UNIVERSIDADE FEDERAL DE PERNAMBUCO

CENTRO DE ARTES E COMUNICAÇÃO

DEPARTAMENTO DE CIÊNCIA DA INFORMAÇÃO

PROGRAMA DE PÓS-GRADUAÇÃO EM CIÊNCIA DA INFORMAÇÃO

TIAGO JOSÉ DA SILVA

INDEXAÇÃO AUTOMÁTICA POR MEIO DA EXTRAÇÃO E SELEÇÃO DE SINTAGMAS NOMINAIS EM TEXTOS EM LÍNGUA PORTUGUESA

Recife

2014

TIAGO JOSÉ DA SILVA

INDEXAÇÃO AUTOMÁTICA POR MEIO DA EXTRAÇÃO E SELEÇÃO DE SINTAGMAS NOMINAIS EM TEXTOS EM LÍNGUA PORTUGUESA

Dissertação apresentada ao Programa de Pós-Graduação em Ciência da Informação da Universidade Federal de Pernambuco como requisito parcial à obtenção do título de mestre em Ciência da Informação.

Área de Concentração: Informação, Memória e Tecnologia

Linha de Pesquisa: Comunicação e Visualização da Memória

Orientador: Prof. Dr. Renato Fernandes Corrêa

Recife

2014

Catalogação na fonte
Bibliotecária Maria Valéria Baltar de Abreu Vasconcelos, CRB4-439

S586i Silva, Tiago José da

Indexação automática por meio da extração e seleção de sintagmas nominais em textos em língua portuguesa / Tiago José da Silva. – Recife: O Autor, 2014.

219 f.: il.

Orientador: Renato Fernandes Corrêa.

Dissertação (Mestrado) – Universidade Federal de Pernambuco. Centro de Artes e Comunicação. Ciência da Informação, 2014.

Inclui referências e apêndices.

1. Ciência da informação. 2. Recuperação da informação. 3. Indexação automática. 4. Processamento de linguagem natural. 5. Gramática gerativa. I. Corrêa, Renato Fernandes (Orientador). II.Titulo.

020 CDD (22.ed.) UFPE (CAC 2014-83)

Serviço Público Federal
Universidade Federal de Pernambuco
Programa de Pós-graduação em Ciência da Informação - PPGCI

TIAGO JOSÉ DA SILVA

INDEXAÇÃO AUTOMÁTICA POR MEIO DA EXTRAÇÃO E SELEÇÃO DE SINTAGMAS NOMINAIS EM TEXTOS EM LÍNGUA PORTUGUESA

Dissertação apresentada ao Programa de Pós-Graduação em Ciência da Informação da Universidade Federal de Pernambuco, como requisito parcial para obtenção do título de mestre em Ciência da Informação

Aprovada em: 28/03/2014

_____

Prof. Dr. Renato Fernandes Corrêa (orientador)
Universidade Federal de Pernambuco

_____

Profª. Drª. Anna Elizabeth G. Coutinho Correia (examinador interno)
Universidade Federal de Pernambuco

_____

Prof. Dr. Renato Rocha Souza (examinador externo)
Escola de Matemática Aplicada - FGV

A Deus que me indicou qual caminho seguir: sem ele eu nada seria. À minha família que sempre esteve comigo em todo estágio da minha vida, dando-me incentivos em tudo. À minha mãe, Edneuza Anacleto, que me apoiou incondicionalmente em todas as decisões de minha vida. Ao meu pai, José Silva, que se orgulha do primeiro filho que entrou na faculdade apesar de todas as adversidades. Aos meus irmãos Tatiane, Tânia, Raquel, Edna e Timóteo. À minha avó Maria e à minha avó Emília que nunca se cansaram de orar por mim. E a algumas pessoas especiais sem as quais eu não teria chegado até aqui: tio José Soares (in memorian), Dr. Joaquim Alcântara, Dr.ª Vanda Alcântara (in memorian) e a minha avó do coração, Raimunda Silva.

AGRADECIMENTOS

Agradeço ao Programa de Pós-Graduação em Ciência da Informação da UFPE pela oportunidade de desenvolver essa pesquisa e pela convivência com grandes pesquisadores da área de Ciência da Informação e áreas afins; isso me fez crescer como profissional. Sou grato ao Corpo Docente do programa pelo amadurecimento que tive ao longo desse período, em especial aos professores Dr. Fábio Mascarenhas e Dr.ª Anna Elizabeth pelas discussões acerca das inovações que a ciência pode proporcionar à indústria brasileira; à professora Dr.ª Cristina Oliveira e ao professor Dr. Lourival Pinto por proporcionarem uma reflexão sobre a função social da informação; à professora Dr.ª Leilah Bufrem por discutir as questões metodológicas da Ciência da Informação; à professora Dr.ª Nadi Presser por me fazer compreender os elos que existem entre Ciência da Informação, Administração e Economia e ao professor Dr. André Fell por me fazer refletir, logo no primeiro momento, sobre a função da minha pesquisa nas organizações. Agradeço também aos professores Dr. Marcos Galindo e Dr.ª Sandra Siebra pelas novas experiências.

Sou agradecido ao meu orientador, Professor Dr. Renato Fernandes Corrêa, que sempre conseguiu me compreender, mesmo quando eu não conseguia ser tão claro. Sua organização foi de extrema importância para que eu não trilhasse caminhos que não fizessem parte dos meus objetivos iniciais. A disciplina ministrada por ele, “Informações em ambientes digitais”, possibilitou a percepção de mais um ambiente onde minha pesquisa pode ser aplicada.

Agradeço ao professor Dr. Raimundo Nonato pelas contribuições que deu em minha formação quanto às questões epistemológicas da Ciência da Informação e às questões metodológicas na Banca de Qualificação do Projeto de Dissertação. Grato também por me ajudar a enfrentar

algumas adversidades que a academia nos proporciona regularmente.

Ao professor Dr. Fábio Pinho que sempre me deu boas orientações em vários trabalhos acadêmicos, apresentou-me as questões epistemológicas quanto a Organização do Conhecimento e contribuiu muito na elaboração dessa dissertação, dando suas sugestões na Banca de Qualificação do Projeto de Dissertação.

Agradeço ao professor Dr. Renato Rocha Souza pelas suas valiosas contribuições na avaliação dessa pesquisa como membro externo da Banca de Defesa, assim como a professora Dr.ª Ana Elizabeth que atuou como avaliador interno.

Outras pessoas se fizeram muito importante na construção dessa pesquisa. A minha grande amiga Gabriela Ortega que sempre esteve comigo desde a adolescência e que sempre conseguiu fazer uma leitura crítica dos meus textos, permitindo, assim, que eu os melhorasse.

Aos meus colegas do curso de Gestão da Informação que me descontraiam quando eu estava muito tenso. Aos meus colegas do mestrado que proporcionaram discussões relevantes acerca do meu tema: Ana Cláudia, Pedro Manoel, Jéssica Monique, Gilvanedja Ferreira, Pollyanna Muniz e Tereza Lucena. Sou grato também a duas pessoas irmãs: - Remi Lapa e Monick Santos, nossas conversas foram muito boas.

Quatro pessoas do mestrado estiveram comigo em todas as situações ao longo dessa jornada, por isso sou muito agradecido: à Rosana Marinho pelo companheirismo; à Aurelina Lopes que se mostrou muito mais que uma colega, é uma verdadeira amiga em quem posso confiar e buscar apoio sempre; à Karla Farias que se apresentou como uma guerreira, a qual admiro muito; ao meu irmão Erinaldo Dias, não esquecerei os dias que ficamos estudando na Biblioteca do Centro de Educação, isso possibilitou que minha pesquisa

evoluísse: – Amigo, obrigado por tudo! Uma das melhores pessoas que Deus permitiu que eu conhecesse. Aproveito para agradecer à Nicácia Lina, outra irmã.

Sou grato a Juliana Cysneiros que me auxiliou na minha jornada de trabalho, permitindo assim que eu pudesse assistir às aulas do mestrado: – Ju, Firma, muito grato a você por tudo, minha amiga, minha irmã!

Agradeço à diretora da escola na qual leciono, Edilene Barreto, que flexibilizou meu horário sempre que eu pedia para poder estar no mestrado. Sem suas ações as coisas teriam ficado mais difíceis para mim. Aproveito para agradecer à Ocidélia (vice-gestora) e à Flávia Lima (secretária) pelo apoio. Também sou agradecido aos meus colegas de trabalho Giovanna Weyne, Andrea Araújo, Laís Lira, Rafaely Maira, Ivson Marques, Andrea Karla, Risonete Martins, Vladirene Lima, entre tantos outros.

Agradeço a todos que de maneira direta ou indiretamente contribuíram para a produção dessa pesquisa.

“Sou humano, não consigo ser perfeito”

(Aderson Freire)

“Saindo do meu conforto
Eu vejo do alto
Sonhando além dos meus sonhos
Enxergo mais longe
Rompendo com meus limites
Até que eu alcance
Mexendo com as estruturas
Amplio horizontes

Digo a mim mesmo
O tempo é o tempo e não volta
Não há mais tempo a perder
Não mais (...)”

(*Marcelo Manhãs)*

RESUMO

Objetiva fazer um levantamento do estado da arte da indexação automática por sintagmas nominais para textos em português. Para tanto, identifica e sintetiza os fundamentos teóricos, metodologias e ferramentas da indexação automática por meio da extração e seleção de sintagmas nominais em textos em língua portuguesa, levando em conta publicações científicas nas áreas da Ciência da Informação, Ciência da Computação, Terminologia e Linguística. Discute as metodologias para indexação automática através de sintagmas nominais em textos em língua portuguesa, no intuito de apontar critérios para extração e seleção de sintagmas que possam ser usados como descritores documentais. Avalia e compara ferramentas de extração automática de sintagmas nominais como o *parser* PALAVRAS, OGMA e LX-Parser, usando como referência a extração manual de sintagmas nominais. Percebe que os trabalhos produzidos depois do ano de 2000 e que trabalham com a extração automática de termos fazem referências ao *parser* PALAVRAS, tendo-o como um bom etiquetador e analisador sintático. Na comparação entre as referidas ferramentas automáticas, percebe-se que apesar do LX-Parser ter tido melhor desempenho em alguns aspectos como extrair um maior número de SNs do que o PALAVRAS, esse ainda consegue ser melhor pelo número menor de erros e a possibilidade de submeter um texto completo à análise do programa, ação que o LX-Parser não permite realizar. Quanto ao levantamento do estado da arte, pode-se dizer que as pesquisas ainda não atingiram um grau de amadurecimento elevado, pois os resultados apresentados pela literatura não alcançam uma taxa de precisão elevada para todos os tipos de corpus. Conclui que os resultados das pesquisas que trabalham com a extração automática de sintagmas nominais devem ser comparados entre si para que se possam detectar os problemas existentes quanto às metodologias e às ferramentas de extração destes sintagmas nominais em língua portuguesa. Tendo, dessa maneira, as ferramentas e as metodologias melhoradas para que efetivamente possam ser aplicadas em sistemas de recuperação de informação, fazendo a seleção de sintagmas nominais que possam ser usados como descritores documentais no intuito de satisfazer as necessidades informacionais do usuário. Sugere, então, algumas possíveis

soluções para os problemas de identificação de sintagmas nominais enfrentados pelas ferramentas automáticas.

Palavras-chave: Sintagmas Nominais. Recuperação de Informação. Indexação Automática. Extração Automática de Sintagmas Nominais. Processamento de Linguagem Natural.

ABSTRACT

This dissertation aims to survey the state of the art of automatic indexing through noun phrases extraction from texts written in Portuguese. To do so, it identifies and summarizes the theoretical foundations, methodologies and tools of automatic indexing through the extraction and selection of noun phrases from Portuguese language, considering scientific publications in the areas of Information Science, Computer Science, Linguistics and Terminology. It discusses the methodologies for automatic indexing through nominal phrases in texts written in Portuguese language, in order to point criteria for the extraction and selection of phrases that can be used as document descriptors. It evaluates and compares automatic extraction of noun phrases parser tools such as PALAVRAS , OGMA and LX –Parser, using as reference the manual extraction of noun phrases. It realizes that the works produced after the year of 2000 and that work with the automatic terms extraction make references to parser PALAVRAS, taking it as a good tagger and parser. In the comparison between the above-mentioned automatic tools, it is seen that in spite of the LX-Parser to have had better performance in some aspects as it extracts a bigger number of SNs than PALAVRAS, this still manages to be better for the smaller number of errors and the possibility of submitting a comprehensive text analysis program, action that the LX -Parser cannot accomplish. As the lifting of the state of the art, it is possible to say that the inquiries still did not reach a degree of elevated maturity, since the results presented in the literature do not reach a tax of precision lifted up for all the types of corpus. It concludes that the results of the inquiries that work with the automatic extraction of noun phrases must be compared between themselves so that the existent problems can be detected as for the methodologies and the tools of extraction of these noun phrases in Portuguese language. In this way, it has the tool and methodologies for improved that can actually be applied in systems of information retrieval, making the selection of noun phrases that can be used as document descriptors in the intention of satisfying the necessities of information of the user. It suggests, then, some possible solutions to the problems of identification of noun phrases faced by automatic tools.

Keywords: Noun Phrases. Information Retrieval. Automatic Indexing. Automatic Extraction of Noun Phrases. Natural Language Processing.

LISTA DE FIGURAS

Figura 1 – Representação arbórea de uma análise sintática..... 58

Figura 2 – Transformação de uma EP em ES..... 66

Figura 3 – Esquema do processo de geração de sentença..... 66

Figura 4 – Regras de transformação de EP em ES..... 68

Figura 5 – Interface do OGMA..... 80

Figura 6 – Forma gráfica da análise do *parser* PALAVRAS..... 82

Figura 7 – Forma arbórea da análise do *parser* PALAVRAS..... 83

Figura 8 – Análise textual feita pelo do *parser* PALAVRAS..... 84

Figura 9 – Interfaces gráficas do Grammar Play..... 85

Figura 10 – Representação arbórea feita pelo LX- Parser..... 87

Figura 11 – Interface gráfica do Curupira..... 88

Figura 12 – Procedimentos de interação usuário ↔ protótipo 106

Figura 13 – Esquema geral de extração dos SNs no método ED-CER..... 114

Figura 14 – Arquitetura do subsistema de filtragem em conjunto com sistema de RI..... 132

Figura 15 – Estrutura em módulos do sistema SISNOP..... 140

Figura 16 – Processo geral de extração automática de conceitos..... 148

Figura 17 – Erro de etiquetagem do LX-Parser..... 182

Figura 18 – Marcação com o pronome relativo “que” pelo LX-Parser..... 183

Figura 19 – Marcação com o pronome relativo “que” pelo Parser PALAVRAS..... 184

LISTA DE GRÁFICOS

Gráfico 1 – Níveis de acertos na identificação de expressões que constituem SNs..... 172

Gráfica 2 – Taxa de acerto na recuperação de SNs semelhantes às palavras-chave dos resumos..... 176

LISTA DE QUADROS

Quadro 1 – Relacionamento entre as classes Gramaticais do FORMA, do PALAVRAS e do SISNOP..... 53

Quadro 2 – Gramática do método ED-CER..... 117

Quadro 3 – Equações utilizadas por Santos (2005)…. 129

Quadro 4 – Regras de extração de SN do OGMA..... 143

LISTA DE TABELAS

Tabela 1 – Etiquetas utilizadas no OGMA e ED-CER. 52

Tabela 2 – NILC TAGSET5 – 36 tags..... 54

Tabela 3 – Diferenças dos etiquetadores morfossintáticos..... 55

Tabela 4 – Regras de formação de SNs 74

Tabela 5 – Valor atribuído ao SN de acordo com sua relevância..... 92

Tabela 6 – Valor atribuído ao SN de acordo com sua estrutura sintática..... 93

Tabela 7 – Símbolos terminais para a gramática ED-CER..... 116

Tabela 8 – Símbolos não terminais para a gramática ED-CER..... 117

Tabela 9 – Comparação entre alguns trabalhos que compõe a literatura de extração de SNs 160

Tabela 10 – Comparação entre alguns trabalhos que compõe a literatura de extração de SNs quanto aos sistemas propostos e usados 162

Tabela 11 – Total de SNs identificados nos resumos por cada programa e pela extração manual..... 167

Tabela 12 – Quantidade de expressões identificadas, mas que não constituem sintagmas..... 171

Tabela 13 – Quantidade de vezes que cada programa e a extração manual recuperam SNs semelhantes às palavras-chave dos resumos 175

Tabela 14 – Quantidade de SNs que continham palavras-chave dos resumos..... 178
Tabela 15 – Quantidade SNs relevantes para a descrição dos resumos..... 179

LISTA DE SIGLAS

BDTD-UFPE – Biblioteca Digital de Teses e Dissertações da Universidade Federal de Pernambuco
BOW – *Bag-of-Words*
BRAPCI – Base de Dados Referenciais de Artigos de Periódicos em Ciência da Informação
Capes - Coordenação de Aperfeiçoamento de Pessoal de Nível Superior
CFG – *Context Free Grammar*
CI – Ciência da Informação
CLEF – *Cross Language Evaluation Forum*
DCG – *Definite Cause Grammar*
DET – Determinante
DITED - Repositório Nacional de Dissertações e Teses Digitais (Portugal)
ENANCIB - Encontro Nacional de Pesquisa em Ciência da Informação
EP – Estrutura Profunda
ES – Estrutura Superficial
EχATOLP – Extrator Automático de Termos para Ontologias em Língua Portuguesa
HPSG – *Head-Driven Phrase Structure Grammar*
HTML – *Hyper Text Markup Language*
IBICT – Instituto Brasileiro de Informação em Ciência e Tecnologia
LKB – *Linguistic Knowledge Builde*
LSA – Análise Semântica Latente
LSI – *Latent Semantic Indexing*
MOD – Modificador
N – Nome
NBR – Norma Brasileira
NILC - Núcleo interinstitucional de Linguística Computacional (Universidade de São Paulo – USP)
NLX-Group – *Natural Language and Speech Group* (Departamento de informática da Universidade de Lisboa)

PDF – *Portable Document Format*
PLN – Processamento de Linguagem Natural
PSG – *Phrase Structure Grammar*
RI – Recuperação da Informação (*Information Retrieval)*
SA – Sintagma Adjetival
SAdv – Sintagma Adverbial
SciELO - *Scientific Electronic Library Online*
SIDSN – Sistema Identificador de Sintagmas Nominais
SISNOP - Sistema Identificador de Sintagmas Nominais do Português
SN – Sintagmas Nominais
SP – Sintagma Preposicional
SRI – Sistema de Recuperação da Informação
SV – Sintagma Verbal
TA – Termos Atômicos
TBL – Aprendizado Baseado em Transformações
TFIDF – *Term Frequency Inverse Document Frequency*
TIC's – Tecnologias Da Informação e Comunicação
UFAL – Universidade Federal de Alagoas
UFSCar – Universidade Federal de São Carlos
VISL – *Visual Interactive Syntax Learning*

SUMÁRIO

1 INTRODUÇÃO..... 23

2 REVISÃO DE LITERATURA..... 34
2.1 Indexação Automática..... 37
2.2 Processamento de Linguagem Natural (PLN)..... 46
2.2.1 Etiquetador Morfossintático (*Tagger*)..... 49
2.2.2 Analisador Sintático (*Parser*)..... 56

2.3 Sintagmas Nominais (SNs)..... 62
2.3.1 Gramática Gerativa..... 62
2.3.2 Extração Automática de SNs..... 75
2.3.3 Identificador ou Extrator de Sintagmas Nominais... 77
2.3.4 Seleção Automática de SNs..... 89
2.3.5 Avaliação da Extração Automática de SNs..... 94

3 PROCEDIMENTOS METODOLÓGICOS..... 96

4 PRINCIPAIS CONTRIBUIÇÕES PARA A EXTRAÇÃO AUTOMÁTICA DE SNs (ESTADO DA ARTE)..... 104
4.1 Contribuições de Kuramoto (1995 - 2002)..... 104
4.2 Contribuições de Bick (2000)..... 109
4.3 Contribuições de Vieira et al (2000)..... 111
4.4 Contribuições de Miorelli (2001) 113

4.5 Contribuições de Othero (2004)..... 118
4.6 Contribuições de Souza (2005)..... 120
4.7 Contribuições de Santos (2005) 125
4.8 Contribuições de Arcoverde (2007)..... 130
4.9 Contribuições de Costa (2007)..... 134
4.10 Contribuições de David (2007)..... 135
4.11 Contribuições de Morellato (2007; 2010)..... 137
4.12 Contribuições de Maia (2008)..... 141
4.13 Contribuições de Pinheiro (2009)..... 145
4.14 Contribuições de Lopes (2011)..... 148
4.15 Contribuições de Corrêa et al (2011)..... 153
4.16 Discussão Acerca do Estado da Arte..... 155

5 IDENTIFICADORES DE SINTAGMAS NOMINAIS: uma análise comparativa entre o OGMA, Parser PALAVRAS, LX Parser e a extração manual 165

6 CONSIDERAÇÕES FINAIS..... 187

REFERÊNCIAS..... 192

APÊNDICES – Extração de SNs dos resumos de teses e dissertações..... 208
APÊNDICE A – Extração manual 209
APÊNDICE B – Extração pelo Parser PALAVRAS 212
APÊNDICE C – Extração pelo LX-Parser 215

APÊNDICE D – Extração pelo OGMA..... 219

1 INTRODUÇÃO

No século XX, iniciou-se o fenômeno denominado de explosão informacional que culminou em grandes transformações da sociedade como a evolução da ciência e da tecnologia.

Entende-se o fenômeno da explosão informacional como o crescimento da informação disponibilizada, ou seja, quando o processo de geração, disseminação, acesso e uso da informação foi impulsionado por meio da tecnologia, o que se somando a outros fatores acarretou numa inefável produção de informação.

A informação, na sua grande parte, não estava mais restrita aos pequenos seletos grupos, agora ela pode ser acessada por outros indivíduos que podem produzir outras informações a partir daquela. Com isso, surge um novo problema, a recuperação de informação relevante em meio a esse grande volume de informação produzida.

Criar novas ferramentas que possibilitem a realização da busca de um usuário e que esse satisfaça sua necessidade de informação é o objetivo dos estudiosos do objeto informação, tanto o cientista da informação que se preocupa com o processo de "produção, seleção, organização, interpretação, armazenamento, recuperação, disseminação, transformação e uso da informação" (GRIFFITH, 1980 apud CAPURRO, 2003, documento não paginado) quanto os cientistas da computação que se preocupam em tornar essas ações viáveis dentro dos sistemas computacionais.

Todavia, a preocupação com o processo informacional já existia bem antes do fenômeno da explosão informacional. Os estudiosos da informação, no século XIX, começaram a estudar métodos para o tratamento informacional. Tem-se, como exemplos, o Sistema de Classificação Decimal de Dewey e o Movimento de Documentação de Paul Otlet e Henri La Fontaine, datados de 1876 e 1890 respectivamente. Já em 1922, são mostrados por Hulme os Estudos Quantitativos de Produção Bibliográfica. A Teoria e Prática da Classificação é apresentada por Bliss em 1929. Três anos depois, Ranganathan elabora os Sistemas de classificação e leis para bibliotecas. Bradford e Lancaster Jones criaram, em 1934, a Distribuição bibliométrica. A "Aplicação de métodos de pesquisas sociais em estudos sobre bibliotecas" foi desenvolvida por Waples, na década de 1930. (VICKERY; VICKERY, 1987; DIAS, 2002; ROBREDO, 2003; NASCIMENTO, 2006).

Mesmo com esses estudos, fazia-se necessário uma ciência que compreendesse as propriedades, comportamentos, relações, desenvolvimento e concepções conceituais da informação. Com esse intuito, surgiu a Ciência da Informação (CI) de natureza interdisciplinar (SARACEVIC, 1996; LE COADIC, 2004), que estuda o objeto informação aplicando métodos e técnicas próprias ou originárias de outras ciências, somando-se ao uso de percursos teóricos de outras áreas. Isso explica o porquê dos autores dizerem que ela epistemologicamente é interdisciplinar.

