XIV Encontro Nacional de Pesquisa em Ciência da Informação (ENANCIB 2013)
GT 2 - Organização e Representação do Conhecimento

Comunicação Oral

CARACTERIZAÇÃO DE TESES DE OITO ÁREAS DE CONHECIMENTO: UMA ANÁLISE PARA O DESEMPENHO DE INDEXAÇÃO AUTOMÁTICA ATRAVÉS DE SINTAGMAS NOMINAIS

Luiz Antônio Lopes Mesquita – Faculdade Estácio de Sá – Faculdade Pitágoras
Renato Rocha Souza - UFMG
Renata Maria Abrantes Baracho Porto - UFMG

Resumo
O objetivo principal desta pesquisa é analisar características linguísticas quantitativas que diferenciam teses de doutorado e que podem influenciar no desempenho da etapa de extração de sintagmas nominais para a sua indexação automática. As características analisadas aqui são relativas a dimensões de grandeza, comportamento linguístico e estrutura do texto. A estrutura do texto considerada foi relativa às 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 de oito áreas de conhecimento de uma mesma universidade. Todos os textos apresentaram comportamentos característicos quando estavam relacionados às ciências naturais ou às ciências sociais. Aqueles relativos às ciências naturais apresentaram menor grandeza, favorecendo assim um melhor desempenho para processadores de indexação automática. Já o comportamento linguístico constatado como mais próximo da linguagem natural, presente sobretudo nas ciências sociais, contribui para o melhor desempenho na indexação automática por gerar menor quantidade de erros de extração de sintagmas nominais. Os textos relativos aos programas de Engenharia Metalúrgica e de Ciência da Informação apresentaram as menores estruturas de introdução e conclusão, fatores que auxiliam no desempenho de processos de indexação automática.
Palavras-chave: Linguística Computacional. Processamento de Linguagem Natural. Indexação automática. Indexação automática por extração. Sintagmas Nominais. Estrutura de texto.

Abstract
The main objective of this research is to analyze quantitative linguistic features that differentiate doctoral theses and that can influence the performance of the step of extracting noun phrases to their automatic indexing. The traits analyzed here are related to dimensions of magnitude, linguistic behavior and structure of the text. The structure of the text was considered relative to their structural parts (introduction, development and conclusion). The terms considered here were only full noun phrases contained in the texts themselves. The texts were considered a total of 98 doctoral theses eight knowledge areas of the same university. All texts showed characteristic behaviors when they were related to the natural sciences or social sciences. Those related to the natural sciences had lower magnitude, thus fostering a better processor performance of automatic indexing. Already linguistic behavior as observed from those of less specialized social sciences contributes to better performance in automatic indexing to generate fewer errors extracting noun phrases. The texts concerning programs Metallurgical Engineering and Information Science presented the smallest structures

introduction and conclusion, factors that assist in the performance of automatic indexing processes.
Keywords: Computational Linguistics. Natural Language Processing. Automatic Indexing. Automatic Indexing For Extracting. Noun Phrases. Text Structure.

1 INTRODUÇÃO

Algoritmos cada vez mais otimizados e processadores cada vez mais rápidos estão permitindo que as pesquisas com indexadores automáticos possam utilizar estruturas linguísticas cada vez mais complexas: uma delas é o sintagma nominal. Tal estrutura, que possui maior valor semântico que a palavra isolada (PERINI *et al.*, 1996), 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 sintagmas nominais como descritores relevantes no processo de indexação automática. A metodologia de Souza foi utilizada por Maia (2008) para o desenvolvimento de uma ferramenta[1] que, dentre outras funcionalidades, extrai tais sintagmas nominais de forma automática.

A utilização do sintagma nominal é responsável por uma significativa evolução nos sistemas usados para a indexação automática atualmente, no entanto a grande maioria desses sistemas é baseada na língua inglesa. A língua portuguesa possui substanciais diferenças em relação ao inglês, o que coloca obstáculos para que tais ferramentas sejam facilmente adaptadas para nossa língua. Logo, faz-se necessária a criação de conhecimento, não apenas sobre, mas para a língua portuguesa para o uso de tais ferramentas.

