**D08**	1,099	2,573	5,090	**8,762**
D	**D09**	2,161	1,064	3,562	**6,786**
D	**D10**	0,096	0,774	0,248	**1,119**
D	**D11**	0,807	0,407	1,357	**2,571**
D	**D12**	0,510	1,461	2,564	**4,534**
E	**E01**	2,322	1,592	0,945	**4,859**
E	**E02**	0,633	3,650	2,662	**6,946**
E	**E03**	0,072	2,077	0,000	**2,150**
E	**E04**	3,137	3,941	1,121	**8,199**
E	**E05**	3,273	2,614	3,199	**9,086**
E	**E06**	0,391	3,544	0,674	**4,609**
E	**E07**	0,130	0,832	0,986	**1,948**
E	**E08**	0,729	0,425	0,000	**1,154**
E	**E09**	2,251	0,838	1,683	**4,773**
E	**E10**	0,349	1,027	1,578	**2,955**
F	**F01**	0,812	1,779	0,000	**2,591**
F	**F02**	0,507	1,373	0,905	**2,785**
F	**F03**	0,769	3,125	1,503	**5,398**
F	**F04**	0,654	2,359	2,101	**5,114**
F	**F05**	0,418	1,759	1,016	**3,193**
F	**F06**	1,059	3,099	3,545	**7,703**
F	**F07**	2,328	2,397	6,228	**10,953**
F	**F08**	1,091	0,671	0,860	**2,622**
G	**G01**	1,337	1,562	3,177	**6,076**
G	**G02**	1,921	0,610	0,423	**2,954**
G	**G03**	1,962	1,330	1,185	**4,476**
G	**G04**	0,310	1,136	1,483	**2,929**
G	**G05**	2,051	1,436	1,826	**5,312**
G	**G06**	1,003	0,424	1,062	**2,490**
G	**G07**	0,968	0,547	0,756	**2,270**

Corpus	DOC	Introdução*	Desenvolvimento*	Conclusão*	Total*
G	**G08**	0,758	0,599	2,763	**4,120**
H	**H01**	3,026	1,670	5,186	**9,882**
H	**H02**	1,475	2,638	7,280	**11,393**
H	**H03**	3,039	2,296	2,902	**8,237**
H	**H04**	0,380	0,815	0,853	**2,048**
H	**H05**	0,000	7,217	5,431	**12,647**
H	**H06**	0,392	0,566	2,353	**3,311**
H	**H07**	2,591	4,570	20,476	**27,638**

Fonte: Elaborado pelo autor.

*** Valores foram multiplicados por 1.000**