Para Pinheiro e Loureiro (1995), foi Nobert Wiener com a obra *Cybernetics or Control and Communication in The Animal and Machine*, em 1948, e no ano seguinte, o livro "The

Mathematical Theory of Communication" de Claude Shannon e Warren Weaver prenunciava o que viria a ser CI.

Vale ressaltar que esses autores trabalharam a questão do surgimento do termo, mas na presente pesquisa, opta-se por afirmar que o marco inicial dessa nova disciplina (CI) foi com a publicação do artigo de Bush, em 1945, "*As we may think*" (SARACEVIC, 1996; BARRETO, 2002). Bush apontou os entraves existentes para organizar e repassar à sociedade as informações sigilosas durante a Segunda Guerra Mundial (BARRETO, 2002).

Cardoso (1996) diz que lidar com o grande volume e a diversificação de informações registradas em diversos suportes e em diversas formas foi o que condicionou a existência da CI. Capurro (2003), baseado nos conceitos de paradigmas de Kuhn, defende que a CI surgiu com o paradigma físico, esse paradigma perdeu espaço para uma visão mais idealista e individualista trazida pelo paradigma cognitivo e que, por sua vez, foi questionado pelo paradigma pragmático e social.

Segundo Barreto (2002, p. 69), Bush "introduziu a noção de associação de conceitos ou palavras na organização da informação", apontou a intuição e as limitações existentes à época nos sistemas de classificação e indexação.

As soluções apontadas por Bush eram operacionalização da associação entre os conceitos, ou seja, "como nós podemos pensar" (*As we may think*), no processo de armazenamento e recuperação da informação, e a formação do profissional da informação. Essa era vista como conservadora na época.

Outro problema apontado por Bush era a literatura sobre a construção dos sistemas de organização da informação que estava ultrapassada e errada, para isso, ele propôs um apetrecho tecnológico que tinha como função o armazenamento e recuperação de documentos mediante associação de palavras (BARRETO, 2002).

Contextualizada a CI no aspecto histórico, cabe defini-la à luz de alguns autores da área, porque Le Coadic (2004) diz que as ciências são atividades sociais que são determinadas por condições históricas e socioeconômicas. Assim, a CI é uma ciência social que se preocupa em esclarecer um problema concreto que é todo o processo que envolve a informação, voltada para um sujeito social que procura informação. Entende-se por informação como um conceito intrínseco à sociedade moderna e que possui um valor político, histórico e socioeconômico.

Pinheiro (1999, p. 156) sintetiza o que se quer falar a respeito da CI nessa introdução. A CI, além de interdisciplinar, é de natureza social que está "relacionada à tecnologia da informação e do novo papel da informação na sociedade e na cultura contemporânea", ou seja, está preocupada com os problemas da sociedade atual que estão relacionados às questões como o acesso e uso da memória coletiva possibilitado pelas Tecnologias da Informação e Comunicação (TIC's).

Retomando Le Coadic (2004), a CI é uma das novas interdisciplinas que se relaciona com outras áreas do conhecimento como: psicologia, linguística, sociologia, informática, matemática, lógica, estatística, eletrônica, economia, direito, filosofia, política e telecomunicações. A

institucionalização se dá pelas revistas científicas, banco de dados, sociedades científicas, profissionais da informação. Esse argumento reforça o posicionamento de Saracevic (1996, p. 41) que, por sua vez, dá fundamento aos argumentos expostos nos primeiros parágrafos dessa introdução.

> Primeiramente, a ciência da informação é interdisciplinar por natureza e a interdisciplinaridade está longe de acabar, em segundo lugar, a ciência da informação está inexoravelmente conectada à tecnologia da informação, terceiro, a ciência da informação é, juntamente com outros campos, um participante ativo na evolução da sociedade da informação.

A CI estuda todo um processo informacional para que documentos sejam preservados, no intuito de que a informação seja recuperada, auxiliando a sociedade na construção e reconstituição de sua memória.

Somando-se a essas concepções, tem-se a argumentação de Ortega (2004) de que a CI tem origem na bifurcação da Documentação/Bibliografia e da Recuperação da Informação. “É uma ciência social cujo objeto é a informação, tendo início no campo da informação científica e tecnológica, passando a atuar também com a informação para fins educacionais, sociais e culturais” (p. 10).

O fato de mencionar a Recuperação da Informação (RI) nesse texto, é que ela é primordial ao estudo proposto por este trabalho, pois ela envolve vários fatores metodológicos e técnicos, um desses processos é a questão da indexação que é uma das fases da organização da informação que, por sua vez, tem como intuito recuperar a informação.

O termo *Information Retrieval* (RI) surgiu em 1950, cunhado pelo pesquisador Calvin Moores, à mesma época da institucionalização da CI, e consiste, *a priori*, na busca de documentos relevantes a uma determinada consulta que foi externada de acordo com a necessidade informacional de um usuário (GONZALES; LIMA, 2003).

Saracevic (1999) diz que a RI, enquanto área de pesquisa, é primordial no processo de disseminação e criação do conhecimento e coopera para o progresso científico no campo da CI.

Por meio desse processo é que o usuário encontra a informação que procura. Se esse processo foi construído de maneira inadequada, toda a expectativa do uso será frustrada. O processo de RI envolve representação, armazenamento, organização e acesso aos documentos. Para tanto, devem ser observadas, também, a indexação, a relevância e os termos de índice, assim como os de consulta, pois com a evolução da internet, a área de RI vem desenvolvendo cada vez mais métodos e técnicas que aprimorem a indexação e a busca de documentos. Os estudos sobre essa vertente muitas vezes estão concentrados na adaptação de métodos e técnicas pesquisados na área de Processamento de Linguagem Natural (PLN), principalmente na subárea de processamento automático de texto.

Meadows, Boyce e Kraft (2000) apontam três conceitos que fundamentam a RI: a representação da informação, interpretação da estrutura de seus símbolos e a identificação de um conjunto de símbolos.

Os sistemas de Recuperação da Informação (SRI) que compreendem a interface de RI, o tratamento da informação,

seu armazenamento e sua organização (KURAMOTO, 1997) adotam termos índices que geralmente são as palavras-chave para indexar documentos (SOUZA, 2006). É esse ponto que o trabalho aqui proposto observa, pois segundo Baeza-Yates e Ribeiro Neto (1999), a requisição do usuário se perde, pois é muito simplório substituir um texto completo por algumas palavras que não carregam valor discursivo suficiente para descrever esses documentos.

Smeaton (1989 apud KURAMOTO, 2006) define SRI como um sistema que tem por objetivo buscar documento no intuito de atender uma solicitação do usuário. Para que um SRI desempenhe sua função corretamente, é necessário que a indexação dos documentos seja feita de maneira adequada, pois envolve padronização, o que remete a uma política de indexação. Essa ação deveria ser feita por um especialista. Contudo, a demanda de informação atual torna isso impossível, ficando a cargo da indexação automática.

A indexação automática pode ser tomada como o tratamento prévio que os documentos devem passar para serem armazenados em uma base de dados, no intuito de que cada documento possa ser posteriormente recuperado. A indexação automática, feita pelos SRIs, extrai os descritores e sua estruturação para representar os documentos, tudo isso feito a partir da análise dos computadores dotados de programas elaborados pela área de Inteligência Artificial, ou Inteligência Computacional, no escopo da Ciência da Computação, e também pela área de Organização da Informação, no escopo da CI.

Um dos problemas semânticos encontrados na RI é quando há polissemia, quando um significante tem vários

significados, e sinonímia, quando o mesmo significado pode ser representado por vários significantes. Portanto, a palavra como descritor de um documento não é suficiente, pois, na maioria dos casos, ela é constituída desses fenômenos (polissêmicos e sinonímicos), além de duas ou mais palavras se combinarem em ordens diferentes, o que pode trazer ideias completamente diferenciadas. Isso faz com que haja um aumento das taxas de silêncio (devido à sinonímia) e ruídos (devido à polissemia e combinação de palavras) na recuperação de um documento.

As palavras, quando são extraídas de um documento, perdem valores atribuídos pela autoria do documento, o que leva a perda de qualquer referência da realidade extralinguística do autor (KURAMOTO, 1995). Sendo assim, é necessário que os descritores de um documento sejam contextualizados e que representem a informação sem descaracterizá-la. Dessa maneira, pode-se pensar nos sintagmas nominais.

Os sintagmas nominais (SNs) são objetos de estudo de várias áreas do conhecimento como a Linguística, Ciência da Computação e Ciência da Informação (CI). Contudo, sob a perspectiva da CI, as pesquisas sobre SNs ainda são muito restritas e muitos estudos sobre essa temática foram interrompidos por causa de diversos fatores. Assim, pretende-se ter uma visão geral acerca das pesquisas sobre os SNs no contexto da indexação automática e RI.

Kuramoto (1995) percebe os SNs como uma abordagem alternativa na recuperação da informação. Os SNs são tidos como essenciais descritores dos documentos, ou seja, facilitam a RI. Em outras palavras, os SNs podem se

tornar descritores em SRIs que são *softwares* buscadores de documentos, o que permitiria uma recuperação melhor.

O **problema da pesquisa** consiste no apontamento de melhorias para as metodologias e ferramentas de extração automática de SNs e no levantamento do estado da arte das pesquisas desenvolvidas acerca da extração de SNs em textos escritos em língua portuguesa.

A **problemática** do trabalho está direcionada para uma reflexão acerca de algumas questões: Mas será que a indexação automática por meio da extração de SNs descreveria documentos? Se os SNs são os melhores descritores de documentos, como utilizá-los nesse processo de indexação e recuperação da informação? Quais as regras de formação de SNs? Todos os sintagmas nominais são bons descritores documentais? Quais são os fundamentos teóricos da CI, da Linguística, da Terminologia e da Ciência da Computação que poderiam guiar a seleção dos SNs como descritores de um documento? Como avaliar a qualidade da indexação automática por SNs? Qual é o estado da arte da indexação automática por meio de sintagmas nominais para textos em língua portuguesa?

Diante dessa problemática, o **objetivo geral** desse trabalho é: Fazer um levantamento do estado da arte da indexação automática por sintagmas nominais para textos em português.

Os **objetivos específicos** são:

- Levantar e sintetizar os fundamentos teóricos da indexação automática por meio da extração de SNs nas áreas do conhecimento como CI, Linguística e Ciência da Computação;

- Coletar e analisar os trabalhos e pesquisas sobre a temática;
- Identificar na literatura critérios de extração e seleção de SNs com valor de descritores, assim como pesquisar e arrolar os processos de identificação, extração e seleção automática de SNs em textos escritos em língua portuguesa;
- Fazer o levantamento de *softwares* extratores e identificadores de SNs para textos em Português, avaliando e comparando tais softwares com base na representação arbórea de sentenças estudada em Linguística, especialmente pela Gramática Gerativa elaborada por Chomsky (1965) e pelas releituras de autores atuais dessa.

Esta dissertação se estrutura em 6 capítulos.

O primeiro sendo a Introdução, onde se abordam o problema, a problemática, os objetivos e a estrutura do trabalho. O segundo trás a fundamentação teórica, em que se aplica o conceito de interdisciplinaridade, pois, buscam-se referências na Ciência da Informação, Ciência da Computação, Linguística e Terminologia para definir os conceitos como Processamento de Linguagem Natural (PLN), etiquetadores morfossintáticos, analisadores sintáticos, identificador ou extrator de SNs, indexação automática; os conceitos de SNs juntamente com a Gramática Gerativa, extração automática de SNs, metodologias de extração automática de SNs, ferramentas de extração automática de

SNs e metodologias de avaliação da extração automática de SNs.

O terceiro capítulo descreve todo percurso metodológico desde a coleta de dados à análise dos mesmos, assim como a construção do estado da arte, o referencial teórico e o experimento.

No quarto capítulo, apresenta-se um panorama sobre o estado da arte acerca dos estudos sobre extração de sintagmas nominais em língua portuguesa, apontando os estudiosos percussores e os projetos desenvolvidos.

No quinto capítulo, é feita uma comparação entre as ferramentas OGMA, Parser PALAVRAS e o LX-Parser, tendo como referência a extração manual de SNs. Nessa comparação, foram observadas a quantidade de SNs identificados, a quantidade de falsos SNs e SNs relevantes. O experimento tem como finalidade mostrar como as ferramentas de extração se comportam na extração e/ou identificação de SN, assim há o apontamento de algumas falhas e sugestões de possíveis soluções para os problemas apresentados por essas ferramentas.

O sexto capítulo é constituído pelas considerações finais em que se faz uma retomada de alguns pontos principais da dissertação e algumas sugestões de trabalhos futuros.

Logo em seguida, têm-se as referências e os sites consultados na elaboração dessa pesquisa, assim como alguns apêndices elaborados durante avaliação das ferramentas automáticas.

2 REVISÃO DE LITERATURA

No referencial teórico, opta-se por fazer uma leitura de alguns autores de quatro disciplinas diferentes para construir um texto interdisciplinar, já que o objeto dessa pesquisa tem seus fundamentos conceituais nestas áreas do conhecimento. As áreas do conhecimento aqui expostas são a Linguística, a Terminologia, a CI e a Ciência da Computação.

A Linguística é, segundo Cunha, Costa e Martelotta (2012), a ciência que estuda a linguagem humana observando sua manifestação oral, escrita ou gestual, tendo como objetivo depreender os princípios que fundamentam e regem essa capacidade humana de expressão por meio de línguas. Para tanto, os linguistas analisam como as línguas naturais se estruturam e funcionam. Podem-se citar algumas áreas específicas como a Linguística Computacional Aplicada que se divide em Linguística de Corpus que, segundo Beber Sardinha (2004), se ocupa de investigar quase todas as áreas da linguística, tendo o léxico a maior atenção dada pelos linguistas de corpus e em Processamento de Linguagem Natural (PLN) que será conceituado na Seção 2.2 dessa dissertação.

A Terminologia pode ser definida como disciplina que trabalha com objeto denominado de “o termo técnico - científico”. Para Krieger e Finatto (2004), esse objeto marca a identidade Terminologia juntamente com a fraseologia especializada e a definição terminológica. Wüster (1998), fundador da Terminologia moderna, argumenta que essa disciplina é autônoma e multidisciplinar e que mantém uma

relação estreita com outras disciplinas a partir da convergência da linguística, da lógica, da ontologia, das "ciências da informação" [1], entre outras disciplinas, e que deve manter um intercâmbio com áreas como a física, engenharia elétrica e a economia.

O conceito de CI foi discutido na introdução dessa dissertação, mas recapitulando: é uma ciência que, além de outras questões sociais, estuda todo o processo de tratamento da informação com o objetivo de que ela seja recuperada, satisfazendo a necessidade informacional do usuário, como é visto em Griffith (1980 apud CAPURRO, 2003, documento não paginado) quando afirma que a CI trata da "produção, seleção, organização, interpretação, armazenamento, recuperação, disseminação, transformação e uso da informação".

A Ciência da Computação, segundo Brookshear (2003), é a disciplina que estuda a construção e a programação de computadores, o processamento de informações, as soluções algorítmicas de problemas, assim como todo o processo algorítmico. Boa parte dos conceitos acerca dos analisadores sintáticos e morfológicos e do Processamento de Linguagem Natural é advinda dessa ciência. As ferramentas de extração automática também são desenvolvidas, tomando-se como referência os estudos construídos dentro da Ciência da Computação e/ou da relação estabelecida com outras ciências.

---

[1] O autor se refere a todas as disciplinas que trabalham com o objeto "informação", o que permite citar, sem entrar na discussão epistemológica, a CI, a Biblioteconomia, a Documentação, a Arquivologia, entre outras.

O estudo proposto nessa dissertação recorre à:

1. Linguística porque dá subsídio teórico para a concepção dos SNs, assim como estuda os fenômenos morfológicos e sintáticos que ocorrem na construção de uma sentença em linguagem natural;
2. Terminologia, pois faz a ligação dos estudos lexicológicos, lexicográficos, terminográficos, terminológicos e documentários (KRIEGER; FINATTO, 2004), tendo sua aplicação na geração de glossários e dicionários especializados, de bancos de dados, de termos técnico-científicos e suas definições, de tradução, na produção de redação técnica e na gestão da informação, entre outras aplicações. Dessa maneira, essa disciplina pode contribuir no processo de escolha dos SNs que servem como descritores documentais e permite trabalhar com os conceitos e definições técnico-científicos acerca dos objetos discutidos nessa dissertação;
3. CI porque trabalha com a organização da informação (do conhecimento), permitindo observar as concepções acerca das técnicas de indexação e representação da informação usando os SNs;
4. Ciência da Computação, pois faz com que essas teorias se tornem uma ação real em um sistema computacional, ou seja, todo o

> processo de tratamento da informação sendo feito a partir de algoritmos.

Tomar esse posicionamento em relação ao referencial teórico se justifica pelo fato dos objetos que compõem a problemática em estudo[2] sejam abarcados por especialidades diferentes, o que permite fazer uma rede entre essas ciências, pois há um objeto comum a essas áreas, o SN.

Os estudos acerca dos SNs na Linguística se dão para mostrar que as combinações de sintagmas levam ao infinito de sentenças; na Terminologia, podem ser usados nas aplicações terminológicas em geral; na CI, para descreverem documentos no processo de indexação; e na Ciência da Computação, para identificarem e extraírem informações de um documento por meio de ferramentas automáticas.

2.1 Indexação Automática

Para tornar a RI de um documento um processo mais simples e eficaz, é necessário representar esse documento de forma que os termos indexadores sejam de fato uma descrição abreviada do conteúdo documental.

Borges (2009) conceitua o ato de indexar como atividade de selecionar ou definir palavras ou expressões que servirão como descritores de um conteúdo de um determinado documento, levando-se em conta as considerações da clientela específica.

---

[2] a indexação automática por meio da extração manual de SNs.

Para Navarro (1988) é a tradução do conteúdo de um documento, feita a partir de palavras que possibilitem a recuperação. Brito (1992, p. 224) corrobora com Navarro, ao dizer que a

> "indexação, tal qual nós a vemos, é uma tradução lexical das unidades da língua, ou ainda uma tradução sintática, quando se trata de exprimir as relações entre as diferentes partes do discurso (...)".

Já Vieira (1988a) define a indexação como técnica de análise de conteúdo que possibilita a condensação da informação significativa de um documento por meio de termos, criando uma linguagem intermediária entre o usuário e o documento.

Dessa maneira, Souza (2005) afirma que há duas fases independentes na indexação: a análise de assunto que também é chamada de análise conceitual; e a tradução. Na primeira fase, determina-se a temática do conteúdo, enquanto na tradução, há a representação dos assuntos pertinentes identificados. Essa representação pode ser feita através de uma linguagem de indexação (códigos de classificação, palavras-chave em um vocabulário controlado, símbolos).

Lancaster (2004) cita três dimensões de indexação: a exaustividade que está relacionada ao adicionar mais termos a indexação; a seletiva que se refere ao processo de quando menos termos são incluídos; e a especificidade que é referida a tarefa de se usar termos mais específicos que façam com que o documento seja compreendido integralmente.

Souza (2005) aponta que se a exaustividade for aumentada há um aumento na revocação[3] e a diminuição da precisão[4] na RI, já se for aumentada a especificidade, haverá o aumento da precisão e a diminuição da revocação.

Araújo Júnior (2007) reforça que o objetivo da indexação é obter um melhor aproveitamento no processo de busca e RI, já que o elemento fundamental é a representação do conteúdo dos documentos. Holanda e Braz (2012) conceituam indexação como a substituição do texto de um documento por uma descrição do conteúdo tratado, no intuito de que as informações contidas nesse documento sejam recuperadas. Segundo Maimone e Francelin (2008, p.75),

> "indexação é um processo de fabricação de *informações documentárias* [...]. É uma das formas de *descrição* de conteúdo informacional. É o trabalho (operação) de "escolha" de termos que *representem* apropriadamente o conteúdo informacional de um documento". [destaques dos autores]

Dessa maneira, pode-se definir indexação como a descrição documental, criando informações que representam o conteúdo informacional de um documento, no intuito de que esse seja recuperado facilmente. Para indexar um documento,

---

[3] Revocação: faz a mensuração do sucesso de um SRI em recuperar documentos, consiste da razão entre o número de documentos pertinentes recuperados e o total do número total de documentos pertinentes.

[4] Precisão: mede o sucesso de um SRI em não recuperar documentos que não sejam relevantes para uma determinada necessidade informacional, consiste da razão entre o número de documentos pertinentes e o número de documentos recuperados.

deve-se adotar uma política de indexação que são ações estabelecidas no processo de análises contextuais (MAIMONE; FRANCELIN, 2008). Como exemplos dessas ações têm-se, segundo a NBR 12676 (ASSOCIAÇÃO BRASILEIRA DE NORMAS TÉCNICAS, 1992, p.2), três estágios fundamentais para indexar um conteúdo: a) exame do documento e estabelecimento do seu conteúdo; b) identificação dos conceitos presentes no assunto; c) tradução desses conceitos nos termos de uma linguagem de indexação.

Há três formas de se fazer indexação de documentos. A primeira é desenvolvida pelo homem, chamada então de manual, a segunda é a automática que é feita pelo computador e a terceira é a híbrida que consiste na utilização das duas primeiras técnicas.

A indexação manual requer um esforço intelectual muito intensa do indexador que, por sua vez, acaba deixando uma carga de subjetividade na escolha dos termos e em todo processo de indexação de um documento. Assim, a literatura aponta para alguns problemas existentes na área, além da morosidade causada pelo tempo limitado que o profissional tem para fazer análise de um documento com qualidade para indexação. Lancaster (2004) aponta que o indexador não dispõe de tempo suficiente, sendo rara a leitura de um documento do começo ao fim. Assim, Borges (2009, p 16) destaca alguns problemas práticos da indexação manual:

> (1) quando diferentes indexadores atribuem diferentes termos a um mesmo documento; (2) quando o mesmo indexador atribui diferentes termos a um mesmo documento, em momentos distintos; (3) o conhecimento do indexador sobre o assunto tratado, que influenciará no nível de consistência atingido na atividade; (4) as

> possibilidades do indexador em acompanhar a dinamicidade do conhecimento e (5) a capacidade de compreensão do idioma do documento tratado.

Esses problemas podem ser minimizados quando a indexação é feita automaticamente, pois os recursos tecnológicos permitem que a representação de um documento seja feita de forma mais rápida, tendo uma redução de tempo na indexação.

Vieira (1988b) define indexação automática como a realização da tarefa diretamente por um sistema de computador que faz a análise do texto, o reconhecimento e a construção de índices para a recuperação do texto em pesquisas.

Hjørland (2008) percebe a indexação automática como um procedimento feito por meio de algoritmos. Para Andreewski (1983) é a técnica de processamento eletrônico de documentos que tem como intuito recuperá-los por meio de informações referentes ao seu conteúdo. Já a definição de Farmer (2006) se refere à tecnologia que tem como objetivo lidar com o grande volume de documentos digitais que não estão estruturados, não indexados e não ordenados, sendo utilizada em conjunto com taxonomias e metadados no intuito de se obter um melhor desempenho das ferramentas de busca.

Os estudos sobre essa modalidade de indexação surgiram a partir da década de 1950 e nesse início a indexação automática estava estritamente relacionada a textos escritos digitalmente, não abrangendo outros gêneros como, por exemplo, vídeos, sons e imagens, (VIEIRA, 1988a;

CAMARA JÚNIOR, 2007; BORGES, 2009). Esses estudos estão direcionados a indicar critérios que aumentem a eficiência das ferramentas de indexação automática. Hjørland (2008) e Lancaster (2004) se referem a Hans Peter Luhn como o autor do primeiro método de indexação automática. Esse método considerava a frequência das palavras dos títulos dos documentos, no intuito de se criar um índice rotativo. As palavras consideradas importantes eram as que tinham uma frequência média, sendo descartadas as palavras vazias como os artigos, preposições e conjunções.