O uso dos sintagmas nominais em um texto em português pode permitir chegar a métodos de escolha automática de descritores que sejam mais relevantes do que simplesmente o uso de palavras isoladas. Tais métodos têm em comum a extração desses sintagmas nominais como etapa anterior à escolha dos mesmos como descritores. Os resultados dessas extrações permitem caracterizar de antemão seus respectivos textos em relação a dimensões de grandeza, diversidade do uso da língua e estilos de estrutura, por exemplo, que influenciam no desempenho desse processo de extração.

Considerando-se a indexação automática de extensas bases digitais de documentos, torna-se relevante analisar as características de seus textos que podem influenciar no desempenho dessa etapa de extração que consome significante parcela do custo computacional de todo o processo de indexação automática de um conjunto de documentos.

[1] 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 Palavras (BICK, 2000), que é fruto de uma tese de doutorado para a análise automática gramatical da língua portuguesa.

Visando contribuir para minimizar esse custo computacional, o objetivo desta pesquisa é analisar características linguísticas quantitativas que diferenciam as teses de doutorado que podem influenciar no desempenho da etapa de extração de sintagmas nominais para a sua indexação automática. As características analisadas aqui são relativas a dimensões de grandeza, comportamento linguístico e estrutura do texto. Os termos considerados aqui são somente sintagmas nominais contidos nos próprios textos. Os textos considerados aqui são teses de doutorado das oito áreas de conhecimento de uma mesma universidade.

2 CONCEITOS GERAIS E REVISÃO DA LITERATURA

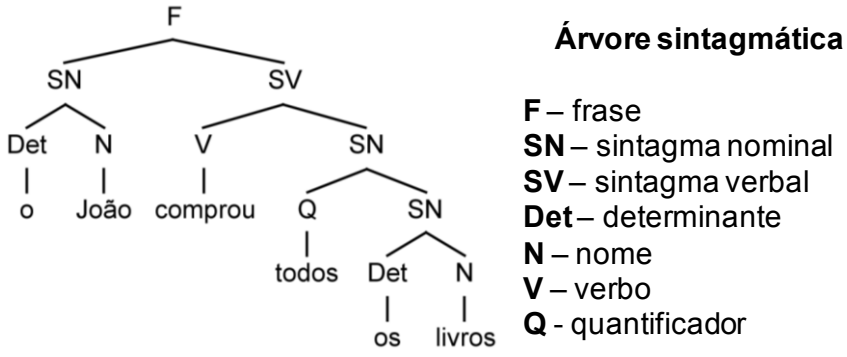
Em todas as partes do texto ocorrem expressões que dependem do contexto para a determinação de seu significado. Essas expressões são denominadas referenciais (LYONS, 1987). Como apresentado adiante, para a indexação automática, a frequência de um termo é usada como peso para determinar a sua relevância como seu descritor. Um problema que as expressões referenciais 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.

Sintagma nominal (*noun phrase*, *NP*) – SN – 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, p. 270).

Abaixo, temos um exemplo de sintagma nominal. É possível observar que existe a estrutura chamada de sintagma nominal aninhado. Na

Figura 1 a seguir o termo “todos os livros” possui tal estrutura, pois ele é composto por outro sintagma nominal (“os livros”) aninhado dentro dele.

Figura 1 - Exemplo de estrutura de um sintagma nominal.



Fonte: Adaptado de Othero (2009).

Os sintagmas nominais em um documento apresentam densidade informacional superior às palavras isoladas, mantendo maior proximidade com o 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). Por sua vez, os sintagmas nominais trazem "em seu bojo o contexto semântico dos discursos" (SOUZA, 2005, p. 136), o que possibilita que tais problemas ocorram menos. Para Baeza-Yates e Ribeiro-Neto (2011, p. 224) os substantivos (que compõem um sintagma nominal) possuem maior valor semântico ao serem usados como termos de indexação. Portanto, o uso de sintagmas nominais como termos de indexação pode apresentar melhores resultados que o uso de palavras isoladas.

Os sintagmas nominais podem ser extraídos automaticamente de textos. Os trabalhos de Kuramoto (1995), Souza (2005), Maia (2008), Corrêa (2011), Mesquita (2012) e outros apresentam como tema central a utilização de sintagmas nominais 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.