Podem-se citar alguns tipos de indexação automática, pois nenhuma técnica até agora existente foi capaz de resolver todos os problemas relacionados à indexação automática. Cada tipo de indexação atua de maneira diferente nesse processo, atuando nos problemas não resolvidos por outros tipos de indexação automática. Entre eles, podem-se citar, baseando-se em Borges (2009), Lancaster (2004) e Wives (1997):

1. **indexação por extração automática**: o conteúdo do documento é representado por palavras e expressões que foram extraídas desse documento. Segundo Lancaster (2004) esse procedimento remete aos mesmos princípios utilizados pela extração manual, pois pode ser feito observando a frequência da palavra dentro do texto, posição dessa palavra no texto (no título, nas palavras-chave, no resumo, etc.) e seu contexto. As atividades desenvolvidas por esse tipo de indexação são

contar as palavras em um documento textual, cotejar essas palavras com uma lista de vocábulos predefinidos e considerados proibidos, não considerar as palavras não significativas, chamadas também de *stopwords,* colocar em ordem as palavras de acordo com sua frequência. Borges (2009) percebe que indexação por extração automática, por ser baseada estritamente em critérios estatísticos, é limitada na realização do processo consistentemente. Dessa maneira, os termos de indexação são encontrados no próprio documento analisado;

2. **indexação por atribuição semântica**: trabalha da mesma maneira que indexação por extração automática, contudo faz uso de um controle terminológico. Assim, desenvolve para cada termo atribuído um perfil de palavras ou expressões nos documentos. Nesse tipo de indexação, os termos de representação do documento são atribuídos independentemente da ocorrência dos termos no documento;
3. **identificação automática de palavras *full text***: o texto é analisado completamente, não sendo levadas em consideração a semântica nem a posição sintática das palavras na sentença;
4. **indexação automática sintática**: para Wives (1997) esse tipo de indexação observa as palavras mais relevantes da oração, para isso,

atenta para as posições sintáticas pré-definidas de termos como sujeito, predicado, local do verbo, entre outros, pois, para o autor, alguns destes termos são mais importantes do que outros, desse modo, somente os termos mais importantes são incluídos no índice;

5. **indexação automática semântica**: identifica as marcações que indicam onde estão os títulos, as palavras-chave e outras estruturas importantes de um documento e identifica os termos mais relevantes dentro dessas marcações. Essas marcações podem ser encontradas em documentos em HTML, por exemplo. Dessa maneira, como afirma Borges (2009), esse tipo de indexação é baseada no princípio de que o documento já possui estrutura de formação no intuito de indicar a semântica dos termos.

Maia (2008) afirma que o objetivo da indexação é representar os conteúdos de um documento, produzindo uma lista de descritores. Para se ter bons descritores, é necessário que esses tragam consigo informações que permitam fazer as relações entre os objetos extralinguísticos e o documento que traz informações sobre esses objetos.

Como aponta Kuramoto (2002), os SRIs convencionais usam termos isolados, o que acarreta problemas linguísticos relacionados à polissemia, sinonímia e a combinação de palavras que em ordens diferentes apresentam significados também diferentes. Perini (1996) já apontava que os SNs

eram mais eficientes na classificação e recuperação da informação e que o uso dos lexemas (palavras) não era interessante.

Miorelli (2001) argumenta que os documentos possuem uma enorme variedade de frases que podem ser usadas na construção de um índice, mas é necessário selecionar um conjunto delas. Assim, os SNs podem ser usados no processo de indexação substituindo as palavras isoladas, objetivando a construção de um índice hierárquico o qual poderia ser usado pelo usuário em uma interface de busca. Essa ideia foi proposta por Kuramoto (2006, p. 128):

> A primeira seria implementar uma indexação automática nos moldes daquela tradicional baseada em palavras, substituindo os índices das palavras isoladas por índices dos sintagmas nominais... Uma segunda alternativa seria o aproveitamento da organização hierárquica, em árvore, dos sintagmas nominais.

Contudo, Kuramoto (2002) ressalta a dificuldade nas pesquisas em relação à extração de SNs, principalmente quando se envolvia um grande volume de textos. Os problemas eram mais de ordem econômica e a inexistência de ferramentas que fizessem a extração em textos em língua portuguesa. Percebe-se que tal argumento motivou trabalhos como os de Miorelli (2001), Souza (2005), Morellato (2007; 2010), Maia (2008), entre outros. Mas já em 2000, Bick defendia sua tese, apresentado uma ferramenta que identifica SNs para textos escritos em português, o *parser* PALAVRAS.

2.2 Processamento de Linguagem Natural

Processamento de linguagem natural (PLN) é uma área de pesquisa dentro da Ciência da Computação que trata da comunicação humana computacionalmente. Percebe-se como linguagem natural o conjunto infinito de frases, as quais são construídas pelas relações lexicais estabelecidas pelos fonemas, morfemas, sintagmas e as questões semânticas, pragmáticas e discursivas. Os aspectos estudados no PLN são os textos produzidos em diversos gêneros textuais e diferentes linguagens (verbais, não verbais, mista, matemática). Esses textos são representados por sons, palavras, sentenças e discursos.

Comarella e Café (2008, p.56) trazem um bom conceito de PLN, além de defini-lo como a habilidade de uma máquina em processar a linguagem humana. Também o define como área de pesquisa que objetiva produzir ferramentas para essa tal ação.

> O processamento de linguagem natural pode ser definido como a habilidade de um computador em processar a mesma linguagem que os humanos usam no dia-a-dia. O PLN visa, também, produzir ferramentas que compreendam a língua, avaliar o impacto da aplicação de tecnologias relacionadas com a língua, etc.

Para Barros e Robin (2001), o PLN tem por objetivo interpretar e gerar textos em linguagens naturais, e que tem característica multidisciplinar, pois agrega estudos da Ciência da Computação, Linguística e Ciências Cognitivas.

No contexto de RI, segundo Pinheiro (2009, p. 04), o PLN tem como principal foco a etapa de pré-processamento que é constituído de fases como “seleção e filtragem de dados, limpeza de dados, normalização e *parsing*, análise semântica e representação numérica dos termos extraídos do documento em um vetor no espaço vetorial (BOW – *Bag-of-Words*)”. Gonzalez e Lima (2003, p. 03) afirmam que “o PLN visa fazer o computador se comunicar em linguagem humana, nem sempre necessariamente em todos os níveis de entendimento e/ou geração de sons, palavras, sentenças e discurso”.

Segundo esses mesmos autores, os níveis linguísticos são: fonético e fonológico; morfológico; sintático; semântico; e pragmático. O foco dessa pesquisa está nos níveis sintático e morfológico, pois o primeiro trabalha a relação que as palavras estabelecem entre si, em que cada uma exerce seu papel estrutural nas frases e como as frases podem ser partes de outras, formando sentenças, e a segunda estuda os morfemas que são as partes que compõem as palavras e suas derivações, todavia, não se nega a importância dos outros níveis na análise e extração de informações de um texto em linguagem natural.

Ressalta-se que, no PLN, a estrutura sintática de uma sentença é obtida também pelas relações morfológicas, que percebe a construção das palavras por meio de unidades de significação e por classificação categóricas das mesmas, tendo, assim, um processamento morfossintático representado por normas gramaticais, ou seja, regido por leis gramaticais.

Quando se tenta interpretar a linguagem natural artificialmente, buscam-se mecanismos que compreendam textos nessa linguagem, traduzindo-os para uma representação que possa ser compreendida e utilizada pelo computador (BARROS; ROBIN, 2001). Assim, se o computador conseguir extrair descritores relevantes para um documento, poupar-se-ia o tempo do usuário na busca de documentos que supram sua necessidade informacional, como já apontava a Quarta Lei da Biblioteconomia, “Poupe o tempo do leitor” (RANGANATHAN, 2009).

Vieira e Lima (2001) explicam que o PLN é a construção de programas que interpretam e/ou produzem informação fornecida em linguagem natural.

O PLN tenta reproduzir, compreender, interpretar e apresentar informações na linguagem que os seres humanos usam no cotidiano, usando leis gramaticais e observando as relações das palavras com suas categorias gramaticais e seus significados.

Pode-se dizer ainda que os métodos estatísticos têm contribuído para os estudos do PLN, pois juntamente com a teoria da probabilidade, têm indicado meios para o processamento de significados dentro das ferramentas de PLN como os *taggers* e os *parsers* que serão expostas a seguir.

Podem-se citar como exemplos de métodos estatísticos: a lei de Zipf que trabalha a frequência de ocorrência dos termos de um determinado texto; e o gráfico de Luhn que considera as palavras com mais frequência de ocorrência e as que aparecem raramente como menos

expressivas, devendo ser consideradas apenas as palavras com ocorrência intermediária (GONZALEZ; LIMA, 2003).

Já a teoria da probabilidade permite a geração de modelos matemáticos para mensurar distribuição, frequência, esperança e variância. Assim, como diz Krenn (1997 apud GONZALEZ; LIMA, 2003) a estatística tem sido utilizada na etiquetagem gramatical, na resolução de ambiguidade e na aquisição de conhecimento lexical.

Pode-se dizer que os métodos estatísticos ainda subsidiam diversas abordagens do PLN; desse modo, lança-se mão dos estudos linguísticos e dos estudos matemáticos para compor uma linguagem mista no intuito de fazer com que a máquina se comunique em linguagem natural, apresentando resultados de análise que auxiliam na busca feita pelo usuário.

2.2.1 Etiquetador Morfossintático (*Tagger*)

O reconhecimento das categorias gramaticais é um dos objetivos do PLN, por isso a análise morfossintática se faz necessária. A morfologia analisa a estrutura das palavras por meio dos morfemas com suas regras de formação e inflexão. Em um sistema de computador, a morfologia permite que se armazenem os radicais das palavras e suas formas flexionadas pela aplicação das regras de formação da palavra. Já a sintaxe estuda as leis que regem a formação de uma sentença de uma determinada língua, apresentando a estrutura sintática da frase. As regras sintáticas determinam a

linearidade e a hierarquia que existem entre os constituintes de uma frase, com base na sua categoria sintática (obrigatória e facultativa).

Gonzalez e Lima (2003) dizem que tanto a morfologia quanto a sintaxe estudam a constituição das palavras e dos grupos de vocábulos que formam os elementos de expressão de uma língua. Esse conceito é construído a partir de uma visão panorâmica das funções das duas áreas dentro de um programa de computador.

Dentro das estratégias do PLN para enfrentar a ambiguidade de termos e buscar os significados dos léxicos, entre outros objetivos, está o etiquetador gramatical (*tagger*) que na língua inglesa é denominado de "*part-of-speech tagger*". O *tagger*, segundo Bick (1998), é um sistema que tem como meta identificar, por meio de uma *tag* (etiqueta) a categoria gramatical de cada léxico do texto analisado.

Pinheiro (2009) aponta que um *tagger* marca, anota e rotula morfossintaticamente um texto escrito em uma determinada língua. Dessa maneira, marcam-se as palavras, os símbolos de pontuação, estrangeirismos e fórmulas matemáticas existentes dentro de um texto de acordo com seu contexto. Assim, um *tagger* pode ser definido, segundo Morellato (2010, p. 52), como um *software* que realiza "o processo de encontrar uma etiqueta, marcar com uma etiqueta cada uma das palavras de um texto baseado em sua definição, assim como em seu contexto".

Em outras palavras, os autores querem dizer que é uma forma de categorizar gramaticalmente a palavra, levando em consideração sua natureza e sua relação com os demais

vocábulos adjacentes a ela na sentença, ou seja, é classificar as palavras em substantivos, adjetivos, pronomes, artigos, numerais, advérbios, verbos, conjunções, interjeições, preposições, assim como as derivações dessas classes gramaticais.

Há um fenômeno que torna mais difícil a tarefa de categorizar gramaticalmente as palavras: a ambiguidade, muito comum em linguagem natural. Sua ação acontece quando uma palavra pode representar mais de uma classe em diferentes contextos.

A etiquetagem consiste em uma descrição detalhada das categorias lexicais de forma que se observam as derivações e as inflexões das palavras.

Podem-se citar como exemplos de etiquetadores: o FORMA que é um programa de código aberto que apresenta bons resultados de precisão porque é composto por uma ferramenta probabilística e por base de dados constituída de um corpus de treino de palavras etiquetadas e com lemas definidos (MORELLATO, 2010); o *parser* PALAVRAS[5] (BICK, 2000) que é de alta precisão e é usado em várias pesquisas, apesar de ser um *parser*, ele também faz etiquetagem; o SISNOP (MORELLATO, 2010) que tem um conjunto de etiquetas que possuem relação com as categorias gramaticais do *tagger*, o que permite o uso de mais etiquetadores, igualando as categorias dos *taggers* e fazendo adaptações, caso sejam necessárias. Podem-se fazer, ainda, referências

[5] VISUAL INTERACTIVE SYNTAX LEARNING – VISL. Disponível: <http://beta.visl.sdu.dk/> Acesso em: 10 jun. 2013.

ao OGMA (MAIA, 2008) que faz etiquetagem morfológica e lexical dos vocábulos, os *taggers* do NILC do projeto Lacio-WEB[6] que fazem a marcação em corpora de textos brasileiros, compostos de textos literários, jornalísticos e didáticos. Miorelli (2001) também propõe etiquetas de palavras no ED-CER. Tomando como exemplo as etiquetas estabelecidas por Miorelli (2001) e Maia (2008), tem-se a Tabela 1:

Tabela 1 - Etiquetas utilizadas no OGMA e ED-CER

ETIQUETA OGMA	ETIQUETA ED-CER	CLASSE GRAMATICAL
AD	AR	Artigo definido
AI	AR	Artigo indefinido
AJ	AJ	Adjetivo
AV	AV	Advérbio
CJ	CO	Conjunção (Aditiva, adversativa, alternativa)
IT	*	Interjeição
NC	NU	Números cardinais
NM	NU	Números multiplicativos
NO	NU	Números ordinais
NP	NP	Nome próprio
NR	NU	Número romano
PS	PS	Pronome possessivo
PD	PD	Pronome demonstrativo
PI	PI	Pronome indefinido
PL	LG	Pronome relativo
PN	PN	Pontuação (exceto virgula)
PP	PP	Pronome pessoal
PR	PR	Preposições
SU	SU	Substantivo
VB	VB	Verbos
VG	CO	Vírgulas
VP	AJ	Verbos no Particípio

Fonte: Maia (2008, p. 64)

[6] NÚCLEO INTERINSTITUCIONAL DE LINGUÍSTICA COMPUTACIONAL. Disponível em: <http://www.nilc.icmc.usp.br/nilc/tools/nilctaggers.html>. Acesso em: 13 abr. 2013.

Percebe-se que cada trabalho, muitas vezes, apresenta um conjunto de etiquetas diferentes (*Tagset*). Isso acontece em todos os trabalhos que objetivam fazer a extração de SNs. Contudo, a maior parte dos trabalhos que, nessa pesquisa, constituem o estado da arte utiliza a etiquetagem das palavras feitas pelo *parser* PALAVRAS (BICK, 2000). No Quadro 1, podem ser vistas algumas etiquetas morfológicas utilizadas pelo FORMA, SISNOP e do PALAVRAS.

Quadro 1 - Relacionamento entre as classes gramaticais do FORMA, do PALAVRAS e do SISNOP

Etiqueta SISNOP	Etiqueta FORMA	Etiqueta PALAVRAS
adjetivo	adjetivo (AJ) verbo no particípio (AP)	adjetivo (ADJ) verbo no particípio (V PCP)
advérbio	advérbio (AV)	advérbio (ADV)
artigo	definido (AD) e indefinido (AI)	determinante (DET-artd)
conjunção	coordenativa (CC) subordinativa (CS)	coordenativa (KC) subordinativa (KS)
nome	substantivo (SU)	substantivo (N) nome próprio(PROP)
verbo	verbos auxiliar (VA) regulares e irregulares (VB)	verbo(V)
preposição	preposição (PR)	preposição (PRP)
prn_pessoal	pronome pessoal (PP)	pronome pessoal (PERS)
prn_demonstrativo	pronome demonstrativo (PD)	demonstrativo (SPEC-dem)
prn_indefinido	pronome indefinido (PI)	quantitavivo (SPEC-quant)
prn_possessivo	pronome possessivo (PS)	possessivo (SPEC-poss)
prn_interrogativo	pronome relativo (PL)	interrogativo (SPEC-interr)
virgula	vírgula, parênteses e traços (VG)	-
ponto	pontuações (PN)	pontuação (PU)

Fonte: Morellato (2010, p. 71)

A Tabela 2 apresenta as etiquetas do NILC do projeto Lacio-WEB que parecem ser mais fáceis de associar às suas respectivas classes gramaticais. Vale ressaltar que existem alguns padrões internacionais de etiquetagem como Brown e o TigerXML.

Tabela 2 - **NILC TAGSET5 – 36 tags**

CLASSE GRAMATICAL	ETIQUETAS
Adjetivo	ADJ
Advérbio	ADV
Artigo	ART
Número Cardinal	NC
Número Ordinal	ORD
Outros Números	NO
Substantivo Comum	N
Nome Próprio	NP
Conj. Coordenativa	CONJCOORD
Conj. Subordinativa	CONJSUB
Pronome Demonstrativo	PD
Pronome Indefinido	PIND
Pronome Oblíquo Átono	PPOA
Pronome Pessoal Caso Reto	PPR
Pronome Possessivo	PPS
Pronome Relativo	PR
Pronome Oblíquo Tônico	PPOT
Pronome Interrogativo	PINT
Pronome Apassivador	PAPASS
Pronome de Realce	PREAL
Pronome Tratamento	PTRA
Preposição	PREP
Verbo Auxiliar	VAUX
Verbo de Ligação	VLIG
Verbo Intransitivo	VINT
Verbo Transitivo Direto	VTD
Verbo Transitivo Indireto	VTI
Verbo Bitransitivo	VBI
Interjeição	I
Locução Adverbial	LADV
Locução Conjuncional	LCONJ
Locução Prepositiva	LPREP
Locução Pronominal	LP
Palavra Denotativa	PDEN
Locução Denotativa	LDEN
Palavras ou Símbolos Residuais	RES

Fonte: adaptada de NILC (2013)

Na Tabela 3, é apresentada a análise manual da frase “A extração de sintagmas nominais em teses e dissertações”, usando as etiquetas das referidas ferramentas com exceção do SISNOP (já que essa usa o nome da classe gramatical por extenso).

Tabela 3 – Diferenças dos etiquetadores morfossintáticos

Ferramenta	SN								
	A	extração	de	sintagmas	nominais	em	teses	e	dissertações
OGMA	AD	SU	PR	SU	AJ	PR	SU	CJ	SU
ED-CER	AR	SU	PR	SU	AJ	PR	SU	CO	SU
NILC	ART	N	PREP	N	ADJ	PREP	N	CONJ COORD	N
FORMA	AD	SU	PR	SU	AJ	PR	SU	CC	SU
PALAVRAS	DET-artd	N	PRP	N	ADJ	PRP	N	KC	N

Fonte: o autor

Percebe-se que as etiquetas do NILC TAGSET5 se assemelham às do PALAVRAS com exceção da etiqueta que se refere à conjunção e ao artigo. As etiquetas que se referem à conjunção se diferenciam em todas as ferramentas, pois são classificadas em conjunções coordenativas e subordinativas, no caso do exemplo, essa categoria gramatical se comporta como coordenativa, tendo uma etiqueta diferente em cada ferramenta. Em relação ao artigo, o PALAVRAS o classifica em categoria geral, “DET” (determinante), especificando-o, depois, com “art” e seguido da indicação de direto, “d” – artd.

O OGMA, o ED-CER e o FORMA também associam etiquetas semelhantes, mas se diferenciam quanto à etiqueta referente à conjunção. Não se podem generalizar essas observações a todos as categorias, contudo quer-se, com essa demonstração, mostrar que os trabalhos têm uma relação entre si. Por exemplo, o OGMA (MAIA, 2008) se baseia no ED-CER (MIORELLI, 2001) para estabelecer suas etiquetas. Morelatto (2010), com o SISNOP, baseia-se no FORMA para construir também as representações das marcações morfológicas de um texto. O processo de marcação das palavras com suas respectivas categorias gramaticais é a primeira etapa para a identificação dos SNs, pois é através da combinação dessas classes que se tem a estrutura de um sintagma.

2.2.2 Analisador Sintático (*Parser*)

A sintaxe trabalha com as leis que regem a estrutura de uma frase em uma língua, ou seja, ela estuda as relações entre os constituintes de uma frase, que podem ser chamados de sintagmas que, por sua vez, podem ser conceituados como grupos de elementos linguísticos que estão ligados a um núcleo como substantivo, verbo, preposição, adjetivo, advérbio, entre outros derivados desses elementos.

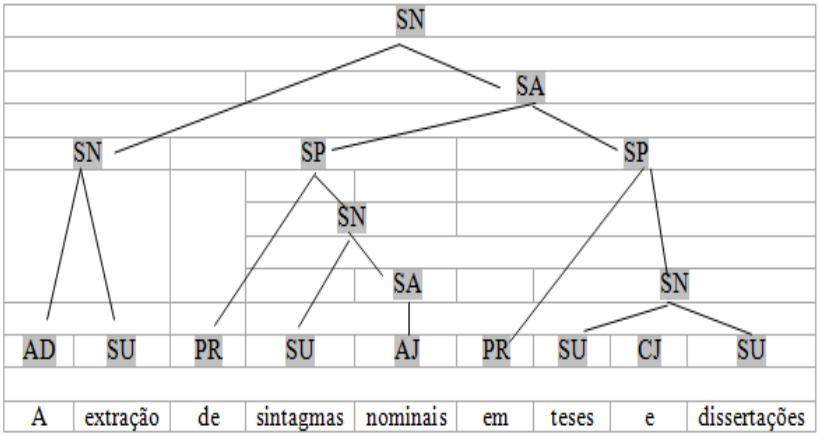
O analisador sintático avalia o agrupamento das palavras, trabalhando a estrutura da frase e é denominado de *parser*, que pode ser entendido como um *software* que tem por finalidade mapear uma sentença, utilizando o léxico e a

gramática do sistema (BARROS; ROBIN, 2001). Beardon (1991 apud MIORELLI, 2001) diz que é um programa que quando fornecido um léxico, uma gramática e um texto, determina se uma sentença é gramaticalmente correta ou não.

Para Vieira e Lima (2001), *parsers* são sistemas que analisam a estrutura das frases e seus constituintes. Opta-se por esse conceito quando se usa o termo sistema para definir *parsers,* pois dá ideia de hiperônimo, uma classe que agrupa várias subclasses. Ainda, para Vieira e Lima (2001, p. 06), "esses sistemas reconhecem estruturas válidas a partir de um léxico que define o vocabulário da língua e um conjunto de regras que definem a gramática da língua".

A análise sintática, chamada de *parsing*, observa as maneiras em que as regras gramaticais e sua combinação dentro de uma sentença se realizam, no intuito de construir uma estrutura arbórea que possa representar a formação sintática da frase analisada. Barros e Robin (2001) mencionam, ainda, que essa representação pode ser feita pelos constituintes da frase agrupados com colchetes. Usando como exemplo a frase "A extração de sintagmas nominais em teses e dissertações" acrescido das etiquetas do OGMA, em representação arbórea criada manualmente, tem-se a Figura 1:

Figura 1 – Representação arbórea de uma análise sintática



Fonte: o autor

A frase analisada na Figura 1 não tem uma estrutura completa, possuindo somente um SN, esse, por sua vez, constituído de sintagma preposicionado (SP) e sintagma adjetival (SA).

Desse modo, Gonzalez e Lima (2003) ressaltam que a análise sintática é muito importante, pois viabiliza o processamento semântico. Para tanto, os *parsers* podem trabalhar com alguns métodos de análise como o *top-down*, *bottom-up*, *left-corner* e tabular.

O *parser top-down* faz análise associado à linguagem de programação Prolog[7], a partir da esquerda para direita,

[7] O Prolog é uma linguagem de programação lógica bastante popular no desenvolvimento de parsers para linguagens naturais, devido seu suporte nativo para buscas e para a operação de unificação, que combina duas estruturas características em uma. (MORELLATO, 2010, p. 25)

respondendo com uma afirmação ou negação sobre a validade da sentença na consulta do usuário, identificando sua estrutura sintática e associando argumentos aos constituintes representados, o que pode ocasionar, como citam Vieira e Lima (2001, p. 18), a transformação em outra gramática, pois “é sabido que qualquer gramática recursiva à esquerda pode ser transformada em outra gramática que gera a mesma cadeia de palavras, mas não é recursiva à esquerda”.

Em se tratando de uma representação arbórea em que a construção representativa da sentença é descendente, ou seja, a partir da raiz, o *parser top-down* começa a análise pelas regras iniciais e desce para os símbolos das sentenças que são as palavras.

Nas palavras de Morellato (2007), esse *parser* encontra o conceito frase e verifica as regras das quais é composto. Desse modo, “a análise de regras feita da esquerda para a direita irá descer pelo ramo mais à esquerda da árvore até encontrar um elemento terminal que a satisfaça” (MORELLATO, 2007, p. 18).

Assim, ao optar por essas regras, produzir-se-á uma estrutura que não está de acordo com as leis gramaticais de uma língua. Quando essa análise é aplicada a uma gramática com produções recursivas à esquerda, pode fazer o *parser* entrar em *loop* (ciclo), não permitindo o término da análise da frase. Esse *parser* permite uma redundância na análise de dados, quando é mister verificar as regras mais de uma vez (VIEIRA; LIMA, 2001; MORELLATO, 2007). Dessa maneira,

para manter a correta estrutura gerada, é necessário mudar o analisador, passando para o *bottom-up*.

O *bottom-up* codifica os léxicos no intuito de combiná-los. Ao encontrar uma palavra, ele a reconhece, encontra a próxima que juntos formam um sintagma (VIEIRA; LIMA, 2001), ou seja, o *parser bottom-up* analisa a frase a partir das folhas da árvore de derivação e tenta montá-la combinando as regras para chegar a uma estrutura sintática, (BARROS; ROBIN, 2001; MORELLATO, 2007), não tendo problemas com *loop* para regras recursivas à esquerda.

Porém, ao trabalhar dessa maneira, o analisador pode se deparar com constituintes vazios (por exemplo, supressão de pronomes) e constituintes facultativos (por exemplo, a ênfase da sentença, adjuntos) com os quais não consegue lidar, o que é muito comum na língua portuguesa. Constituintes são grupos de palavras que fazem parte de sequências maiores, mas que estão coesas entre si (PERINI, 1995).

Combinando as estratégias dos dois analisadores anteriores (*top-down* e *bottom-up*), o *left-corner*, ao encontrar um vocábulo, observa a que categoria de constituintes se inicia tal palavra e lida com as regras recursivas à esquerda, de modo que as regras são exploradas uma de cada vez (VIEIRA; LIMA, 2001), não tendo problemas com *loop* (MORELLATO, 2007).

Já o analisador tabular, chamado também de *chart parser*, atenta para as subestruturas já analisadas e se um retrocesso for mister, a repetição pode ser evitada. Pois,

guarda as estruturas que já foram analisadas e as reutiliza, se for necessário.

Pode-se concluir nessa seção, fazendo referência a Comarella e Café (2008, p. 64) quando citam Rich, que na análise sintática,

> constroem-se árvores com a derivação de cada sentença, objetivando relacionar as palavras entre si e verificar a adequação das seqüências de palavras às regras de construção da linguagem, tais como a concordância verbal e a regência nominal, bem como o posicionamento de termos na frase. Ou seja, a análise sintática procura identificar os relacionamentos sintáticos entre as seqüências lineares de palavras, produzindo a estrutura sintática da sentença (identifica, por exemplo, qual o sujeito, o predicado, o objeto direto, etc.). Normalmente utilizam se para a análise sintática abordagens baseadas em Gramáticas, ou seja, em um conjunto de regras que descrevem quais sentenças fazem parte de uma determinada linguagem. Estas regras estipulam se uma sentença (S) é sintagma nominal (SN) e/ou um sintagma verbal (SV), possibilitando a construção da estrutura em árvore mostrando sem ambigüidade como as palavras interagem numa sentença.

Existem diversos analisadores sintáticos, podendo ser citados entre outros: o *Nptool* de Voutilainen (1995) que extrai SNs em textos escritos em língua inglesa, no intuito de reconhecer termos representativos na construção do índice do texto (esse programa não está mais disponível no mercado, há apenas atualizações para quem já o adquiriu); o *Multi-Stage* do Projeto CLARIT (EVANS, 1991) é um *parser* que objetiva ser rápido na identificação de sintagmas nominais em grandes corpora. Há um *parser* chamado de Curupira que é

um produto derivado de um revisor gramatical e é constituído de um conjunto de regras que geram todas as análises sintáticas possíveis para uma dada sentença em português brasileiro (MARTINS; HASEGAWA; NUNES, 2003). Há também outros *parsers* como O *parser* PALAVRAS (BICK, 2000) e o LX-*parser* que analisam sintaticamente textos em língua portuguesa. As funções desses *parsers* serão descritos na seção “Identificador ou Extrator de Sintagmas Nominais (SNs)”.

2.3 Sintagmas Nominais (SNs)

Nesta seção será descrita a teoria da Gramática Gerativa, a qual estuda as estruturas sintáticas a partir dos agrupamentos dos constituintes morfológicos que, por sua vez, formam os sintagmas. Também serão abordadas a extração automática de SNs, algumas ferramentas de extração automática desses termos, algumas metodologias de seleção de SNs e metodologias de avaliação de extração automática de SNs.

2.3.1 Gramática Gerativa

Em 1957, Noam Chomsky, professor do Instituto de Tecnologia de Massachussets publica o livro *Estruturas*

Sintáticas[8] que traz uma nova percepção de como a linguagem deve ser analisada, ocorrendo o foco no cognitivo do falante e percebendo a linguagem como componente criativo do ser humano. Nessa concepção, a natureza da linguagem está ligada a estrutura biológica humana, assim, a experiência com outros indivíduos estimula a faculdade da linguagem, ou seja, como diz Martelotta (2012, p. 58), a linguagem é vista pelos gerativistas como "reflexo de um conjunto de princípios inatos – e, portanto universais – referentes à estrutura gramatical das línguas".

Em Kenedy (2012), percebe-se que o gerativismo passou por várias modificações e reformulações ao longo desses 50 anos, no intuito de elaborar um modelo teórico formal que, por sua vez, é inspirado nos modelos matemáticos, assim, é-se capaz de descrever e explicar, de maneira abstrata, o que é e como funciona a linguagem humana.

Como toda corrente ou movimento teórico vem criticar os modelos vigentes, a gramática gerativa rejeita o modelo behaviorista que dominava as ciências em geral, que dentro da linguística estava relacionado ao estruturalismo. A linguagem era vista como um condicionante social, ou seja, a linguagem era produzida mediante os estímulos recebidos da interação social[9].

---

[8] CHOMSKY, Noam. **Syntact Structures**. Haia: Mouton, 1957.

[9] Pode parecer meio contraditório o fato dessa pesquisa se desenvolver no âmbito de uma ciência social e usar uma corrente que vê essa questão em segundo plano. Isso pode ser explicado quando se vê que aplicação das

Para Borges Neto (2011), Chomsky percebe a necessidade de se supor a existência de algo que antecede à língua dos estruturalistas, em outras palavras, as regras que regem os corpora representativos, “a capacidade que os falantes têm de produzir exatamente os enunciados que *podem* ser feitos” (BORGES NETO, p. 99). [destaque do autor].

O gerativismo pretende construir um mecanismo computacional que formalize e transforme representações e que, segundo Borges Neto (2011, p.97), “‘simule’ o conhecimento linguístico de um falante de uma língua natural, ‘registrado’ em sua mente/cérebro”. Sistema computacional é visto como hipóteses explicativas e suas consequências empíricas que, por sua vez, devem ser observadas de maneira dedutiva.

Pode-se conceituar a gramática gerativa como um programa de investigação que pretende construir um modelo para perceber como se constrói esse conhecimento linguístico.

Chomsky (1965) precisava de uma ferramenta que descrevesse os fenômenos das línguas naturais e explicasse as suas formações estruturais, o sistema computacional é visto por esse teórico como a implementação da gramática gerativa que deve dar conta de boa formação de uma língua

---

regras do objeto de pesquisa em um sistema automatizado depende de meios exatos para que funcione, e a percepção da linguagem baseada no cognitivismo expresso por modelos matemáticos torna o desenvolvimento da reprodução da linguagem natural pela máquina mais palpável.

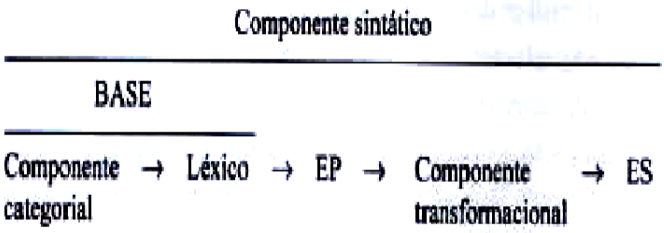
qualquer. Assim, ele reconhece seis níveis de descrição linguística: fonemas, morfemas, palavras, categorias sintáticas, estrutura frasal e transformações.

Na teoria do gerativismo, a semântica e a fonologia são componentes interpretativos, enquanto a sintaxe é o componente que gera representações. Dessa maneira, a sentença é formada inicialmente pelos componentes sintáticos que são constituídos de uma base responsável pela geração das estruturas profundas e pelos componentes transformacionais que convertem essas estruturas profundas em superficiais.

A base é construída a partir de componentes categoriais, os quais quando aplicados ao axioma, no primeiro momento, geram estruturas arbóreas etiquetadas. A base também contém um léxico que insere itens lexicais nos terminais das árvores. Assim, a entrada e saída da base são as estruturas profundas.

No processo de mutação, o componente transformacional recebe as estruturas profundas, como entrada, e as converte em superficiais através de leis transformacionais, como podem ser vistas na Figura 2, em que EP é estrutura profunda (voz ativa do verbo) e ES é estrutura superficial (voz passiva do verbo), focalizando-se no componente sintático. A estrutura superficial está relacionada à forma física de realização concreta da oração e à interpretação fonológica (SILVA; KOCH, 2009), enquanto a estrutura profunda se relaciona a representação da frase abstratamente, na qual se observa as relações semânticas existentes entre os itens lexicais (HOUAISS, 2001).

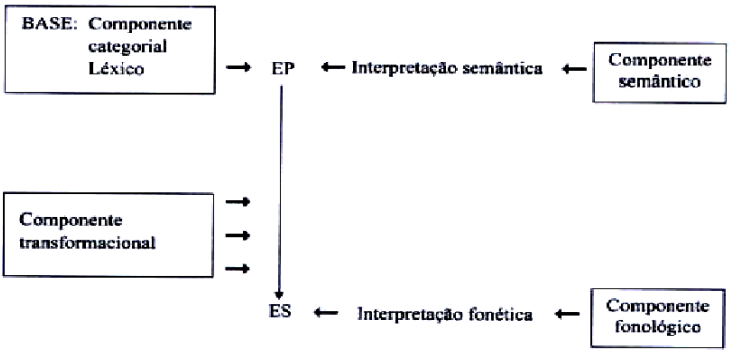
Figura 2 - Transformação de uma EP em ES



Fonte: Borges Neto (2011, p. 112)

Na Figura 3, há o esquema completo de geração de sentença, a EP recebe as relações das interpretações quanto ao componente semântico, já a ES associa as interpretações quanto ao componente fonológico.

Figura 3 - Esquema do Processo de Geração de Sentença



Fonte: Borges Neto (2011, p. 112)

Assim,

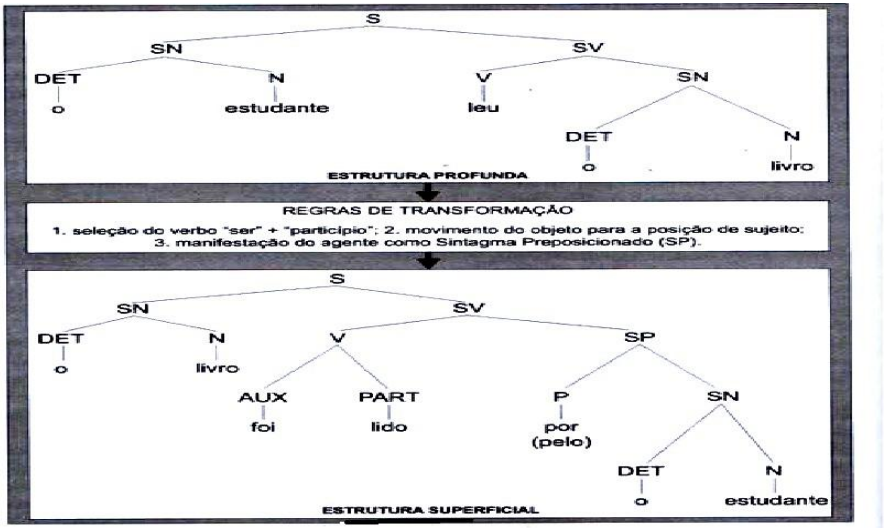
> O componente sintático gera pares ordenados <EP, ES> e os dois componentes interpretativos associam representações aos elementos dos pares: o componente semântico associa interpretações semânticas às EPs, e o componente fonológico associa interpretações fonéticas às ESs. [...] A EP deve conter todos os elementos necessários para interpretação semântica de sentença enquanto a ES deve conter as transformações para a sua leitura fonética. (BORGES NETO, 2011, p. 112)

Essas eram as ideias inicias acerca da gramática transformacional dentro do gerativismo. A descrição de como os constituintes das sentenças se formavam e como eles se transformavam em outros, por meio de aplicação de regras, era o intuito da gramática transformacional. Nas palavras de Kenedy (2012), tem-se a concepção do que vem a ser essa gramática, assim como seu contexto de formação:

> Os gerativistas perceberam que as infinitas sentenças de uma língua eram formadas a partir da aplicação de um infinito sistema de regras (a gramática) que transformava uma estrutura em outra (sentença ativa em sentença passiva, declarativa em interrogativa, afirmativa em negativa, etc.) – e é precisamente esse sistema de regras que, então, se assumia como o conhecimento linguístico existente na mente do falante de uma língua, o qual deveria ser descrito e explicado pelo linguista gerativista. (KENEDY, p. 131)

Na Figura 4, há uma representação da transformação de uma sentença em EP em ES com as regras específicas para essa ação.

Figura 4 - Regras de transformação de EP em ES



Fonte: Kenedy (2012, p. 132)

A regra de transformação pode ser descrita da seguinte maneira, tomando como exemplo a sentença “O estudante leu o livro”: seleciona-se o verbo auxiliar “ser”, no caso da sentença, conjugado na terceira pessoa do modo indicativo no pretérito perfeito “foi” mais o verbo no particípio “lido”, o objeto “livro” é alocado na posição de sujeito, enquanto esse é alocado para a função de objeto, representado pelo SP com a introdução da preposição por (pelo).

A gramática gerativa, resumidamente, é um conjunto de teorias que se preocupam com o dispositivo mental inato do falante. Há vários modelos na gramática gerativa, sendo os mais importantes para essa pesquisa os estados finitos, o transformacional e o sintagmático. Enquanto os dois primeiros

versarão sobre os conceitos da descrição da gramática, o sintagmático regerá a discussão na concepção da descrição das regras de aplicação para extração de SNs.

O modelo dos estados finitos trabalha aplicando regras recursivas sobre um vocabulário finito, assim, as frases são geradas por um conjunto de escolhas feitas pelo formulador da frase. No modelo sintagmático, a oração é formada por sintagmas que se organizam de acordo com regras determinantes. Já o modelo transformacional explica as sentenças construídas por constituintes separados por morfema, aplicando regras que transformam uma estrutura profunda em uma superficial, como foi visto nos parágrafos anteriores.

Posta uma breve reflexão sobre os pressupostos teóricos, os quais constituem a essência da gramática gerativa, cabe conceituar os elementos que formam os textos verbais, esmiuçando cada constituinte.

Texto vem do vocábulo latino *textum*, que significa tecer, entrelaçamento de tecidos. Analogicamente, o texto é um entrelaçamento de sentenças/frases. A frase é tida como expressão do pensamento emitido verbalmente e que tem sentido em si, ou seja, consegue estabelecer a comunicação. A frase é usada como sinônimo de sentença que, segundo Santos (2005), é a sequência gramatical de uma língua.

A frase é uma estrutura formada por alguns constituintes denominados de sintagmas. Esses, por sua vez, são tidos, para Silva e Koch (2009), como conjunto de elementos que se organizam em torno de um núcleo e assim

constituem uma unidade significativa dentro da oração, mantendo relações de dependência e ordem.

Nas palavras de Perini (2003, p. 44), os constituintes são “certos grupos de unidades que fazem parte de sequências maiores, mas que mostram certo grau de coesão entre eles”. Para esse mesmo autor, é necessário ter um bom entendimento sobre estruturação das sentenças em constituintes, pois toda análise se baseia nas estruturas das frases.

Os sintagmas são definidos por Dubois-Charlier (1977) como sequências de palavras que constituem uma unidade, dessa maneira, pode-se dizer que são associações de elementos compostos em conjuntos, organizados e funcionando conjuntamente. Sintagma significa organização e relações de dependência e de ordem em torno de um elemento essencial, o núcleo.

Os sintagmas são classificados como essenciais e facultativos. Os essenciais são tidos como elementos básicos de uma oração[10], o sintagma verbal (SV), cujo núcleo é um verbo ou uma locução verbal, e o SN (sintagma nominal) que tem como núcleo o substantivo ou palavra substantivada. Esses dois tipos de sintagmas são compostos por outros sintagmas como o sintagma adjetival (SA), o qual tem como núcleo um adjetivo ou locução adjetiva, o sintagma

[10] Oração é um enunciado que gira em torno de um verbo, enquanto frase é qualquer enunciado com sentido completo, dessa maneira, toda oração é uma frase, mas nem toda frase é uma oração. Assim, só haverá sintagmas verbais em oração.

preposicional (SP) que é composto por um núcleo chamado preposição e o sintagma adverbial (SAdv) cujo núcleo é um advérbio ou uma locução adverbial.

O SN é definido como a menor unidade de sentido do discurso e que possui uma estrutura sintática e lógico-semântica (KURAMOTO, 1999). Perini et al (1996) conceituam SN como uma classe gramatical que se comporta sintaticamente como sujeito, objeto direto e se for o caso de ser precedido de uma preposição, pode se comportar como um adjunto adnominal ou um objeto indireto. Já para Liberato (1997), o SN é a parte do enunciado que representa um conceito ou referente.

No contexto da Ciência da Computação, Miorelli (2001) vê o SN como uma informação que é composta de um conjunto de símbolos e que possui uma estrutura, assim, ele é muito significativo na identificação do tema central de um documento.

Para a CI, especificamente na RI, o SN é tido como descritor no processo de classificação e recuperação da informação, pois, como afirma Kuramoto (1995), é o conjunto que traz o tema do enunciado, por isso é considerado promissor por esse autor.

Quando o SN é extraído do texto, ele mantém o mesmo significado. Daí a importância de tê-lo na organização da informação, pois são elementos identificadores de informação com alto poder discriminatório.

Os SNs podem ser organizados hierarquicamente em forma arbórea, o que, para Kuramoto (2002), permite que o

usuário filtre com mais fidelidade os resultados de sua consulta, assim, os níveis da estrutura dos SNs vão do mais generalizado ao mais específico, resultando em uma maior interação do usuário com o sistema e em um maior retorno de documentos relevantes. O SN denominado de nível 1 é considerado simples, o de nível 2 é aquele a partir do qual foi extraído o de nível 1 e, assim por diante.

Ex. 1: A extração de sintagma nominais em teses e dissertações

- **Nível 3**: A extração de sintagma nominais em teses e dissertações
- **Nível 2**: A extração de sintagmas nominais
- **Nível 1**: Sintagmas nominais

O Ex. 1 é de um SN, em que são expostos seus níveis hierárquicos. Desse modo, o usuário navegaria por esses níveis até satisfazer a sua busca.

Para se construir a estrutura arbórea de uma sentença, Miorelli (2001) aponta que se devem conhecer as estruturas de uma língua.

> Para construir a estrutura em árvore de uma sentença, devemos conhecer quais as estruturas legais da língua. Um conjunto de regras de reescrita descreve quais as estruturas permitidas. Essas regras dizem que um certo símbolo pode ser expandido em árvore, pela seqüência de outros símbolos. Cada símbolo é um constituinte da sentença que pode ser composto de uma ou mais palavras. (MIORELLI, 2001, p. 18)

Os SNs possuem duas estruturas, assim denominadas: uma à esquerda do núcleo do sintagma, o qual pode ser constituído por determinantes (possessivos, quantificadores e outras classes de palavras); a segunda é uma estrutura à direita do núcleo, a qual é composta de modificadores que, por sua vez, podem ser classes abertas ou outros sintagmas. (PERINI, 2003). Contudo, opta-se por ver a estrutura do SN composta por três partes: determinantes (artigos, pronomes, numerais, e outros), núcleo (substantivo) e modificadores (adjetivo, advérbio)

Entendem-se como palavras abertas aquelas que possuem mecanismos de produção de novos vocábulos, utilizando a derivação e a composição. Estão inclusas nessa classe substantivos, adjetivos e verbos. Já a classe fechada se refere a palavras que já são consagradas dentro de uma língua, como aponta Santos (2005), e onde as mudanças são quase inexistentes como é o caso dos artigos, pronomes, numeral, advérbio, preposição, conjunção e interjeição.

Há várias possibilidades na composição de um SN, ou seja, ele não é estático. Sua composição pode ser feita ora por único substantivo/pronome (núcleo), ora pela composição entre determinantes (DET), núcleo (N) e modificadores (MOD), como podem ser vistos nas regras de formação apresentadas na Tabela 4, vale ressaltar que a numeração das regras não representa necessariamente o grau de importância de uma em relação à outra, é apenas um recurso didático de apresentação das mesmas:

Tabela 4 – Regras de formação de SNs

Regras	Exemplos
Regra geral: DET + MOD + N + MOD	A interdisciplinar Ciência da Informação
Regra 1: DET + N + MOD	A Ciência da Informação
Regra 2: N + MOD	Informação estratégica
Regra 4: DET + N	A informação
Regra 5: **N**	**Informação**
Regra 6: DET + N + DET + N + MOD	A filosofia e a ciência juntas
Regra 7: DET + DET + N + MOD	A minha recuperação da informação
Regra 8: MOD + N + MOD	Grande área da informação
Regra 9: DET + DET + **N**	Uma certa área

Fonte: baseada em Miorelli (2001) e Santos (2005)

Os determinantes são formados por artigos, pronomes demonstrativos e pronomes possessivos, sendo assim, uma classe de palavra fechada. A sua função é especificar um nome, sendo anteposto a esse, o que permite uma construção de sua valoração de referência, fazendo com que a extração de informações acerca das propriedades semântico-sintáticas dos objetos ou entidades as quais sejam referentes se realize. Se for levada em consideração que o DET pode ser subdividido em pré-det, det-base ou pós-det, as regras 7 e 9 se modificariam, ficariam DET + N + MOD e DET + N, semelhando-se as regras 1 e 4 respectivamente.