Baeza-Yates e Ribeiro-Neto (2011) apresentam que um documento pode ser pré-processado seguindo cinco operações: a primeira consiste na denominada análise léxica, que consiste no tratamento de acentuações (*accents*), espaços *(spacing*), marcas de pontuação, números, hífens etc.; em seguida 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*), outra operação utiliza os sintagmas nominais (*noun groups* ou *noun phrases*) 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.); em seguida o *stemming* consiste na transformação de uma palavra para a sua raiz. Uma técnica para isso consiste na retirada de prefixos e sufixos; e finalmente os termos restantes são eleitos como descritores através de um processo que pode ser automático ou manual.

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 ao 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. 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" (BORKO e BERNIER, 1978, p.8).

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, p. 32), 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 ao crescente volume de documentos eletrônicos e de forma mais consistente que a manual. A questão mais recorrente nos critérios de seleção de descritores é aquela que pode ser considerada como essencial para a indexação automática: o uso de estratégias e técnicas baseadas em cálculos, estatísticas e probabilidades.

3 METODOLOGIA

É apresentado aqui em detalhes o método utilizado de seleção, obtenção e tratamento do *corpus* de teses de doutorado, assim como o processo para a extração dos sintagmas nominais.

Seleção, obtenção e tratamento do *corpus*

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 da UFMG.

Para uma tese, que possui aproximadamente entre cem e quatrocentas páginas relacionadas a uma área de estudos (ECO, 2007, p. 27), acredita-se aqui que essa ordem de grandeza textual pode favorecer o estudo da extração dos sintagmas nominais como descritores. Essa hipótese é baseada nos seguintes aspectos: as repetições de um mesmo sintagma nominal 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 as características de cada texto.

A escolha aqui de teses como elementos de pesquisa implica em maior **custo computacional** de processamento da extração dos sintagmas nominais, em comparação a artigos, uma vez que estes últimos, geralmente, possuem um tamanho da ordem de dez vezes menor (MESQUITA, 2012). No entanto, com o desempenho dos recursos computacionais atuais em relação aos mais antigos[2] usados em outras pesquisas, que se basearam em artigos, o processamento de teses mostrou-se viável (cerca de 12 horas para 98 teses), como pode ser visto adiante na análise de resultados.

Inicialmente foram levantadas todas as quantidades de teses na Biblioteca Digital da UFMG, encontrando-se 1.921 referências pertencentes a 54 programas de pós-graduação (outros 13 programas só apresentaram dissertações de mestrado).

Para atingir um maior grau de representatividade e um menor erro amostral, foi utilizada uma amostragem estratificada, ou seja, os elementos de pesquisa (teses) foram agrupados de modo a representar sua heterogeneidade (BABBIE, 1999, p. 137), 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 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 os 54 distintos programas e eleger aqueles que possuíssem mais teses dentro da sua área de conhecimento.

A equação utilizada para determinar o tamanho da amostra para uma proporção (n) foi "$\mathbf{n = Z^2 p(1\text{-}p)/e^2}$" (LEVINE *et al,* 2000, p. 301). Admitiu-se aqui o nível de confiança (relativo a Z) como 90%, a verdadeira proporção (relativo a *p*) como a proporção para todas as teses e o nível de erro de amostragem (relativo a *e*) como 10%.

Para cada programa de pós-graduação, foram selecionadas teses que foram disponibilizadas na Biblioteca Digital da UFMG mais recentemente. O **recorte temporal** aqui, que faz parte de qualquer processo de amostragem (BABBIE, 1999, p. 114), é importante pois existe a possibilidade de variações de comportamentos linguísticos ao longo das gerações de autores que podem influenciar na análise de dados. Portanto foi utilizada uma amostragem sistemática iniciando-se da publicação mais recente em direção à mais antiga.

[2] 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.

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.

Cada tese foi obtida a partir da Biblioteca Digital da UFMG no formato PDF[3]. 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[4];
3. Foram eliminados espaços em branco consecutivos;
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;
5. Foram inseridos demarcadores 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*.txt*"*.

Extração dos Sintagmas Nominais

Para cada texto, foram obtidos seus sintagmas nominais e apresentados, um em cada linha, em um novo texto. Considerou-se aqui cada sintagma nominal máximo, desconsiderando-se os sintagmas nominais aninhados, ou seja, aqueles que são sintagmas nominais, porém fazem parte de um sintagma nominal maior (máximo). Essa escolha deve-se ao fato de a ferramenta Ogma fornecer a listagem sequencial de sintagmas somente nesse formato.