O núcleo de uma SN sempre será um vocábulo com a função nominal. Assim, as classes de palavras que se enquadra nessa função são os substantivos, pronomes, numerais e palavras substantivas que, por sua vez, são aquelas pertencentes a outras classes, mas que se tornam substantivos quando precedidas por um artigo ou por pronome (demonstrativo ou possessivo). Desse modo, o SN

pode ter como núcleo um substantivo próprio ou comum, pronome substantivo, pronomes (pessoal, demonstrativo, indefinido, interrogativo, possessivo, relativo).

Os modificadores, diferentemente dos determinantes que sempre antecedem o núcleo, podem estar antepostos ou pospostos ao núcleo, exercendo a função de caracterizar, quantificar, enfatizar e sendo compostos por adjetivos, locuções adjetivas, advérbios e locuções adverbiais.

Por fim, como conceitua Morellato (2007), o SN é a unidade sintático-semântica cujo núcleo é constituído por um nome ou pronome. Para Miorelli (2001), os SNs são percebidos como forma sintática (em que se privilegia a forma) ou semântica (em que se buscam os significados com suas especificidades e implicações).

2.3.2 Extração Automática de SNs

Corrêa et al (2011) afirmam que os SNs são extraídos do texto e analisados a fim de facilitar o processo de indexação automática. Dessa maneira, a utilização dos sintagmas nominais como recurso de acesso à informação contida em uma base de dados textual se apresenta como uma forma alternativa aos SRIs tradicionais, como também apontou Kuramoto (1995, p. 03):

> O processo de indexação produzindo uma lista de descritores visa à representação dos conteúdos dos documentos. Ou seja, este processo tem como objetivo extrair as informações contidas nos documentos, organizando-as para permitir a recuperação destes

> últimos. Assim, os descritores deveriam ser, obrigatoriamente, portadores de informação de maneira a relacionar um objeto da realidade extra-linguística com o documento que traz informações sobre este objeto. Contudo, na maioria dos SRI convencionais, os descritores não passam de uma simples lista de palavras extraídas dos documentos que constituem as bases de dados.

Como já foi exposto, os SNs podem ser definidos como grupos nominais constituídos de uma organização hierárquica em árvore e, diferentemente das palavras (tidas como isoladas), quando extraídos do texto mantém o significado.

Souza (2005) coloca que a identificação de sintagmas pode ou não ser fácil, pois, segundo Perini (1995), depende da intuição para que a oração seja separada em seus constituintes imediatos, feitos a partir de critérios puramente formais. Já para Liberato (1997), essa tarefa é realizada completamente através de uma abordagem cognitiva e contextual que é permitida pela análise do discurso e pela pragmática. Enquanto para outros autores, é feita a partir de uma análise transformacional, como se vê em Ruwet (1975 apud SOUZA, 2005).

A semântica também deve ser levada em conta, pois ela se depara com as questões de anáforas, as quais acontecem quando há substituições de estruturas mais complexas por uma mais simples ou vice-versa dentro de um texto. Para que haja compreensão, as anáforas são interdependentes.

Morellato (2010) afirma que a identificação de SNs implica em soluções para diversas áreas como RI, resolução

de anáforas, construção de ontologia, entre outros. Para a realização da tarefa de extração de SNs, várias propostas foram oferecidas por diversos autores, dos quais podem ser citados Kuramoto (1995), Miorelli (2001), Souza (2005), Maia (2008), Morellato (2010) entre outros. Esses autores construíram alguns métodos para extração de SNs, aplicando-os em sistemas automatizados.

2.3.3 Identificador ou Extrator de Sintagmas Nominais (SNs)

Quando o PLN está preocupado com as tarefas relacionadas à recuperação e extração de informações, análise sintática, resolução de correferência e análises semânticas, ele deve observar a identificação de SNs que pode ser um apontamento para resoluções dessas problemáticas, como afirma Santos (2005).

Quando se fala de identificação automática de SNs, significa observar programas computacionais que identifique as sequências de léxicos que constituem SNs. Sua aplicação pode ser feita na RI para criar termos de indexação.

Para Santos (2005), a identificação de SNs é um problema de classificação, pois associa a cada item do corpus uma etiqueta adicional que o classifique como pertencente ou não a um SN.

Pode-se entender um identificador de SNs como um programa computacional que tem a função de retornar sintagmas nominais contidos em um texto, tendo como

entrada a frase que passa por um pré-processamento em que os léxicos são etiquetados (recebendo categorias gramaticais) e submetidos às regras gramaticais de uma língua natural e tendo como saída o texto com os SNs marcados.

Um exemplo de identificador de SNs é o Sistema Identificador de Sintagmas Nominais do Português (SISNOP) de Morellato (2010). Essa ferramenta é compreendida como um extrator de SN com um conjunto de módulos ou de programas que interpretam textos através de análises morfológicas e sintáticas, permitindo a identificação e a extração dos SNs de cada frase contida nos textos. Outro exemplo de extrator é o OGMA. Já os *parsers* PALAVRAS, o LX-Parser, o Curupira e o Grammar Play são também ferramentas que podem ser utilizados para identificar os SNs, contudo para fins posteriores de extração. Assim como O TBL - *Transformation-Based Learning* (Aprendizado Baseado em Transformações) de Santos (2005) que também foi desenvolvido para identificar SNs.

Ferramentas de extração automática de SNs são *softwares* desenvolvidos para fazer o processo de etiquetagem que consiste em marcar as palavras de acordo com suas categorias gramaticais, dando-lhes traços linguísticos e a partir daí, com a combinação dessas classes gramaticais, têm-se regras que regem as estruturas formadas pelos constituintes dos SNs. Dessa maneira, as ferramentas extraem os SNs, aplicando no texto as regras de estrutura dos SNs e detectando, assim, se um determinado conjunto de constituintes está de acordo com a regra do sintagma ou não.

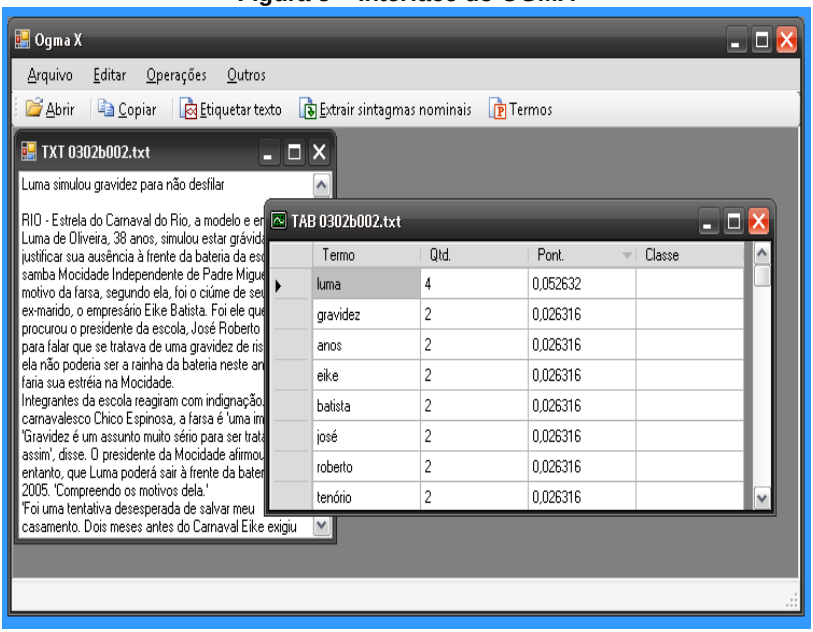
O programa mostra como saída uma lista de sintagmas nominais retirados do texto.

Os *parsers* são *softwares* que realizam a análise sintática das frases, construindo uma representação arbórea dos constituintes das mesmas. Os *parsers* podem ser utilizados como identificadores de SN, pois recebem os textos e dão como retorno textos em formatos de árvores, ou em gráfico, ou em representação de análise de texto simples, em que os SNs são identificados.

Nessa seção serão descritas superficialmente algumas ferramentas de extração automática de SNs como o OGMA e *parsers* como o PALAVRAS, o Grammar Play e o LX-Parser.

O OGMA faz análise textual, cálculo de similaridade entre documentos e extração de SN, identificação da classe do SN e o cálculo da pontuação do mesmo como descritor de forma automática. A Figura 5 traz a representação da interface dessa ferramenta.

Figura 5 – Interface do OGMA



Fonte: OGMA. Disponível em: <http://www.luizmaia.com.br/ogma/>. Acesso em: 16 maio 2013.

Cita-se o trabalho de Corrêa et al (2011) que utiliza o OGMA para analisar a aplicabilidade da extração de SNs em teses e dissertações no contexto da BDTD-UFPE. A ferramenta é fácil de ser manuseada e tem a vantagem de ser disponibilizada para *download* livremente e a desvantagem é que não gera representação arbórea.

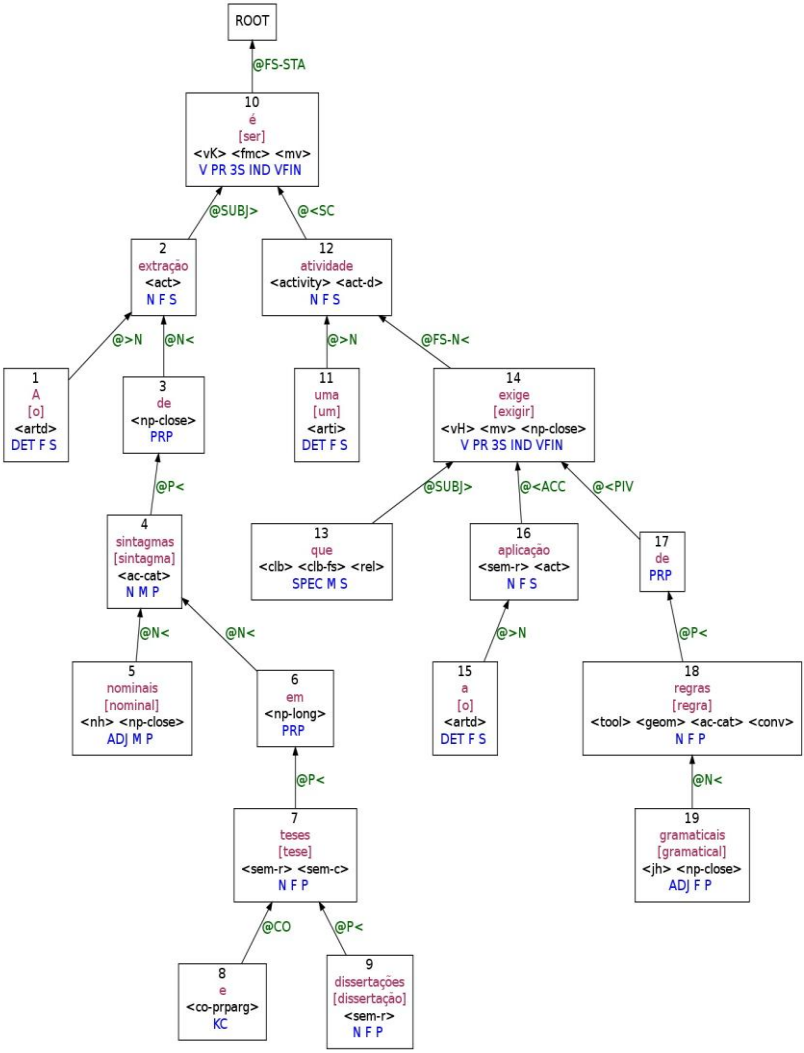
O *parser* PALAVRAS (BICK, 2000), segundo Maia (2008), é um programa que faz parte de um conjunto de várias ferramentas multilíngues que compõe o V*isual Interactive Syntax Learning* (VISL). Elas estão disponibilizadas na internet, às quais o usuário submete um texto ou até mesmo uma sentença e recebe como resultado o texto com a

marcação. As marcações indicam elementos linguísticos diversos no texto como categorias gramaticais, flexões de gênero e número, sujeito (SN) e predicado (SV).

O *parser* PALAVRAS apresenta como saída três formatos:

1- forma gráfica, conforme a Figura 6, em que há a representação para a frase “A extração de sintagmas nominais em teses e dissertações é uma atividade que exige a aplicação de regras gramaticais”;

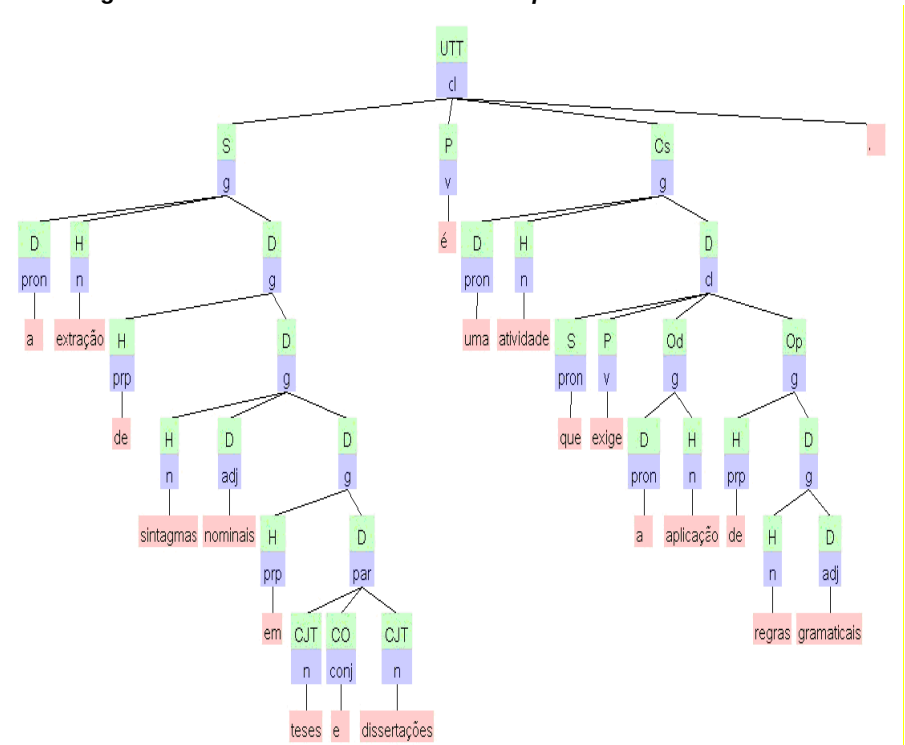
Figura 6 – Forma gráfica da análise do *parser* PALAVRAS



Fonte: VISL. Disponível em: <http://beta.visl.sdu.dk/visl/pt/parsing /automatic/dependency. php>. Acesso em: 30 maio 2013.

2- forma arbórea, conforme a Figura 7, em que há a representação para a mesma frase;

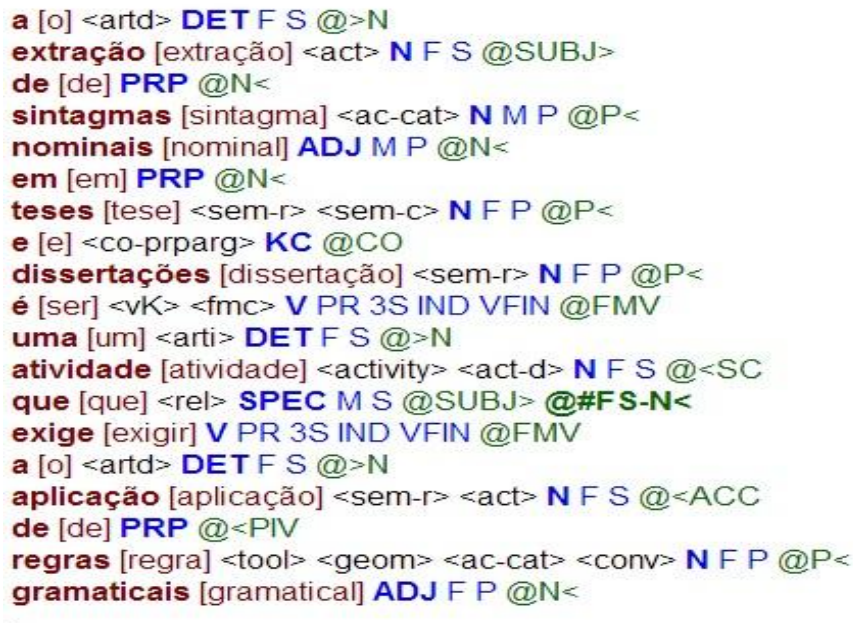
Figura 7 – Forma arbórea da análise do *parser* PALAVRAS



Fonte: VISL. Disponível em: <http://beta.visl.sdu.dk/visl/pt/parsing/automatic/trees.php>. Acesso em: 30 maio 2013.

3- formato de representação da análise de texto, como é vista a Figura 8.

Figura 8 – Análise textual feita pelo do *parser* PALAVRAS



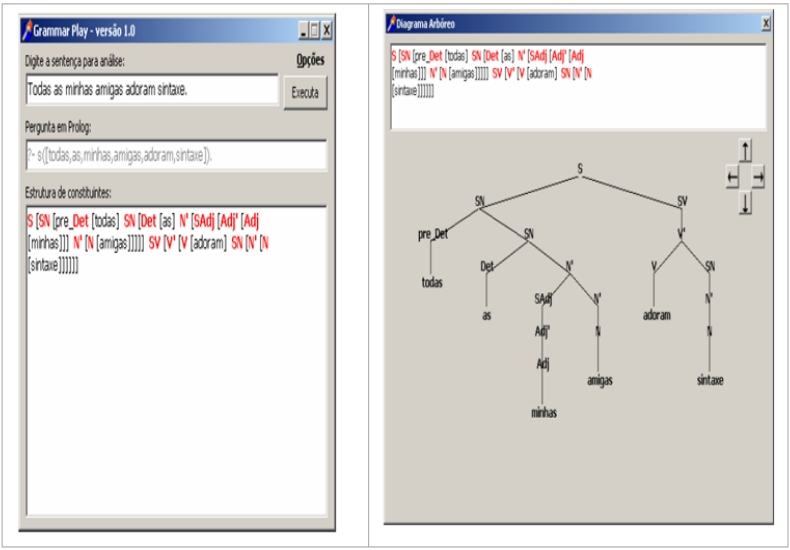
Fonte: VISL. Disponível em: < http://beta.visl.sdu.dk/visl/pt/parsing/automatic/parse.php>. Acesso em: 30 maio 2013.

O VISL trabalha com uma gramática de restrições que é feita a partir de uma análise em que se observam os morfemas, os grupos de palavras e a composição da oração, o que permite uma observação da ortografia, da semântica e da sintaxe. Quando feita a identificação dos morfemas, essa ferramenta elimina as ambiguidades encontradas em cada léxico. A eliminação se dá por meio de uma aplicação de regras que identificam e eliminam as possibilidades de

estruturas sintáticas inexistentes. O *parser* PALAVRAS é usado por Souza (2005), Santos (2005), Miorelli (2001), Morellato (2007; 2010), Lopes (2011), entre outros.

Outra ferramenta é o Grammar Play (OTHERO, 2004) que faz análise apenas de sentenças e usa o mesmo princípio de restrição utilizado pelo VISL. A técnica aplicada para fazer as análises vai da identificação de um léxico que permita perceber as várias possibilidades de cada palavra à identificação dos sintagmas, aplicando as regras gramaticais. Desse modo, o *parser* consegue identificar os sintagmas (nominais, adjetivais, preposicionais, verbais e adverbiais). As interfaces gráficas do Grammar Play estão representadas na Figura 9.

Figura 9 – Interfaces gráficas do Grammar Play



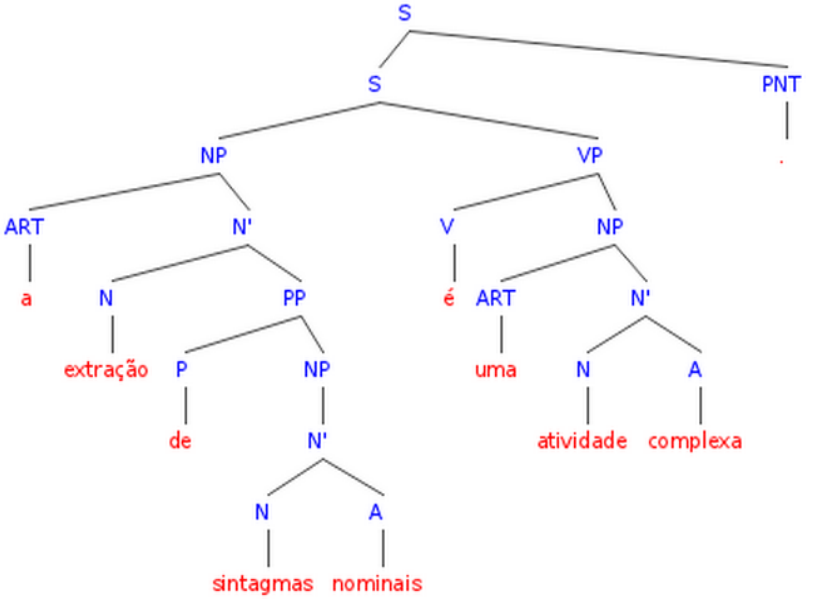
Fonte: Othero (2009)

Na imagem à esquerda, há representação gráfica dos colchetes rotulados, já na imagem à direita, há representação da estrutura arbórea. Vale ressaltar que esse *software* só pode ser usado em um único tipo de frase, a declarativa.

Outro analisador é o LX-Parser[11] que é uma parte de um conjunto chamado *lx suíte.* Essa ferramenta é constituída de vários programas que fazem algumas tarefas: o *lx-lemmatizer* que identifica a conjugação dos verbos fornecidos; o *lx-conjugator* que faz a conjugação dos verbos; o *lx-tagger* que marca as categorias morfossintáticas do texto; e o *lx-infletor* que identifica as variantes de gênero e número dos termos. Assim, o LX-Parser é um analisador sintático de textos escritos em português baseado em uma abordagem estatística. Na relação de ferramentas baseadas em estatísticas como MXPost, Build Tiger, Gate, entre outras. A Figura 10 traz a representação da estrutura em árvore resultante da análise do LX-Parser para a sentença “A extração de sintagmas nominais em teses e dissertações é uma atividade que exige a aplicação de regras gramaticais".

[11] LX-Parser. Disponível em: <http://lxcenter.di.fc.ul.pt/tools/pt/conteudo/LXParser.html>. Acesso em 17 jun. 2013.

Figura 10 – Representação arbórea feita pelo LX- Parser



Fonte: LX-Parser. Disponível em: <http://lxcenter.di.fc.ul.pt/tools/pt/conteudo/LXParser.html>. Acesso em 17 jun. 2013.

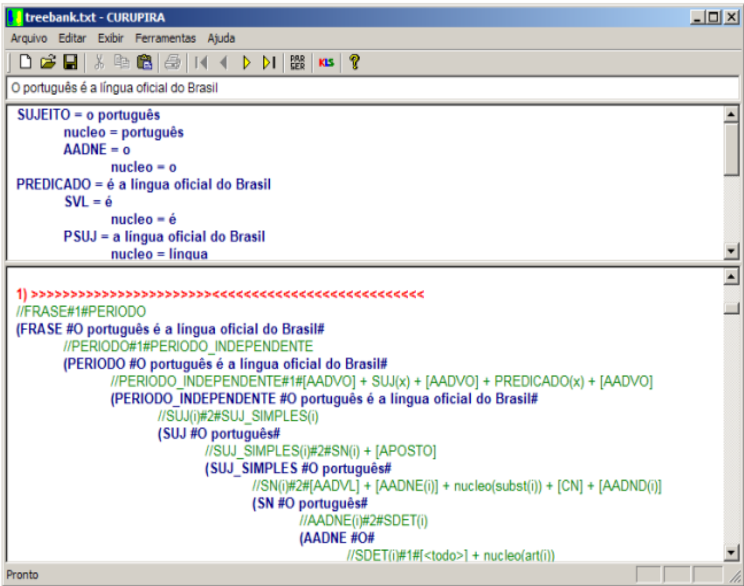
O LX-Parser é desenvolvido pelo NLX-Group. O *site* do LX-Center, no qual o *parser* está agregado, disponibiliza várias outras ferramentas de trabalho de análise linguística feita automaticamente.