A ferramenta Ogma 0.10[5] e o *software* Microsoft Office Word 2007 foram utilizados para a extração dos sintagmas nominais 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

[3] O PDF é um padrão aberto de arquivo (*Portable Document Format*) desenvolvido pela *Adobe Systems*.
[4] Os textos contidos em formatos digitais não textuais, tais como em imagens ou figuras, também foram descartados.
[5] 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.

intermediária para a extração dos sintagmas nominais. 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).

2. Extração dos sintagmas nominais: 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 sintagmas nominais 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 sintagmas nominais extraídos é o mesmo do original acrescido de "-s". Exemplo: ogma s a01-e.txt a01-s.txt).
3. Limpeza dos sintagmas nominais: a partir de cada listagem de sintagmas nominais foi realizado um procedimento para a melhoria dos resultados baseado na elaboração pelo autor de macros de aplicação[6] dentro do Microsoft Office Word 2007 (o nome do arquivo gerado com a sequência de sintagmas nominais extraídos já limpos é o mesmo do original acrescido de "-sl". Exemplo: a01-sl.txt). A limpeza dos sintagmas nominais considerou os seguintes resultados encontrados a partir do Ogma:
    - Alguns sintagmas nominais 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 sintagmas nominais extraídos pelo Ogma foram números puros (como aqueles decorrentes das numerações de páginas) ou até mesmo compostos somente por *stopwords*.

Ao final desses procedimentos descritos, para cada tese obteve-se a listagem final de todos os sintagmas nominais já com os procedimentos de limpeza aplicados (arquivos com a seguinte sintaxe "*ann*-sl.txt").

4 ANÁLISE DOS RESULTADOS

A metodologia descrita no capítulo anterior e aplicada nesta pesquisa teve como principal pressuposto avaliar a diferença de comportamento linguístico entre os oito programas de pós-graduação, tais como: proporção entre início/desenvolvimento/conclusão, quantidade média de sintagmas nominais por tese (e seu consequente tamanho numérico médio de palavras) e seus aspectos relacionados ao desempenho da extração.

[6] As *macros* de aplicação consistem na automatização da execução de funções.

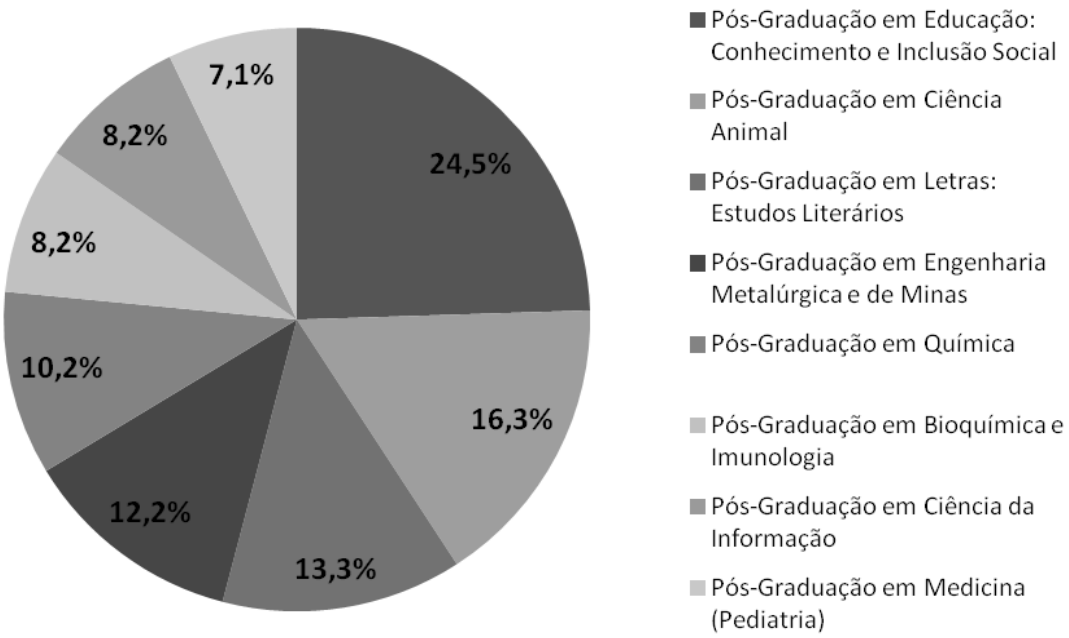
O *corpus* foi constituído de oito seções, sendo que cada uma delas representou uma das oito áreas de conhecimento da UFMG. O total de teses analisadas foi noventa e oito, distribuídas para cada programa de pós-graduação conforme a Tabela 1 e o