O *parser* Curupira[12] (MARTINS; HASEGAWA; NUNES, 2003) é um *parser* que analisa as frases em português brasileiro da esquerda para a direita, de cima para baixo por meio de uma gramática de restrição e livre de contexto. O

[12] NÚCLEO INTERINSTITUCIONAL DE LINGUÍSTICA COMPUTACIONAL - NILC. Curupira. Disponível em: <http://www.nilc.icmc.usp.br/nilc/tools/curupira.html>. Acesso em: 13 abr. 2013.

programa tem um conjunto de 1,5 milhões de léxicos em formas livres, incluindo palavras flexionadas e derivadas, o que permite compreender informações morfossintáticas como pessoa do discurso, número, gênero, tempo, transitividade, entre outras. Na Figura 11 é apresentada a interface do Curupira.

Figura 11 – Interface gráfica do Curupira



Fonte: MARTINS; HASEGAWA; NUNES, 2003

Das ferramentas apresentadas, três são desenvolvidas por estudiosos brasileiros, o OGMA, o Curupira e o Grammar Play. Maia (2008) desenvolveu o OGMA no âmbito do Programa de Pós-Graduação em Ciência da Informação da Universidade Federal de Minas Gerais. Othero (2004) construiu o Grammar Play no Programa de Pós-Graduação

em Linguística Aplicada da Pontifícia Universidade Católica do Rio Grande do Sul. Já Martins, Hasegawa e Nunes (2003) desenvolveram o Curupira no NILC da Universidade Federal de São Carlos.

2.3.4 Seleção Automática de SNs

Nem todo SN é um descritor de um documento. Na busca por um melhor termo que descreva o conteúdo de um dado documento, deve-se levar em consideração a relevância do termo extraído. Um termo é relevante quando contém informações buscadas pelo usuário, respondendo a questão formulada por estes e quando descreve os assuntos tratados nos documentos.

Na terminologia, “termo” é tido como elemento constitutivo da produção do saber, cujas propriedades permitem a univocidade da comunicação especializada (KRIEGER; FINATTO, 2004). Busca-se o conceito de “termo” na terminologia no intuito de se nortear a concepção do que é descritor documental que, desse modo, pode ser concebido como o termo (técnico) que descreve o conteúdo de um documento.

Dentro do estado da arte que se está trabalhando nessa dissertação, percebe-se que essa descrição deve ser feita pela extração automática de termos que, segundo uma aplicação de métodos e técnicas, podem ser refinados, tornando-se assim perfeitos descritores informacionais.

Em SRI, o usuário precisa recuperar informações condizentes com sua consulta, sua necessidade informacional deve ser suprida. Portanto, para efetivar um melhor resultado de busca, os documentos devem estar bem indexados, no intuito de que durante a sua consulta, o usuário seja atendido de forma que o satisfaça. Baeza-Yates e Ribeiro Neto (1999) classificam a consulta em alguns tipos como: por palavras-chave; padrões de comparação (palavras, prefixo, sufixo, conjunto de caracteres, expressões regulares, entre outros); ou por consulta estruturada ou por linguagem natural que, por sua vez, é constituída de análises léxicas e com a eliminação de *stopwords* (artigos, preposições, conjunções e outras palavras que podem ser eliminadas do texto) e *stemming* (resultados após a remoção dos sufixos e prefixos).

Independentemente do tipo de consulta, o documento deve ser descrito por termos que tragam informações relevantes. Assim, como afirmam Kuramoto (1995; 1999), Souza (2005), Maia (2008), Lopes (2011), entre outros, os SNs são termos candidatos a descritor documental porque são portadores de informação conceitual, ou seja, a menor unidade de sentido do discurso (KURAMOTO, 1995).

Dos trabalhos apresentados, poucos trouxeram algum método para selecionar, dentre os SNs extraídos, os melhores descritores. A preocupação primordial da maioria era a extração automática dos SNs, o que é interessante, pois se criaram várias maneiras de extrair SNs automaticamente, permitindo que novas ideias surjam para aperfeiçoar as técnicas.

Contudo, alguns autores incluíram nos seus objetivos critérios de seleção de SNs como descritores de documentos como Kuramoto (1995; 1999), Souza (2005) seguido de Maia (2008) e Pinheiro (2009), entre outros. A maneira como os autores mensuravam os resultados da extração de SNs sempre remetia a dois princípios primordiais: precisão e revocação.

Kuramoto (1995; 1999) não trabalhou diretamente com a seleção dos melhores SNs para indexação e sim com a automação da extração desses. Contudo, o autor faz apontamento para o que poderia ser um melhor descritor documental, tendo em vista o tipo de usuário. Para o usuário novato ou que não possui definida sua necessidade informacional, deve-se ter os centros do SNs de primeiro nível, já para os usuários considerados especialistas, cientes de sua necessidade informacional, os SNs de níveis mais elevados são recomendados.

Souza (2005) propõe uma maneira de escolher, dentre os SNs extraídos, os melhores descritores. A mesma metodologia de seleção desse autor foi usada por Maia (2008). A metodologia consiste na verificação da relevância semântica por meio da quantificação da freqüência de ocorrência dos SNs no texto, a incidência desses no conjunto de documentos, percepção dos níveis dos SNs, análise das suas estruturas sintáticas e sua ocorrência em algum vocabulário controlado ou tesauro de um domínio.

Para tanto, Souza (2005) usa conceitos de “Pontuação” e “Taxa de Relevância” para atribuir valores aos SNs depois de ordená-los de acordo com sua frequência no

texto e destacar os com ocorrência inferior a um patamar de 1% sobre o total de SN únicos do documento. Os valores quanto à relevância estão expostos na Tabela 5.

Tabela 5 – Valor atribuído ao SN de acordo com sua relevância

Relevância descritiva do SN	Símbolo	Valor associado
SN extremamente relevante como descritor	SN***	1,0
SN razoavelmente relevante como descritor	SN**	0,5
SN moderadamente relevante como descritor	SN*	0,25
SN não relevante como descritor	SN –	0,0

Fonte: Souza (2005, p 93)

Para se chegar a esses valores atribuídos a cada tipo de SN, aplica-se a fórmula que resulta na seleção dos SNs quanto a sua qualidade como descritor, segundo Souza (2005):

$$Pontuação\ (desc) = (Núm.SN^{***}) + 0{,}5 \times (Núm.SN^{**}) + 0{,}25 \times (Núm.SN^{*})$$

Atribui-se um valor numérico aos SNs segundo a sua relevância como descritores. O valor numérico máximo é 1, sendo atingido quando a totalidade de descritores fosse extremamente relevante.

Na classificação manual dos SNs, deve-se se observar a estrutura de cada um, como sugere a Tabela 6 em que a categoria do SN recebe um valor segundo sua estrutura sintática e seu nível.

Tabela 6 – Valor atribuído ao SN de acordo com sua estrutura sintática

CSN	Estrutura e Nível do SN	Valor associado
1a	Nível 1, estrutura (D + N)	0,25
1b	Nível 1, qualquer estrutura exceto (D + N)	0,75
2	Nível 2, qualquer estrutura	1,0
3	Nível 3, qualquer estrutura	0,75
4	Nível 4, qualquer estrutura	0,5
5	Nível 5 ou superior, qualquer estrutura	0,25

Fonte: Souza (2005, p. 123)

Assim, Dado um SN que tem num documento uma frequência (F) e é classificado em uma categoria (CSN), sua pontuação é: P= F ***x*** valor (CSN). Dessa, forma, têm-se os valores de acordo com sua estrutura sintática, ou seja, é calculada a pontuação atingida por todos os SNs, fazendo a multiplicação da frequência deles no documento e no corpus de acordo com o peso atribuído à categoria desse tipo de SN. Souza (2005), a partir do corpus da sua tese, sugere os pesos das categorias (valor atribuído) de SN: 0,25 para nível 1a; 0,75 para nível 1b; 1,0 para nível 2; 0,75 para nível 3; 0,5 para nível 4; e 0,25 para nível 5.

Então, é feita uma comparação entre a extração manual com os SNs de maior pontuação, permitindo que se tenha a taxa de relevância na saída. Maia (2008) utiliza toda essa metodologia na confecção do OGMA. Com a atribuição de valor numérico aos SNs, Souza (2005) argumenta que quanto maior a taxa de relevância, melhor é a representação do assunto por descritores.

Para uma melhor seleção de descritores, Corrêa et al (2011) apresentam algumas sugestões como a extração de

sintagmas deveria ser acompanhada de estratégias de ordenação por relevância dos sintagmas, sendo levado em conta critérios de frequência e posicionamento, semelhantemente às propostas existentes para palavras isoladas (p. 20).

Já Lopes (2011) utiliza regras heurísticas de descarte para selecionar os SNs, ordenando-os, logo em seguida, para extrair conceitos de acordo com a relevância de domínio.

A seleção de SNs para fazê-los descritores é fundamental no processo de indexação, pois se as questões quanto à relevância não forem levadas em consideração, os problemas de indexação continuarão os mesmos, recuperando documentos irrelevantes para o usuário.

2.3.5 Avaliação da Extração Automática de SNs

A avaliação da extração automática de SNs é verificar se essas unidades de sentidos são de fato SN. Para tanto, alguns autores lançam mão de fórmulas tradicionais, como as de precisão e revocação. É necessário fazer essa avaliação para saber a real performance do método ou ferramenta de extração. Em alguns trabalhos, havia a extração manual para que fosse feita a comparação com os resultados das extrações automáticas.

Kuramoto (1995-2002) usa um índice de equivalência entre a extração automática e manual, assim como Morellato (2007; 2010) e Miorelli (2001).

Souza (2005) compara os SNs simples extraídos automaticamente com as palavras-chave atribuídas pelas autorias dos artigos. Depois da seleção dos descritores, o autor compara estes com as palavras-chave e resumos dos documentos originais.

Já Santos (2005) faz a avaliação por meio de equações de acurácia na classificação dos SNs, precisão e abrangência na identificação de SNs. Para finalizar a avaliação, o autor compara os SNs identificados pelo *parser* PALAVRAS com os classificados pelo TBL.

Corrêa et al (2011), depois de submeterem os resumos ao OGMA, fazem a verificação manual dos resultados, verificando se os SNs identificados pela ferramenta são de fato SNs, a quantidade de SNs relevantes, os que contém as palavras-chave e os que coincidem com as palavras-chave dos textos analisados. As métricas utilizadas para avaliação foram o cálculo e a análise dos percentuais de precisão em extrair SNs relevantes como descritores, a taxa de erro ao extrair cadeias de caracteres que não configuram SNs e o percentual de SNs extraídos não relevantes como descritores, precisão e abrangência.

As avaliações feitas das metodologias de extração automática dos SNs dos outros trabalhos se restringiram a submeter os resultados das pesquisas a outras ferramentas (ou até mesmo à avaliação manual) no intuito de se fazer a comparação entre os resultados. Vale ressaltar que esses procedimentos iam de acordo com o contexto de cada autor (objetivos, metodologias, referenciais teóricos, ambiente da pesquisa).

3 PROCEDIMENTOS METODOLÓGICOS

Para classificar essa pesquisa em seus diversos aspectos, tomaram-se como preceitos os fundamentos da metodologia científica abarcadas por Gil (2002), Marconi e Lakatos (2009), pois trabalham no intuito de explicitar os procedimentos sistemáticos e racionais, condensando a metodologia científica, técnicas de pesquisa e metodologia do trabalho científico.

Dessa maneira, essa pesquisa se classifica, quanto à sua natureza, como básica, pois permite gerar novos conhecimentos para o avanço da indexação automática a partir da observação da literatura sobre extração automática de SNs de documentos textuais e avaliação de sistemas de indexação automática, sem haver uma previsão da aplicação prática.

Quanto ao seu objetivo, esse estudo é exploratório, porque, como afirma Gil (2002), ele possibilita uma maior familiaridade com o problema, no intuito de torná-lo explícito ou até mesmo desenvolver hipóteses. Dessa maneira, a indexação automática por meio SNs é explorada a partir da extração de informações advindas da literatura com o objetivo de perceber as dificuldades encontradas pelas ferramentas e metodologias de extração automática de SNs e do experimento realizado em que se obtiveram alguns dados empíricos, permitindo, dessa maneira, que possíveis soluções sejam apontadas.

Em relação à forma de abordagem do problema, essa pesquisa se caracteriza pelos métodos de análise quantitativa e qualitativa. Ela é quantitativa, quando trabalha com a quantificação de SNs extraídos dos textos pelas ferramentas de extração automática e a quantidade desses SNs que podem ser usados como descritores documentais, apresentando os resultados por meio de técnica estatística e fazendo uso, também, de fórmulas na obtenção de taxa de acerto, taxa de erro, taxa de revocação e taxa de relevância dos SNs. O enfoque passa a ser qualitativo quando se faz a interpretação desses resultados por meio de análise reflexiva que estabeleça a relação das variáveis, observando o contexto em que cada trabalho da literatura explorada se desenvolveu, permitindo que se apontem soluções hipotéticas para os problemas apresentados pelas ferramentas e métodos de extração automática de SNs. Vale ressaltar que o levantamento do estado da arte também é qualitativo

Nos procedimentos, foi utilizada a pesquisa bibliográfica para atender quatro objetivos específicos: fazer o levantamento do estado arte das pesquisas sobre a indexação automática por meio dos SNs; identificar e sintetizar os fundamentos teóricos da referida temática; discutir as metodologias para a extração automática de SNs no intuito de se buscar critérios de seleção de SNs descritores; e pesquisar o processo de identificação, extração e seleção automática de SNs em textos escritos em língua portuguesa.

A pesquisa bibliográfica permite identificar o conhecimento acerca da indexação automática por meio dos SNs existente na literatura. Nessa fase, foram levados em

consideração alguns contextos acerca do tema: o histórico para a compreensão da evolução dos conceitos relacionados à indexação; o empírico para perceber as metodologias que são utilizadas para o estudo e aplicação da indexação automática por meio dos SNs; o determinístico do estado da arte para visualizar o quadro literário desenvolvido até o momento sobre a indexação automática a partir da extração e seleção automática de SNs em textos escritos em língua portuguesa.

Para o levantamento bibliográfico, foram usadas algumas ferramentas de busca da *Web*. Para buscar e recuperar teses e dissertações escritas nas instituições de pesquisa do Brasil, foi usada a Biblioteca Digital Brasileira de Teses e Dissertações (BDTD – IBICT) [13], utilizando como termos para estratégias de busca:

- sintagmas nominais;
- “sintagmas nominais”;
- sintagma + nominal;
- recuperação + informação + sintagma + nominal;
- extração + sintagma + nominal;
- sintagma + nominal + indexação;
- indexação automática;
- extração automática de sintagmas nominais.

[13] INSTITUTO BRASILEIRO DE INFORMAÇÃO EM CIÊNCIA EM TECNOLOGIA – IBICT. BIBLIOTECA DIGITAL BRASILEIRA DE TESES E DISSERTAÇÕES – BDTD. Disponível em: <http://bdtd.ibict.br/>. Acesso em: 05 fev.2013.

Esses mesmos termos foram usados para fazer a busca no Google Acadêmico[14], na SciELO[15], na BRAPCI[16], no Portal de Periódicos da Capes[17] e no Repositório Científico de Acesso Aberto de Portugal[18] no intuito de encontrar artigos de periódicos científicos, além de outros materiais científicos (relatórios, anais de eventos, entre outros). Vale ressaltar que na recuperação de artigos de periódicos, foi feita a busca, também, com os referidos termos traduzidos para a língua inglesa se restringindo a trabalhos que tratem da extração de SNs em textos escritos em português:

- *noun phrases;*
- *"noun phrases";*
- *noun + phrase;*
- *Information + retrieval + noun + phrase;*

---

[14] GOOGLE ACADÊMICO. Disponível em: <http://scholar.google.com.br/>. Acesso em: 07 fev. 2013.

[15] SciELO. Disponível em: <http://www.scielo.org>. Acesso em: 20 fev. 2013.
SciELO BRASIL. Disponível em: <http://www.scielo.br>. Acesso em: 20 fev. 2013.

[16] BASE DE DADOS REFERENCIAIS DE ARTIGOS DE PERIÓDICOS EM CIÊNCIA DA INFORMAÇÃO – BRAPCI. Disponível em: < http://www.brapci.ufpr.br>. Acesso em: 15 jan. 2013.

[17]MINISTÉRIO DA EDUCAÇÃO. COORDENAÇÃO DE APERFEIÇOAMENTO DE PESSOAL DE NÍVEL SUPERIOR – CAPES. PORTAL DE PERIÓDICOS DA CAPES. Disponível em: <http://www.periodicos.capes.gov.br>. Acesso em: 15 fev. 2013.

[18] MINISTÉRIO DA EDUCAÇÃO E CIÊNCIA (PORTUGAL). Repositório Científico de Acesso Aberto de Portugal. Disponível em: <http://www.rcaap.pt>. Acesso em: 06 fev. 2013.

- *Extraction + noun+ phrase;*
- *noun + phrase + index;*
- *automatic indexing;*
- *automatic extraction of noun phrases.*

A busca de teses e dissertações escritas sobre o tema em instituições de pesquisa de Portugal foi feita no Repositório Nacional de Dissertações e Teses Digitais (DITED) (Biblioteca Nacional de Portugal) [19].

Dentre os trabalhos recuperados, foram também extraídas informações das Referências nas quais os autores se basearam para recuperar outros trabalhos, assim como a filtragem das pesquisas que trabalham a indexação automática por meio de SNs em textos escritos em língua portuguesa.

A partir das sínteses dos trabalhos selecionados, constrói-se o estado da arte do tema abordado nessa dissertação. Muitos trabalhos foram lidos, mas os trabalhos escolhidos foram os que, de fato, falavam sobre a identificação e/ou extração de SNs em textos escritos em língua portuguesa. Assim, têm-se: as teses de Kuramoto (1999), Bick (2000), Souza (2005), Maia (2008), Lopes (2011) e Pinheiro (2009); as dissertações de Miorelli (2001), Othero (2004), Santos (2005), Arcoverde (2007), Costa (2007), David (2007) e Morellato (2010); o Trabalho de Conclusão de Curso

[19] BIBLIOTECA NACIONAL DE PORTUGAL. REPOSITÓRIO NACIONAL DE DISSERTAÇÕES E TESES DIGITAIS – DITED. Disponível em: <http://dited.bn.pt>. Acesso em: 05 fev 2013.

(TCC) de Morellato (2007); os artigos de Kuramoto (1995; 1997; 2002), Bick (1998), Vieira et al (2000), Souza (2006), Maia e Souza (2010) e Corrêa et al (2011); além do capítulo de livro escrito por Kuramoto (2006).

Para atender ao objetivo de avaliar e comparar softwares extratores de SNs, usando como referência a representação arbórea de sentenças estudadas em Linguística, foi usado nos procedimentos o estudo de caso, pois permite estudar a extração automática de SNs de forma profunda e exaustiva no intuito de descrever os contextos em que as aplicações das metodologias foram realizadas, formular as hipóteses possíveis acerca dos problemas não resolvidos pelas ferramentas, explicando as variáveis que determinam os vários comportamentos tomados por esses softwares.

Para avaliar e comparar as ferramentas de extração automática de SNs, fez-se extração manualmente dos SNs presentes no conjunto de textos usados para compor o corpus dessa pesquisa que é constituído de 30 resumos de teses e dissertações de três áreas diferentes (Direito, Nutrição e Ciência da Computação), indexados na BDTD/UFPE, sendo 10 resumos para cada área. A escolha desses se deu de forma aleatória, tendo como objetivo observar o comportamento das ferramentas automáticas diante de documentos de áreas de domínios diferentes.

Feita a extração de SNs dos textos que compõem o corpus dessa dissertação, compararam-se as extrações automáticas de SNs feitas pelo OGMA, pelo *parser* PALAVRAS e pelo LX – Parser, a fim de apresentar as falhas

existentes nessas ferramentas e possibilitar o melhoramento dos recursos para extração automática desses símbolos linguísticos. Esses *softwares* foram escolhidos porque estão disponíveis livremente na *Web*.

Os resumos constituem de texto com formato simples, sendo tiradas as palavras-chave, o nome do autor e os dados institucionais, contemplando apenas o título e o corpo em si do resumo. Já no LX-Parser, o título do resumo e cada sentença do corpo do resumo (sem o nome da autoria nem os dados institucionais) foram submetidos um a um, já que o programa analisa a sentença até o primeiro ponto final encontrado. Enquanto no PALAVRAS e no OGMA, o título e o corpo do resumo (sem nome da autoria nem os dados institucionais) foram submetidos em única vez.

O único programa que faz a extração de SNs é o OGMA, dessa maneira, após a extração feita por essa ferramenta, os SNs foram copiados e colocados em editores de texto e de planilha. Já o LX-Parser e o PALAVRAS apenas identificavam os SNs, sendo feita a extração manualmente para também editores de texto e planilha.

Logo após a essa etapa, foram computados os SNs totais extraídos por cada ferramenta e a extração manual, classificando cada SN de acordo com sua característica e seguindo as seguintes categorias: expressões que não constituem SNs; SNs compostos por palavras semelhantes às palavras-chave; SNs semelhantes às palavras-chaves.

O procedimento para a seleção de expressões que, de fato, constituem SNs levou em consideração o descarte dos que começavam com preposição ou conjunção, que

continham verbos em suas estruturas e numerais que não exerciam função de determinantes dos nomes. Em seguida foram contabilizados os SNs que eram semelhantes às palavras-chave e os que continham essas em sua estrutura.

As métricas utilizadas para a computação dos resultados foram: a taxa de acerto que é a razão do total de expressões que de fato constituem SN sobre o total de expressões extraídas por cada programa; e a taxa de revocação que, por sua vez, é a razão do número de expressões que de fato são SNs extraídos pelos *softwares* sobre o número total de SNs extraído pela técnica manual.

Os apêndices foram confeccionados a partir de um recorte do corpus dessa pesquisa no intuito de exemplificar a extração de expressões feitas pelos *softwares* OGMA, Parser PALAVRAS, LX-Parser e pela técnica manual.

4 PRINCIPAIS CONTRIBUIÇÕES PARA A EXTRAÇÃO AUTOMÁTICA DE SNs (ESTADO DA ARTE)

Nessa seção, serão descritas algumas metodologias de extração e seleção de SNs, bem como da avaliação das ferramentas na realização destes processos, utilizadas pelos autores em seus trabalhos acadêmicos como artigos, teses e dissertações. Os autores abordados são Kuramoto (1995; 1999), Vieira et al (2000), Bick (2000), Miorelli (2001), Othero (2004), Souza (2005), Santos (2005), Arcoverde (2007), Costa (2007), David (2007), Morellato (2007; 2010), Maia (2008), Lopes (2011) e Pinheiro (2009). O final da seção é feita uma breve análise dos trabalhos, estabelecendo algumas relações entre eles.