Gráfico 1 a seguir:

Tabela 1 - 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%

Gráfico 1 - Quantidade de teses analisadas por programa de pós-graduação.



O período de publicação de todas as teses analisadas corresponde a aproximadamente 4,5 anos (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[7] entre as publicações na Biblioteca Digital de Teses e Dissertações da UFMG – BDTD/UFMG para cada programa foi de 2,5 meses, conforme a Tabela 2 seguir:

Tabela 2 - Datas de publicação das teses analisadas na BDTD da UFMG.

Seção do *corpus*	Publicação da Tese no BDTD da 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).

[7] Para alguns programas, algumas teses dentro do período não foram analisadas por não estarem disponíveis integralmente na BDTD/UFMG.

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.

Análise da extração dos sintagmas nominais no *corpus*

Para a extração dos sintagmas nominais foram realizados, como descrito anteriormente, os processos de: escolha das teses, 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 sintagmas nominais, foram utilizadas as ferramentas Ogma, *macros* no Microsoft Word e *macros* no Microsoft Excel, como também descrito anteriormente. Na Tabela 3 a seguir é possível verificar que a média de tempo para a extração foi de aproximadamente 9 horas e 52 minutos (83% do tempo total). O tratamento dos sintagmas nominais através de macros do Word criadas pelo autor durou cerca 2 horas (17% do tempo total).

Tabela 3 - Tempo de processamento para extração dos sintagmas nominais.

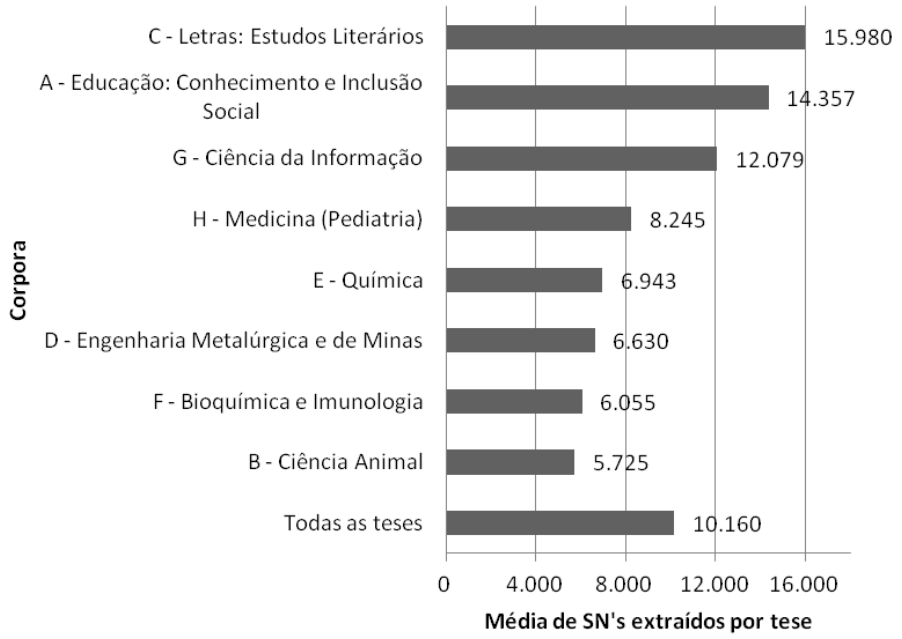
Grupos	**A**	**B**	**C**	**D**	**E**	**F**	**G**	**H**	**Total**	**Total (%)**
Processamento do Ogma	03:32	00:53	02:14	00:36	00:58	00:25	00:50	00:24	09:52	83,15%
Processamento de Macro do Word	00:30	00:13	00:25	00:14	00:09	00:11	00:13	00:05	02:00	16,85%
Tempo Total	**04:02**	**01:06**	**02:39**	**00:50**	**01:07**	**00:36**	**01:03**	**00:29**	**11:52**	**100,00%**
Quantidade de Teses (unid.)	24	16	13	12	10	8	8	7	98	
Quantidade de SN's extraídos	344.576	207.746	96.631	91.599	79.560	69.429	57.714	48.436	995.691	
Média de tempo por tese (hora:min.)	**00:10**	**00:04**	**00:12**	**00:04**	**00:06**	**00:04**	**00:07**	**00:04**	**00:07**	
Média de tempo por 1.000 sintagmas nominais extraídos (min.:seg.)	**00:42**	**00:19**	**01:39**	**00:33**	**00:51**	**00:31**	**01:05**	**00:36**	**00:43**	