4.1 Contribuições de Kuramoto (1995 - 2002)

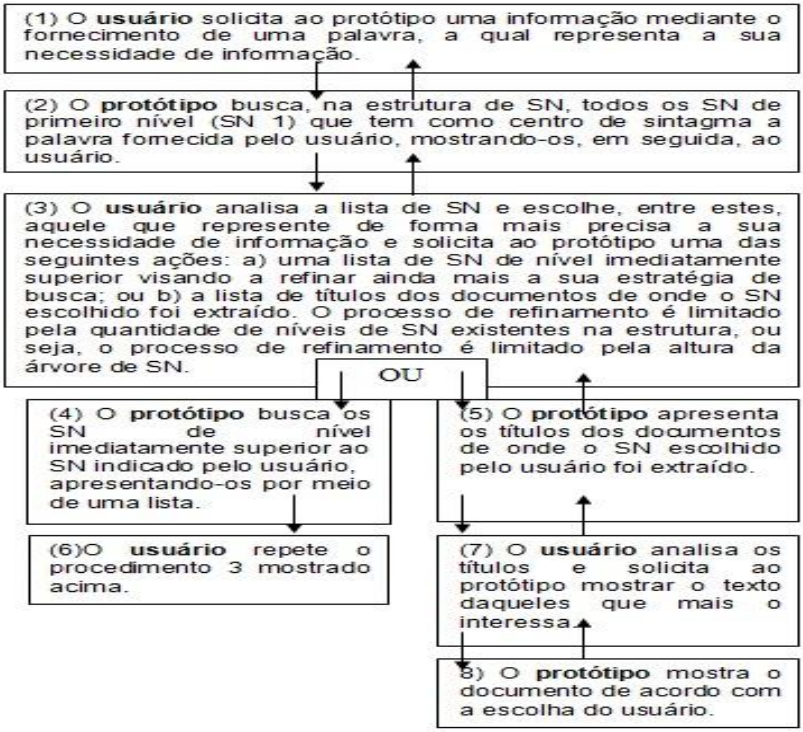
Kuramoto (1995) foi um dos precursores sobre os estudos de identificação e extração de SNs com a utilização de PLN em textos escritos em língua portuguesa e a maioria dos trabalhos analisados nessa dissertação fazem referência a esse autor no que diz respeito à identificação e extração de SNs, assim como a abordagem de se ter os SNs como descritores documentais. Seus trabalhos foram e são desenvolvidos na área de Ciência da Informação.

O autor propõe uma nova interface nos sistemas de buscas textuais e uma nova alternativa no tratamento de RI utilizando SNs. Para fazer os testes, a extração de SNs foi

feita manualmente devido à inexistência de ferramenta que fizesse esse processo automaticamente, tal tarefa foi feita simulando uma extração feita automaticamente. Os SNs foram extraídos de um corpus composto por uma amostragem de 15 artigos escritos em língua portuguesa e extraídos da revista Ciência da Informação. Para selecionar os artigos, Kuramoto (1995) levou em consideração a importância da definição de um domínio já que isso está relacionado estritamente com o critério de obtenção de menor ocorrência de ambiguidades entre os SNs. Para a extração de SNs, o autor considerou algumas questões como resoluções de anáforas e de elipses, assim como problemas relacionados à identificação de SNs sem determinação (quando não são precedidos por nenhum tipo de artigo), o que é bem comum na língua portuguesa.

Sendo assim, Kuramoto (1995) cria um protótipo que tem a interface de busca baseada no encadeamento hierárquico existente entre os SNs, dessa maneira, a interface varre a estrutura arbórea dos SNs de modo interativo com o usuário, como pode ser detalhado na Figura 12.

Figura 12 - Procedimentos de interação usuário ⟺ protótipo



Fonte: adaptada de Kuramoto (1995)

O autor ressalta que a forma escolhida para a interação entre o usuário e a interface de busca do protótipo foi orientada por menus porque possibilita que a implementação e ajustes sejam feitas facilmente, contudo nada impede o desenvolvimento de um SRI que seja orientado a comandos em linguagem natural. A ideia, a priori, era construir uma interface capaz de criar menus dinamicamente à medida que os SNs fossem sendo recuperados e apresentados na tela.

Na extração de SNs, Kuramoto (1995) simula uma extração automática. Como há algumas semelhanças na estrutura sintática das línguas portuguesa e francesa, o autor utiliza algumas regras básicas de reescritura dos SNs adotadas. Alguns problemas dificultam a identificação de SNs como as questões de anáforas e de elipses.

Na identificação dos SNs, Kuramoto (1995) atenta para as seguintes dificuldades na extração de SNs: os SNs escondidos em frases com fatoração (as que são constituídas por sequência de palavras que antecedem outro conjunto de palavras coordenadas pelas conjunções “e / ou” ou isoladas por parênteses); a ausência de determinantes em muitos SNs da língua portuguesa, tal fato que exige aplicação de regras diferentes de outras línguas como a francesa; os elementos anafóricos que aparecem frequentemente na língua portuguesa mediante pronomes (o autor não conseguiu resolver problemas de anáforas relacionados à ausência de fontes explícitas no corpo do texto, como por exemplo, “nesse sentido” (que sentido?), e às fontes que se encontram na história dos acontecimentos, como por exemplo, “esse período pré-industrial”; e atentou também para as elipses que consiste na falta de complemento de algumas palavras.

Nos estudos posteriores, Kuramoto (1999; 2002) considera os SNs como recursos na RI já que são oferecidas duas alternativas para seu uso. A primeira está relacionada à indexação automática em que se reproduziria o método tradicional, mas substituindo os índices contendo palavras isoladas por índices contendo SNs, o que permite utilizar modelos de classificação e ranking. A segunda alternativa

está relacionada à organização hierárquica em modelos arbóreos dos SNs, o que possibilita criar um novo conceito em termos de indexação e introduzir inovação quando se fala de interface de busca.

Depois da análise dos SNs extraídos do corpus de sua pesquisa de doutorado, Kuramoto (2006) descreve uma regra geral na formação de SNs:

> N'' = D' + N' onde:
>
> N'' é o símbolo que representa um sintagma nominal;
>
> D' representa um determinante;
>
> N' representa um predicado relacionado, ou no mínimo um nome ou substantivo;
>
> N representa um nome, um substantivo.
>
> (...) É um determinante: D' = artigos│ pronomes possessivos │ pronomes demonstrativos │determinante vazio.
>
> (KURAMOTO, 2006, p. 132)

Kuramoto (2006) também apresenta a definição de um predicado relacionado (N'):

> N' = Nome próprio │N + EP │N + SP │
>
> N representa um nome, um substantivo;
>
> EP significa expansão preposicional;
>
> EP = P + N (de fogo, de sistema)
>
> SP = P + N (da informação, do negócio etc.)
>
> P representa as preposições.
>
> (KURAMOTO, 2006, p. 131-132)

O autor alerta que a regra geral não é suficiente para descrever os SNs, ela apenas faz parte de um conjunto maior de regras.

Nos trabalhos de Kuramoto (1995-2002), os resultados obtidos por meio do protótipo comprovaram que é possível tecnicamente implementar uma interface de busca que permite a navegação por meio das estruturas arbóreas construídas hierarquicamente dos SNs.

4.2 Contribuições de Bick (2000)

Bick (2000), com seu trabalho na área de Linguística, descreve um analisador sintático (*parser*) que trabalha com a gramática e o léxico da língua portuguesa irrestritamente. O *parser*, chamado de PALAVRAS, é destinado para aplicações para observações semânticas e sintáticas, além do estudo sobre a tradução automática.

Para a construção da ferramenta, o autor formula regras gramaticais nos moldes da *Constraint Grammar* (gramática de restrições), focando na desambiguação e no tratamento de vários níveis de análise linguística. O PALAVRAS é composto por um conjunto de etiquetas altamente diferenciado; contudo o analisador atinge uma taxa de mais de 99% de acerto para análise morfológica e cerca de 97% na análise da função sintática. A análise morfológica é feita na prosa literária “O tesouro” de Eça de Queiroz e análise da função sintática é feita nos textos publicados no dia 09 de

dezembro de 1992 pela revista Veja. As taxas de acerto são referentes aos dois corpora.

Uma das inovações trazidas por Bick (2000) é permitir a visualização de árvores sintáticas, marcação em cores, entre outras opções, possibilidades dada pelo *site* do conjunto de programas, ao qual o trabalho de Bick está inserido atualmente, o VISL.

Ao optar por trabalhar com a gramática de restrições, quer-se fazer a análise do texto morfologicamente, de grupos de palavras e da composição da oração, o que permite analisar o texto em três níveis: o ortográfico, sintático e semântico.

No primeiro momento, cada constituinte da oração é etiquetado, observando-se os níveis sintáticos e semânticos por meio de um *tagger* que faz análise baseado em um léxico. Isso gera uma lista cheia de ambiguidades que é resolvida por meio de uma análise que observa o contexto da oração, quais as estruturas sintáticas que podem ser descartadas, quais são possíveis e quais são escolhidas, ou seja, eliminam-se possibilidades de estruturas sintáticas inexistentes. Aplicando-se essas regras repetida e sucessivamente, resolvem-se, em partes, o problema das ambiguidades quanto à classificação sintática da frase, permitindo que, ao final, reste apenas uma sentença e somente uma classificação para cada palavra. Isso é o que permite chamar o método do *parser* PALAVRAS de robusto.

Atualmente, além de trabalhar como *tagger* e *parser*, o PALAVRAS trabalha com desambiguadores: morfológico, sintático, de valência e de classes semânticas. Quando

sucessivamente, a sentença é submetida a estes módulos (analisador sintático, etiquetador e desambiguadores), tem-se um resultado singular para a classificação morfossintática de cada sentença. O PALAVRAS não trabalha extraindo sintagmas, mas os identifica dentro da estrutura sintática da frase.

4.3 Contribuições de Vieira et al (2000)

Vieira et al (2000), objetivando tratar a extração semiautomática de SN para resolver correferências[20] textual em língua portuguesa, constroem um corpus, constituído por 15 artigos do jornal Correio do Povo de Porto Alegre formando um conjunto de 5 mil palavras que servem como base para anotação automática de correferência em textos escritos em língua portuguesa. A resolução de correferência é uma atividade muito importante porque possibilita diversas aplicações de PLN, como citam os autores ao se referenciarem a geração automática de resumos, tradução automática, extração de informações e recuperação de informações.

Para atingir o objetivo, Vieira et al (2000) submete os textos ao *parser* PALAVRAS que gera uma árvore sintática em que os SNs se encontram, assim, a identificação de SNs é feita por esse *software*. Logo após, um *software* em

[20] Correferência é "a relação entre elementos lingüísticos que se referem a uma mesma entidade de mundo. Sua resolução tem como intuito deixar a relação entre as entidades evidenciada" (COREIXAS; VIEIRA, 2008, p. 02)

linguagem C[21] gera lista com SNs em linguagem Prolog. A lista de SNs é revista manualmente, permitindo que erros encontrados, cerca de 10%, sejam corrigidos. Desse modo, a lista é usada como entrada em um sistema de anotação automática de correferência, o que possibilitou um acerto de 50% nos resultados, considerados adequados em relação ao objetivo ao qual se propuseram a atingir. Contudo, algumas dificuldades foram encontradas pelos autores em relação às conjunções coordenativas, às vírgulas que coordenam os SNs, às dificuldades em reconhecer SP (sintagma preposicionado) e analisar expressões numéricas, às funções do pronome relativo "que" e às conjunções integrantes[22].

Esse trabalho foi desenvolvido nas áreas de Ciências Exatas e Tecnologias e Ciências da Comunicação, permitindo uma interdisciplinaridade que possibilitou ver as questões de correferência de ângulos diferentes, tanto da questão teórica quanto da aplicabilidade de ferramentas que identifique as retomadas de termos por outros.

---

[21] Linguagem de programação que permite a criação de processadores de textos, planilhas eletrônicas, sistemas operacionais, entre outros programas computacionais.

[22] São as conjunções "que" e "se", quando iniciam uma oração subordinada, em que elas funcionam sintaticamente como sujeito, objeto direto ou indireto, complemento nominal, aposto de outra oração ou predicativo.

4.4 Contribuições de Miorelli (2001)

Miorelli (2001), no Departamento de Ciência da Computação da Pontifícia Universidade Católica do Rio Grande do Sul, constrói um método chamado ED-CER para extração de SNs aplicando as regras estruturais de sintagmas nominais de Perini (1995). Com esse trabalho a autora objetiva formalizar um método para a extração de SNs em língua portuguesa, aplicando ferramentas do PLN no intuito de possibilitar uma maior funcionalidade dos SRIs. Assim, o trabalho tem como meta encontrar palavras-chave ou expressões-chave para que faça a representação dos conteúdos dos resumos de dissertações do Programa de Pós-Graduação em Ciência da Computação da Pontifícia Universidade Católica do Rio Grande do Sul, em formato digital, os quais constituem o *corpus* da pesquisa.

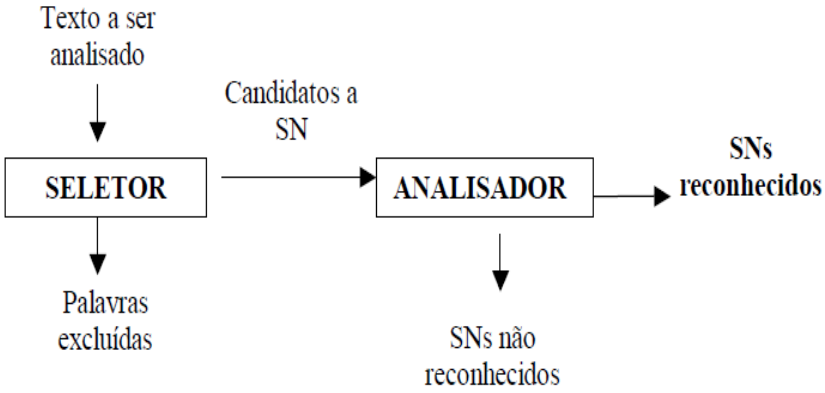
A problemática que direciona o trabalho está relacionada às palavras ideais para a tarefa de indexação.

Para validar os resultados, a autora utiliza os trabalhos de Bick (2000) e de Vieira et al (2000) no intuito de fazer uma comparação entre os seus resultados e os desses autores. Ela também faz uma retomada a alguns trabalhos correlatos como o *Nptool* de Voutilaine (1995) e o trabalho de Cardie e Pierce (1998).

O método de Miorelli (2001) é constituído de dois módulos. O primeiro é o Seletor que elimina as palavras que certamente não são SNs, fazendo a decomposição das sentenças em partes menores, de onde serão extraídos os

possíveis SNs. O segundo módulo é denominado de Analisador que analisa sintaticamente os possíveis SNs, verificando se são condizentes com a gramática de formação de SN ou não. A Figura 13 apresenta o esquema de funcionamento do ED-CER.

Figura 13 - Esquema geral de extração dos SNs no método ED-CER



Fonte: Miorelli (2001, p 47)

O funcionamento do ED-CER consiste em ler as palavras do texto da esquerda para a direita, logo em seguida há um corte e uma exclusão de palavras para que ao fim haja o reconhecimento do SN.

O processo de etiquetagem do corpus foi feita manualmente para que as ambiguidades fossem resolvidas. Miorelli (2001) ressalta que alguns detalhes devem ser observados no momento da marcação do corpus, como:

> 1) Os itens lexicais devem estar em letras minúsculas, seguidos de um sinal "_" e da etiqueta em letras

> maiúsculas. Os itens do texto, com suas respectivas etiquetas, devem estar verticalizados. [...]
>
> 2) As locuções (prepositivas, adverbiais, conjuntivas) constituídas de mais de uma palavra deverão ser reconhecidas como um único item lexical. [...]
>
> 3) As contrações devem ter suas palavras separadas. [...]
>
> 4) As siglas, anotações bibliográficas, nomes de software, *links* etc podem ser tratados como substantivos, adjetivos ou nomes próprios, dependendo da sua representatividade no texto. [...] (MIORELLI, 2001, p. 49-50)

O modulo seletor realiza duas atividades no intuito de seguir a formação do SN máximo proposto por Perini (1996): a de corte e a de exclusão de palavras. O processamento é feito lendo palavra por palavra, começando da palavra mais à esquerda.

A atividade de corte é feita quando são encontradas etiquetas que se referem ao verbo (_VB), pronomes relativos e locuções – prepositivas, comparativas, entre outras (_LG), pontuações – exceto a vírgula (_PN), ou quando a etiqueta da primeira palavra se refira à preposição (_PR) ou vírgula e conjunções aditivas, adversativas e alternativas (_CO).

A exclusão de palavras é feita em três momentos: na eliminação da palavra de corte (_VB, _LG ou PN); ou na eliminação de palavras anteriores ao corte que possuem etiquetados do tipo _PR, _AV (advérbios), _CO, _AR (artigos) e NU (numeral); ou na eliminação da primeira palavra lida, se ela for etiquetada com _PR ou _CO.

O módulo analisador reconhece ou não um candidato a SN como um SN, fazendo, inicialmente, a leitura das

palavras que constituem o candidato a SN e verificando se ele está em conformidade com a gramática definida. Dessa forma, quando um candidato não está de acordo com a gramática (ou não é reconhecido por ela) não é extraído como SN.

Miorelli (2001) utiliza regras de combinação e ocorrências das funções já definidas por Perini (1995; 1996) juntamente com as etiquetas definidas anteriormente pela autora para as categorias gramaticais, no intuito de definir os símbolos terminais (chamados de *tokens*) e símbolos não terminais conforme a Tabela 7 e Tabela 8.

Tabela 7 - Símbolos terminais para a gramática ED-CER

Símbolos terminais	
tokens	etiquetas
det	_AR _PD _PI
qua	_AJ _NU _PS
adv	_AV
con	_CO
pre	_PR
ref	_SU _PP _NP

Fonte: Miorelli (2001, p. 61)

Os símbolos não terminais são considerados intermediários na gramática para a obtenção de SN.

Tabela 8 - Símbolos não terminais para a gramática ED-CER

Símbolos não terminais
PN - Pré-núcleo
MD - modificador
AV - advérbio
NS - núcleo do SN
SN - sintagma nominal

Fonte: Miorelli (2001, p. 62)

A partir dessas informações, tem-se a gramática utilizada no método ED-CER para extrair SNs no módulo analisador conforme o Quadro 2.

Quadro 2 - Gramática do método ED-CER

NS	→	ref			
MD	→	qua			
AV	→	AV	adv		
AV	→	adv			
MD	→	AV	MD		
MD	→	MD	con	MD	
NS	→	NS	MD		
NS	→	MD	NS		
NS	→	NS	pre	NS	
NS	→	NS	pre	det	NS
NS	→	NS	con	NS	
NS	→	NS	con	det	NS
NS	→	AV	NS		
SN	→	det	SN		
SN	→	NS			

Fonte: Miorelli (2001, p. 62)

Nas palavras de Miorelli (2001, p. 63-64), o analisador, inicialmente, começa lendo a palavras mais à esquerda,

> verifica, através da etiqueta da palavra, a qual *token* a etiqueta pertence, e empilha o *token*. Os elementos da pilha são confrontados com a gramática (de cima para baixo da pilha) em busca de uma regra que reduza os elementos da pilha a um símbolo não terminal. Se existir esta regra os elementos da pilha que concordam com a regra são reduzidos ao símbolo não terminal e continuam na pilha. Este processo é executado até que todas as palavras do candidato a SN sejam lidas. Se o candidato a SN for reconhecido o que restará na pilha é um único elemento não terminal, cujo símbolo é "SN". As duas últimas regras da gramática somente são validadas quando todas as palavras do candidato a SN forem lidas.

Por fim, a autora sugere a utilização de um etiquetador que gere etiquetas definidas em seu trabalho, o que foi feito por Maia (2008). A avaliação dos resultados dos seus métodos foi feita de duas maneiras, a primeira manualmente e a segunda pelo PALAVRAS.

4.5 Contribuições de Othero (2004)

Othero (2004) desenvolve um *parser*, na área de Letras, que faz análise de sentenças simples do português, o Grammar Play. O tipo de frase analisada por essa ferramenta é a declarativa que contenha um único verbo e que não esteja na forma interrogativa, salva quando estiver estruturada por meio dos pronomes de interrogação "QU" (que, qual) no

momento em que exigem resposta do tipo sim ou não, pois a estrutura da frase continua a mesma.

O Grammar Play diz se a sentença é gramatical ou agramatical, atribuindo uma estrutura sintagmática e se baseando no modelo da teoria X-barra[23] tradicional. A interface do *parser* apresenta a estrutura das sentenças em formato de colchetes rotulados e também uma representação arbórea.

A gramática utilizada foi *Phrase Structure Grammar* (PSG) livre de contexto acrescida de traços lexicais. O autor adapta todas as regras de descrição de sentenças simples em português brasileiro à linguagem Prolog, por meio da gramática DCG (*Definite Cause Grammar*). O léxico utilizado foi o corpus DELAS_PB do NILC que contém 67.500 palavras simples da língua portuguesa. O Grammar Play faz análise sintática por meio de técnicas de *parsing top-down*.

Os resultados atingidos por Othero (2004) com o Grammar Play não foram adequados para o modelo de gramática X-barra, a qual o autor adotou. Contudo, Othero (2004) sugere que o *parser* possa ser usado como uma ferramenta em cursos de Linguística Formal, Linguística Computacional, Sintaxe e Processamento do Português. Além disso, o autor aponta os problemas apresentados pelo

[23] Ramo da Gramática Gerativista que trabalha com a inclusão de categorias de nível intermediário existentes entre as categorias lexicais e as categorias sintagmáticas.

Grammar Play que servem para nortear estudos de implementação computacional de modelos teóricos, no intuito de que se atinja uma maior eficiência e compatibilidade com o conhecimento corrente dos modelos gramaticais.

Vale salientar que o objetivo do autor com o desenvolvimento do Grammar Play não era disponibilizar uma ferramenta robusta de análise morfossintática em textos escritos em língua portuguesa, mas implementar regras linguísticas coerentes baseadas em hipóteses teóricas consensuais em teoria linguística e em resultados de descrição sintática do português brasileiro.

4.6 Contribuições de Souza (2005)

Souza (2005), desenvolvendo pesquisa na área de Ciência da Informação e com intuito de criar uma metodologia que seleciona descritores automaticamente em textos digitais escritos em língua portuguesa por meio de SNs, trabalha com metodologias (prospectiva e consolidada) que consistem em processos e seus produtos.

O primeiro processo da metodologia prospectiva de Souza (2005) está relacionado à escolha de um corpus inserido em uma área de domínio, o que gera um produto com textos originais digitais. O corpus inicialmente utilizado por este autor é composto de quinze artigos trabalhados por Kuramoto (1999), no intuito de avaliar a metodologia. Para validação da metodologia, Souza (2005) trabalha com um corpus composto por 60 documentos extraídos de periódicos

da área de CI, 29 artigos do periódico DataGramaZero e mais 31 artigos do periódico Ciência da Informação, publicados entre 2002 e 2003 e escritos em língua portuguesa.

O segundo passo consiste em converter esses textos, que geralmente estavam em formatos como PDF ou HTML, em formato de texto simples, resultando em outro produto. No terceiro momento, o autor retira os resumos e as palavras-chave atribuídas pelos autores dos artigos, tal tarefa foi considerada por Souza (2005) como um artifício metodológico com objetivo de poder se verificar, ao fim da metodologia, o sucesso do procedimento automático de extração de descritores, comparando os SNs extraídos automaticamente com as palavras-chave atribuídas pelas autorias dos artigos.

A extração dos sintagmas nominais do corpo dos textos acontece na quarta etapa da metodologia de maneira quase que inteiramente automática, utilizando o *parser* PALAVRAS e o *software* Xtractor. O produto gerado dessa extração foram arquivos em formato HTML os quais continham os SNs ordenados de acordo com a ocorrência nos textos originais, a partir daí, foram criadas planilhas (em MICROSOFT EXCEL) contendo pastas específicas para cada texto, tendo essas planilhas como o sexto produto da metodologia.

A quinta ação é ordenar os SNs nas planilhas observando a frequência de ocorrência dos SNs nos documentos. Já a sexta atividade é o descarte de SNs que apresentem valores de frequências de ocorrências inferiores a um valor preestabelecido. Na sétima etapa, é feito o agrupamento dos SNs remanescentes, ainda manualmente,

observando os determinantes iniciais e a representação assumida pelo agrupamento de acordo com as regras de construção de tesauros (SOUZA, 2005, p. 79-80). Essa etapa gerou uma planilha com os SNs agrupados, tendo sido descartados os que não atingiram ao patamar preestabelecido.