A média de tempo de processamento para a extração dos sintagmas nominais foi de sete 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 43

(quarenta e três) segundos para cada 1.000 (mil) extrações, conforme pode ser visto anteriormente na Tabela 3.

As seções do *corpus* que apresentaram maiores médias de tempo por tese, apresentadas na Tabela 3, também foram aquelas que apresentaram as maiores médias de sintagmas nominais extraídos por tese, conforme pode ser visto no Gráfico 2 a seguir:

Gráfico 2 - Média de sintagmas nominais extraídos por tese em cada seção do corpus.



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[8] (SANTOS, 1996), pôde-se perceber, no Gráfico 2, que nas seções do *corpus* de programas de pós-graduação mais relacionados às Ciências sociais houve uma quantidade acima da média de sintagmas nominais extraídos, assim como, em todas as seções do *corpus* relacionadas às Ciências naturais, essa quantidade foi abaixo da média. Para Dubois *et al* (1973, p. 247) há uma concepção distinta de estruturas 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. Tal estabilidade é considerada aqui como fator primordial para a constatação da maior objetividade das teses relacionadas às Ciências naturais considerando-se o seu menor uso em quantidade de sintagmas nominais.

[8] 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).

Em relação à quantidade de sintagmas nominais, dentre as principais pesquisas referenciadas aqui e que realizaram extração de sintagmas nominais 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 sintagmas nominais extraídos entre todas essas pesquisas conforme Tabela 4 a seguir:

Tabela 4 - Comparação de extração de sintagmas nominais entre pesquisas.

Pesquisa	Quantidade de Documentos	Tipo de Documentos	Modo de Extração	Sintagmas Nominais extraídos	Média de Sintagmas Nominais 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 e outros (2011)	30	resumos de teses e dissertações	automática	**951**	**32**
Esta pesquisa	98	teses	automática	**995.691**	**10.160**

A quantidade de sintagmas nominais extraídos nesta pesquisa corresponde a aproximadamente 6,5 vezes mais que a maior quantidade observada nas demais pesquisas. Esse fato é devido ao tipo de documento escolhido (tese). Assim como em outras pesquisas, durante a extração de sintagmas nominais, ocorreram extrações automáticas que não resultaram propriamente em sintagmas nominais devido a falhas nos processos de extração. Corrêa *et al*. (2011) explicitaram uma taxa de erros de extração através do Ogma de 42% (ibidem, p. 18). Devido à pequena quantidade de sintagmas nominais extraídos em tal pesquisa, os autores puderam constatar manualmente a efetividade de cada resultado da extração.

Para esta pesquisa, os erros puderam ser contatados de forma automática através da retirada de *stopwords* residuais com o uso de *macros* do Microsoft Word, usando-se para isso *macros* do Microsoft Excel, também desenvolvidas pelo autor.

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 5 a seguir:

Tabela 5 - 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%**

Uma análise manual em cada um dos sintagmas nominais 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 sintagmas nominais 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.

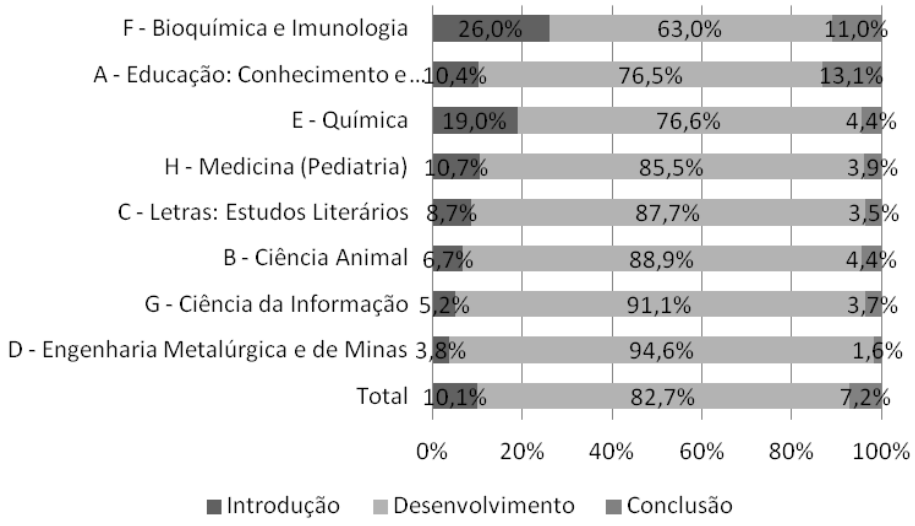
A seção do *corpus* que apresentou maior taxa de erros foi a correspondente ao programa de pós-graduação em Química, que possui como característica de seu sistema linguístico o uso de fórmulas químicas. No entanto, os fatores que influenciaram a sua elevada taxa de erros aqui foram: a elevada presença 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 pelos autores 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, aquele que usa a própria língua como

objeto de seu discurso (DUBOIS *et al*, 1973, p. 471), 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 de caracterização de teses de doutorado, foi considerada para cada sintagma nominal extraído a sua posição estrutural correspondente às partes de introdução, desenvolvimento e conclusão. Dentre essas, a de desenvolvimento conteve 82,7% dos sintagmas nominais, 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 3 a seguir:

Gráfico 3 - Distribuição de sintagmas nominais por partes da tese.



A maior distribuição de sintagmas nominais nas partes de introdução e conclusão ocorreu no programa de pós-graduação em Bioquímica e Imunologia, enquanto o programa que concentrou mais sintagmas nominais 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. 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.

Foi possível também concluir aqui que um mesmo sintagma nominal ocorre, em média, aproximadamente duas vezes em uma mesma tese. O total de sintagmas nominais identificados em cada tese correspondeu a 53,5% do total dos que foram extraídos. Ou seja, esse valor corresponde à quantidade de sintagmas nominais que são distintos entre si frente ao

total extraído. A Tabela 6 a seguir apresenta um detalhamento desses dados por seção do *corpus*.

Tabela 6 - 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%**

A respeito da relação entre a quantidade de sintagmas nominais identificados e o total de extraídos, Kuramoto (1999) obteve manualmente 8.818 destes e identificou 75,2% deles como sem repetições (ibidem, p. 65, calculado pelo autor). Souza (2005), assim como Kuramoto, utilizou artigos da Ciência da Informação e extraiu automaticamente 76.739 sintagmas nominais, sendo que 78,9% destes eram únicos (ibidem, p. 127, calculado pelo autor). 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 sintagmas nominais deriva da quantidade de palavras de uma língua, que é limitada sincronicamente[9]. 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". 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.

[9] 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.

Novamente, pôde ser observada uma maior singularidade na seção do *corpus* correspondente ao programa de pós-graduação em Letras – Estudos Literários, cuja porcentagem de sintagmas nominais 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 sintagmas nominais 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.

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 sintagma nominal. Foi considerada a seguinte hipótese para a causa deste fato: em tal comunidade ocorreria um uso do sistema linguístico mais especializado e mais controlado que os outros. Ou seja, foi considerado 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[10], denominado também por "*Gold Book"*, adotado internacionalmente e disponibilizado livremente pela IUPAC - *International Union of Pure and Applied Chemistry*. 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 sintagmas nominais dentre os demais programas de pós-graduação.

Dentre esses sintagmas nominais 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.

Embora a média de repetição de um mesmo sintagma nominal tenha sido apresentada aqui como aproximadamente duas, foi possível perceber que somente um quinto dos sintagmas nominais identificados ocorre mais de uma vez ao longo de uma tese (19,4%). Foi

[10] IUPAC - International Union of Pure and Applied Chemistry. Compendium of Chemical Terminology. Gold Book. Disponível em: <http://goldbook.iupac.org/PDF/goldbook.pdf>.

possível também comprovar o comportamento da distribuição de frequências de acordo com a Lei de Zipf[11] (BAEZA-YATES; RIBEIRO-NETO; 2011, p. 221).

A seção do *corpus* do programa de pós-graduação em Letras – Estudos Literários apresentou a maior média de sintagmas nominais 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, 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 sintagmas nominais com mais de uma ocorrência, assim como o maior índice de repetições de um mesmo sintagma nominal (2,0%). Esse fato pode estar, mais uma vez, relacionado ao uso do que se assemelha a um vocabulário controlado internacionalmente (*Gold Book*, divulgado pela IUPAC).

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 8 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.

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.

Os programas que apresentaram menor quantidade de sintagmas nominais na introdução e na conclusão 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.

Mesmo adotando teses de doutorado como documentos, o tempo total de processamento chegou a ser menor que em outras pesquisas. Podemos concluir que, com o crescente avanço de recursos de processamento as pesquisas de indexação automática podem tender a adotar documentos cada vez maiores, assim como coleções também cada vez maiores.

---

[11] 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.

Programas que possuem uma linguagem mais especializada, como no caso da Química, que utiliza um vocabulário controlado da língua inglesa e apresentou a maior média de exclusões, necessitam de processadores mais especialistas que o Ogma. É recomendável também 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.

O comportamento distinto entre as teses relativas às ciências naturais e aquelas relativas às ciências sociais abre espaço para novas análises. 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.

REFERÊNCIAS

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.

BDTD/UFMG - BIBLIOTECA DIGITAL DA UFMG. Disponível em: < http://www.bibliotecadigital.ufmg.br/dspace/browse-date>. Acesso em novembro de 2011.

BICK, E. *The Parsing System Palavras:* Automatic Grammatical Analysis of Portuguese in a Constraint Grammar Famework. Aarhus: Aarhus University Press, 2000.

BORKO, Harold. Toward a theory of indexing. Information Processing and Management, v. 13, p. 355-365, 1977.

BORKO, H.; BERNIER, C. **Indexing concepts and methods**. New York: Academic Press. 1978.

DIAS, Eduardo Wense; NAVES, Madalena Martins Lopes. **Análise de assunto**: teoria e prática. Brasília: Thesaurus, 2007. 116p.

DUBOIS, J.; GIACOMO, M.; GUESPIN, L.; MARCELLESI, C.; MARCELLESSI, J.; MEVEL, J.. Dicionário de lingüística. São Paulo: Cultrix, 1973. 657p.

ECO, U. *Como se faz uma tese em ciências humanas*. 13ª Ed. Lisboa - Presença. 2007. 238 p.

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

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.

LEVINE, D. M.;BERENSON, M. L.; STEPHAN, David. Estatística: Teoria e Aplicações usando Microsoft Excel em Português. Rio de Janeiro: LTC, 2000.

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

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

MESQUITA, L. A. L. *SINTAGMAS NOMINAIS NA INDEXAÇÃO AUTOMÁTICA: uma análise estrutural da distribuição de termos relevantes em teses de doutorado da UFMG.* Dissertação de Mestrado. Orientador Prof. Dr. Renato Rocha Souza. UFMG, ECI, 2012.

OTHERO, G. A. A gramática da frase em português: algumas reflexões para a formalização da estrutura frasal em português. Dados eletrônicos. 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, Julho / Dezembro 1996. p. 43-56.

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

SOUZA, R. R. *Uma proposta de metodologia para escolha automática de descritores utilizando sintagmas nominais*. Tese de Doutorado. Orientadora Profª. Dr. Lidia Alvarenga. UFMG, ECI, 2005.

TRASK, R. L. *Dicionário de Linguagem e Lingüística*. Tradução e adaptação de Rodolfo Ilari. Revisão Técnica de Ingedore Villaça Koch e Thaís Cristófaro Silva. São Paulo: Contexto. 2004. 364 p. ISBN 85-7244-254-5.