No oitavo procedimento, Souza (2005) analisa manualmente os SNs pré-escolhidos e decide sobre sua relevância como descritores, no intuito de se criar uma lista de *stopwords*. A incidência dos SNs foi verificada na nona ação, partindo do pressuposto de quanto maior a incidência de um SN no conjunto de documentos, menor é a sua relevância como descritor.

No décimo procedimento, são analisados a estrutura e o nível dos SNs, o que é muito importante na análise de relevância. No décimo primeiro momento, o autor verificou a ocorrência destes SNs em tesauro específico. A avaliação da relevância dos SNs como descritores é feita no décimo segundo procedimento, levando em consideração a frequência de ocorrência dos SNs no texto do documento, a incidência dos SNs no conjunto de documentos, seus níveis, suas estruturas sintáticas e sua ocorrência em tesauro especializado, no caso de Souza (2005), o tesauro está relacionada à área de domínio da CI. O produto desta etapa é a ordenação dos SNs, segundo critérios de relevância estabelecidos, nas tabelas das planilhas com os candidatos a descritores.

A penúltima etapa da metodologia é a comparação entre as palavras-chave e resumos dos documentos originais

e os SNs escolhidos como descritores. A última fase é a análise feita por especialistas. As partes da metodologia de avaliação utilizadas por Souza (2005) são mais detalhadas na seção “Metodologias de avaliação de extração automática de SNs”. Já na metodologia consolidada, têm-se as seguintes etapas:

> 1. Escolher *corpus* significativo de documentos reconhecidamente inseridos dentro de uma área de conhecimento, como universo empírico desta pesquisa;
>
> 2. Converter os formatos de arquivo para texto simples;
>
> 3. Retirar os resumos e as palavras-chave atribuídas pelos autores;
>
> 4. Extrair os sintagmas nominais do corpo do texto;
>
> 5. Ordenar os SNs nas planilhas através da verificação da freqüência de ocorrência dos sintagmas nominais nos documentos;
>
> 6. Descartar os SNs que apresentam freqüências de ocorrência inferiores a um patamar preestabelecido;
>
> 7. Agrupar os SNs remanescentes a partir dos determinantes em suas formas “canônicas”, e reordená-los;
>
> 8. Analisar manualmente os SNs pré-escolhidos e decidir sobre a sua relevância como descritores, para fins de construção de uma *stoplist* e verificar se algum SN escolhido consta em uma *stoplist*, dinamicamente construída, para, se for o caso, descartá-lo (em 11);
>
> 9. Verificar a incidência dos SNs nos outros documentos do *corpus*;
>
> 10. Analisar a estrutura e o nível dos SNs;
>
> 11. Atribuir pontuação e ranquear os SNs remanescentes de acordo com fórmula estabelecida [...], levando em conta a freqüência de ocorrências no texto e a freqüência de saturação

> definida, e a quantidade de textos do corpus em que ocorrem, a estrutura sintática e o nível do SN. Esses critérios de relevância são regidos por parâmetros [...] a serem sintonizados com a sucessiva aplicação da metodologia;
>
> 12. Em caso de “empates” nos valores da pontuação dos SNs, considerar a ocorrência no tesauro da CI como fator de desempate;
>
> 13. Caso ainda ocorram “empates” nos valores da pontuação dos SNs, considerar os seguintes critérios de desempate:
>
> a. Maior valor absoluto da freqüência de ocorrência;
>
> b. Menor valor absoluto da ocorrência no *corpus*;
>
> c. Maiores nível e estrutura do SN;
>
> d. Maior quantidade de letras do SN;
>
> 14. Apresentar tantos descritores quanto forem desejáveis, a partir da lista ranqueada de candidatos a descritores. (SOUZA, 2005, p. 121-122)

Os resultados obtidos pelo autor foram satisfatórios, pois a utilização dos SNs como descritores apresenta mais vantagens, já que eles são mais significativos do que as palavras-chave isoladas e são carregados de significados contextuais acerca da semântica dos discursos. Souza (2005) ressalta que a extração automática de SNs se mostra extremamente viável, apesar de não estar em “pé de igualdade” na comparação qualitativa com a extração manual.

4.7 Contribuições de Santos (2005)

Santos (2005), na área de Ciência da Computação, apresenta uma nova abordagem de molde de regras para o TBL (Aprendizado Baseado em Transformações), o termo atômico com restrição (TA com restrição) que consiste na verificação de uma possível dependência entre a preposição a ser classificada e o verbo precedente, dessa maneira, geram-se regras específicas para observar a relação dessa preposição e o verbo antecedente. Isso é importante ser observado porque muitos SNs que estão na função de objetos são precedidos por uma preposição, estando inseridos dentro de um SP.

Assim, o objetivo do TA com restrição é capturar o valor de um traço X de um item, se outro traço Y, no mesmo item, atender a um determinado teste condicional. O teste é feito verificando se o valor do traço Y é igual a um valor predefinido, possibilitando a especificação de regras que tenham função de restringir itens específicos, assim como gerar regras que verificam apenas os elementos importantes de um problema que esteja sendo tratado.

Para fazer o teste, o autor escolheu como corpus o Mac-Morpho que é constituído de textos de 10 cadernos do jornal Folha de São Paulo e que constitui um corpus do NILC. Dentro desses textos, foram reconhecidos os SNs, assim, cada palavra do Mac-Morpho recebeu a etiqueta SN. Para tanto, o autor desenvolveu alguns programas em linguagem Java.

O primeiro programa, chamado de *RemoveTags.Java*, recebe como entrada um arquivo no formato do Mac-Morpho, no qual o texto está etiquetado e apresentado verticalmente, e gera um arquivo apresentado horizontalmente, eliminando-se as etiquetas morfossintáticas e se realizando as concatenações das palavras, como por exemplo a junção da preposição “de” com o artigo “o” (de o => do). Os arquivos gerados pelo *RemoveTags.Java* foram submetidos ao *parser* PALAVRAS. A partir dos resultados obtidos pelo *parser* PALAVRAS, o autor criou dois conjuntos de etiquetas, tendo como primeiro aquele que inclui as etiquetas que poderiam iniciar um SN e o segundo como o que contem as etiquetas que poderiam estar dentro de um SN.

Dessa forma, foi desenvolvido um programa que identifica os SNs a partir das anotações feitas pelo *parser* PALAVRAS, o *IdentifySNs.java*, que recebe como entrada o arquivo gerado pela análise sintática e tem como saída um arquivo com o texto na forma vertical, em que cada linha contém uma palavra e sua etiqueta SN separadas pelo caractere “_” (SANTOS, 2005, p. 63). Contudo, tal programa apresenta falhas ao não identificar SNs que contenham orações subordinadas adjetivas e/ou que contenham vírgula, dessa forma, o SN é quebrado no ponto em que essas situações acontecem. Para resolver o problema, o autor inclui um pseudocódigo que não atinge a 100% de solução, mas consegue resolver problemas dessa natureza na maior parte dos casos.

O terceiro programa desenvolvido por Santos (2005) é o *AttachSNtag.java* que acrescenta a etiqueta SN a cada

arquivo do Mac-Morpho, permitido pela comparação com arquivo gerado pelo *IdentifySNs.java*, tendo como entrada os arquivos destes e como saída um arquivo etiquetado com a etiqueta SN. O *AttachSNtag.java* pode gerar, também, um arquivo em HTML " no qual o texto aparece na horizontal, sem etiquetas e com os SNs destacados pelas cores azul e vermelho (o vermelho é usado para identificar um SN que aparece imediatamente após um outro)" (SANTOS, 2005, p. 64). Outra característica do *AttachSNtag.java* que o diferencia do *IdentifySNs.java* é a possibilidade de identificar o momento em que se devem ser incluídos nos SNs os parênteses ou as aspas.

O quarto programa criado por Santos (2005) é o *FormatCorpus.java* que formata os arquivos do Mac-Morpho depois de terem passado pelos outros três programas (*RemoveTags.Java*, *IdentifySNs.java* e *AttachSNtag.java*) de maneira que separa as sentenças por uma linha em branco, considerando os caracteres "." (ponto final), "!" (ponto de exclamação) e "?" (ponto de interrogação) com delimitadores das frases. Outra tarefa realizada pelo *FormatCorpus.java* é retirar algumas etiquetas complementares e indicadores de ênclises[24], contrações[25], disjunção[26] e mesóclise[27], as quais não eram importantes para o que o autor se propõe trabalhar.

---

[24] Quando um pronome oblíquo átono (me, te, o, os, a, as, lhe,, lhes, nos e vos) é colocado depois do verbo.

[25] A junção de duas palavras que acabam perdendo fonema (de + a = da).

[26] É falta de conectivos entre orações ou palavras coordenadas, pode ser definido também como a substituição de constituinte por outro (ex. olhos

Assim, com o uso desses quatro programas, Santos (2005) cria dois corpora de treino (compostos por todos os textos dos cadernos jornalísticos do Mac-Morpho) e um para testes (composto pelos cadernos Agricultura, Brasil, Cotidiano, Dinheiro, Esportes e Ilustrada).

Para a identificação de SNs, na classificação inicial, atribui-se uma etiqueta SN a cada item do corpus de teste, usando estatísticas de coocorrências observadas nos corpora de treino a partir do aprendizado de máquina (TBL). Para Santos (2005), houve uma redução de tempo de treinamento e uma melhora nos resultados obtidos, quando se usou o método lexicalizado na classificação inicial. Dessa maneira, o autor usa alguns moldes de regras[28] para identificação de SNs, tendo os resultados obtidos por meio de equações que encontram a *accuracy* (precisão geral) da classificação item a item, precisão de identificação de SNs, abrangência da identificação de SNs e à medida que é calculada a partir das equações de precisão de identificação de SNs e abrangência,

---

grandes, nariz afilado, cabelos lisos; a perfeição.).

[27] Quando um pronome oblíquo vem entre o verbo e a terminação verbal (far-**se**-á).

[28] Os *moldes de regras* determinam os tipos de expressões condicionais que comporão as regras. Dessa forma são eles que determinam o espaço de regras que podem ser geradas, e por isso são os elementos que provocam maior impacto no comportamento do aprendizado com TBL. Os moldes de regras devem codificar as informações linguísticas do contexto que são importantes para a classificação de um item, informações estas que são dependentes do problema tratado. (SANTOS, 2005, p. 67)

como podem ser vistas no Quadro 3. A avaliação se dá entre os SNs identificados pelo *parser* PALAVRAS e os que o classificador TBL encontra.

Quadro 3 – Equações utilizadas por Santos (2005)

$$\text{Precisão geral} = \frac{\text{Total de itens classificados corretamente}}{\text{Total de itens}}$$

$$\text{Precisão} = \frac{\text{Total de SNs classificados corretamente}}{\text{Total de SNs identificados}}$$

$$\text{Abrangência} = \frac{\text{Total de SNs classificados corretamente}}{\text{Total de SNs existentes no corpus}}$$

$$F_{\beta=1} = \frac{(\beta^2 + 1) * \text{Precisão} * \text{Abrangência}}{\beta^2 * \text{Precisão} + \text{Abrangência}}$$

Fonte: adaptado de Santos (2005)

Essas métricas foram utilizadas por Santos (2005) para avaliar o desempenho do seu método na identificação de SNs.

Os resultados alcançados por Santos (2005) permitiram alcance do objetivo do autor que era criar uma ferramenta de identificação de SNs do português brasileiro com o uso da técnica de aprendizado de máquina. Ele compara os seus resultados com os obtidos em inglês por Ramshaw e Marcus (1995), dizendo que abordagem utilizada foi diferente no tocante a identificação de SNs contendo os pós-modificadores adjetivos e SPs. O que o diferencia do

estudo de Ramshaw e Marcus (1995) é o uso de apenas SNs básicos, pois Santos (2005) objetivava identificar os SNs mais complexos. Para evitar erros com o uso dos moldes, Santos (2005) cria os Termos Atômicos (TA) com restrição como foi apontado no início dessa subseção com os objetivos de tratar problemas de PLN que requerem tamanhos variáveis e de gerar novas regras que corrijam erros específicos. Dessa maneira, os resultados atingidos pelo autor foram mais eficientes do que uso dos conjuntos de moldes que usavam apenas TAs tradicionais. Santos (2005) não cria uma metodologia de seleção automática de SNs, contudo é percebido que o autor aponta para essa necessidade, quando faz sugestões para trabalhos futuros.

4.8 Contribuições de Arcoverde (2007)

Arcoverde (2007), com o objetivo de representar textos, constrói um modelo híbrido que utiliza dois tipos de conhecimento, o linguístico e o estatístico, aplicados a sistemas de RI, o que possibilita um processo de pós-filtragem de informação. Na construção do modelo, o autor usa a Categorização de Textos[29] integrado a um sistema de RI, nos quais, recursos de PLN proporcionassem ao usuário experiências na busca de informações relevantes. Seu projeto foi desenvolvido na área de Ciência da Computação.

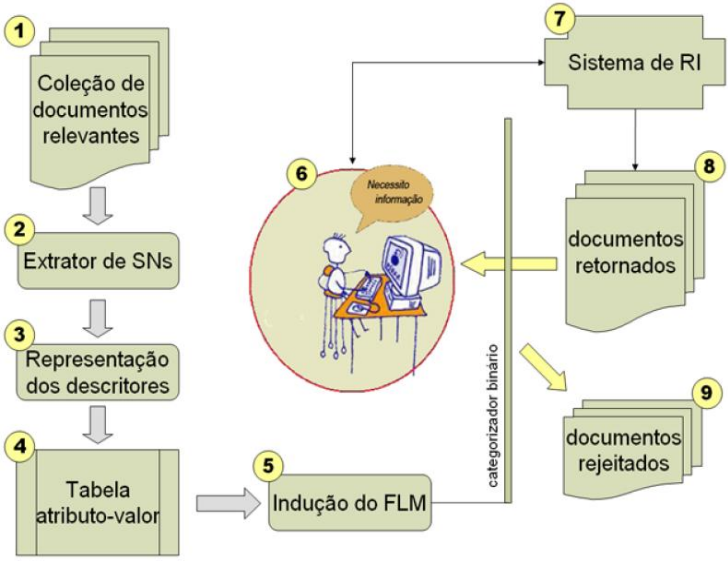
[29] Técnica que classifica automaticamente documentos textuais em formato digital, de acordo com categorias pré-estabelecidas.

O corpus utilizado pelo o autor é composto de uma coleção de documentos extraídos da *Web* escritos no português brasileiro (CLEF 2006 - *Cross Language Evaluation Fórum*) e um conjunto de tópicos de consultas e julgamentos de relevantes. Arcoverde (2007) usa as mesmas técnicas que Santos (2005), Aprendizagem de Máquina, especificamente com o método de Categorização de Textos, para que houvesse o processo de Filtragem de Informação, permitido por uma representação de documentos baseada no uso de SNs como descritores multitermos com alto poder informativo. Estes descritores são utilizados para compor tabelas de atributo-valor, o que permite uma melhora no desempenho de classificação automática de texto.

Os tópicos para a realização do experimento de classificação automática de textos são retirados do CLEF 2006, observando alguns critérios de seleção, o que permitiu a seleção de apenas 8 dos 50 tópicos disponíveis. Arcoverde (2007) ressalta que a quantidade não é ideal para uma base de treino, mas por questões de delimitações de pesquisa, os tópicos permitiram avaliar o resultado da filtragem.

Arcoverde (2007) defende que a elaboração do espaço de descritores formado por SNs possibilitaria empregar técnicas avançadas de reconhecimento de padrões em textos, relacionando certos descritores como, por exemplo, “Presidente Fernando Henrique Cardozo” e “Pres. Cardoso, F. Henrique”. A Figura 14 apresenta a arquitetura do subsistema de filtragem formulado por Arcoverde (2007).

Figura 14 – Arquitetura do subsistema de filtragem em conjunto com sistema de RI



Fonte: Arcoverde (2007, p. 60)

Segundo a descrição feita pelo o autor, no primeiro momento, há um conjunto de documentos relevantes que sintetiza o interesse do usuário em determinado tópico. Na segunda etapa, cada documento é processado pelo identificador de SN. Na pesquisa de Arcoverde (2007), são utilizados identificadores construídos por técnicas de Aprendizado de Máquina como o de Santos (2005). Na terceira fase, há a escolha e construção dos descritores, assim como o respectivo cálculo dos pesos. Já na quarta etapa, é feita uma tabela atributo-valor que é definida pelo autor como a representação estruturada dos documentos que determinam o perfil de busca do usuário. A indução de Filtro Linguisticamente Motivado (conceito introduzido pelo autor) é

feita na quinta fase e é realizada por meio de técnicas de aprendizagem de máquina supervisionada, tendo suas características dependentes da família de algoritmos adotada.

Na sexta etapa, o usuário explora o resultado por meio da elaboração de sua necessidade informacional. Já na sétima etapa, o sistema de RI acessa sua base de dados para atender a consulta do usuário, o que já remete a oitava etapa caracterizada como a devolução de um conjunto de documentos ordenado por algum critério de relevância. Na última fase, o Filtro Linguisticamente Motivado bloqueia os documentos irrelevantes a partir da inferência entre decisões sobre a semelhança entre cada documento e o perfil de busca do usuário.

Os resultados atingidos por Arcoverde (2007) foram satisfatórios, pois permitiram afirmar que a Categorização de Textos atua no processo de pós-filtragem de documentos permitindo uma precisão no bloqueio aos documentos cujo conteúdo não esteja de acordo com os interesses do perfil de busca expresso pelo usuário. Como sugestão para trabalhos futuros, o autor aponta para a replicação da experiência em um ambiente real, o que permitiria que a técnica enriquecesse a experiência do usuário na busca por informação relevante.

Embora Arcoverde (2007) não trabalhe exclusivamente com técnicas de extração automática de SNs, seu trabalho é importante, pois também utiliza a abordagem de Santos (2005) com uso de Aprendizagem de Máquina para fazer a extração automática de SNs, além de permitir ver uma aplicação prática do uso dos SNs como descritores em sistemas de RI.

4.9 Contribuições de Costa (2007)

Costa (2007), com seu projeto na área de Ciência da Computação, apresenta uma gramática computacional para o português, chamada de LXGram[30] e que está em desenvolvimento na Universidade de Lisboa. Os objetivos do LXGram são analisar frases no intuito de se produzir uma descrição formal de seu significado e gerar frases a partir de representações do significado. O autor foca na modelagem e implementação da sintaxe e semântica de SN de língua portuguesa.

Para tanto, Costa (2007) usa a plataforma *Linguistic Knowledge Builder* (LKB) para desenvolver o LXGram, pois esta plataforma utiliza algoritmos eficientes de análise e geração. Mas, a priori, o autor descreve a forma como os vários elementos constituintes de SN podem combinar uns com os outros, formando novos SNs ou apenas invertendo posições. Como os outros autores, Costa (2007) defende o uso de SNs no PLN no intuito de haver representações de significado desses SNs.

[30] LXGram foi desenvolvido junto ao NLX-Group (Natural Language and Speech Group) da Universidade de Lisboa. Essa gramática faz parte das ferramentas do LXSuite que será descrito na seção “Ferramentas de Extração automática de SNs”. O LXGram está disponível em: <http://nlxgroup.di.fc.ul.pt/lxgram/>. Acesso em: 15 jun. 2013.

Na constituição do LXGram, é usada a *Head-Driven Phrase Structure Grammar* (HPSG) que é uma estrutura gramatical adequada a aplicações computacionais, já que seus recursos são fundamentais para construção de linguagens de programação. O HPSG foi utilizado para modelar a sintaxe dos SNs, apontando as restrições sobre o padrão de ordem dos elementos constituintes dos SNs e sua ocorrência. Dessa maneira, um sistema de restrições é desenvolvido para dar conta dos fenômenos linguísticos como a estrutura do SN e a semântica.

Os resultados obtidos por meio de testes feitos em 186 itens (frases extraídas da própria dissertação) são satisfatórios, pois se trata de uma abordagem nova e que pode ser aplicado para desenvolvimento de *parsers* como o LXParser.

4.10 Contribuições de David (2007)

David (2007), no intuito de analisar expressões nominais da língua portuguesa, desenvolve um programa de computador, no Programa de Pós-Graduação em Linguística da Universidade Federal do Ceará, que atribui a estas expressões sua estrutura de constituintes e sua representação por meio de colchetes rotulados e árvores. Outro objetivo expresso pela autora é testar a hipótese de que

sintagmas determinantes[31] têm como núcleo pronomes pessoais.

Como corpus, David (2007) usa 36 expressões nominais extraídas do *corpus* NILC, dentre estas expressões, há algumas que têm na sua estrutura interna pronomes pessoas que assumem o papel de DET, tendo os SNs como complemento. Para coletar os dados, a autora usa um programa de computador chamado de "O Constructor" [32] que é uma ferramenta interativa e que permite que usuários leigos façam pesquisa em corpora que foram codificados em ambiente IMS-CWB. A partir daí, a autora dirige comando ao O Constructor com o objetivo de que ele procure no corpus do NILC as 36 expressões elencadas, levantando todas as ocorrências.

Dessa maneira, é feita uma descrição linguística aplicando os pressupostos da teoria X-barra "contemporânea". Assim, cria-se um fragmento de gramática computacional do português brasileiro que gera, reconhece e analisa possíveis expressões validadas por esta língua.

Essa descrição é feita por meio da representação arbórea da estrutura das expressões nominais, o que permite que seja constituído um fragmento de gramática livre de contexto. A partir desta gramática, a autora aplica as mesmas

---

[31] Um DET (pronome pessoal) + SN = DP (sintagma dominante). Um exemplo retirado do trabalho da autora é "A percepção como, **vocês psicólogos** a estudam, não pode afinal de contas ser diferente da observação física, pode?" (DAVID, 2007, p. 54)

[32] O CONSTRUCTOR. Disponível em: <http://www.reocities.com/Athens/crete/1546/indexc.htm>. Acesso em: 20 jun. 2013.

regras a um compilador Prolog, utilizando o formalismo DCG (*Definite Cause Grammar*) para transformar a gramática descrita em gramática livre de contexto (*Context Free Grammar* – CFG) dentro de um reconhecedor gramatical automático.

A partir destas transformações gramaticais, são acrescentados à gramática em DCG de David (2007) traços morfossintáticos aos itens lexicais, no intuito de que o analisador sintático feito pela autora posteriormente fornecesse expressões nominais que tivessem suas estruturas gramaticais reconhecidas por meio dos traços morfossintáticos unificados.

Os resultados alcançados foram bons, pois o analisador gramatical elaborado por David (2007) consegue gerar e reconhecer expressões nominais em língua portuguesa, assim como analisar as estruturas dessas. Outra questão atingida pela autora é que os pronomes pessoais “eu, você, nós e vocês” são núcleos de um sintagma determinante que selecionam sintagmas nominais como complemento.

4.11 Contribuições de Morellato (2007; 2010)

Em sua monografia do curso de bacharelado em Ciência da Computação, Morellato (2007) desenvolve um sistema identificador de SNs, o qual foi denominado pela autora de SIDSN. Ao construir esse sistema, almejava que fosse utilizado em ferramentas de PLN. Assim, o objetivo principal do trabalho era confeccionar um sistema que fosse

capaz de identificar e extrair SNs automaticamente em textos escritos em língua portuguesa.

SIDSN é conceituado de conjunto de programas que faz o reconhecimento e o retorno de SNs existentes em um texto. Essa ferramenta possui dois módulos chamados de Pré-processador e Identificador de Sintagmas Nominais. No primeiro módulo, submetem-se textos digitais que estão em formato *.txt* na entrada do sistema que retoma um arquivo com frases que estão em forma de lista de palavras. No segundo módulo, como entrada o Identificador de Sintagmas Nominais recebe as frases fornecidas pelo Pré-processador e verifica se cada uma está de acordo com as regras definidas da língua portuguesa, assim, ao final desse processo, o programa retorna uma lista com os SNs como as respectivas funções desempenhadas nas frases.

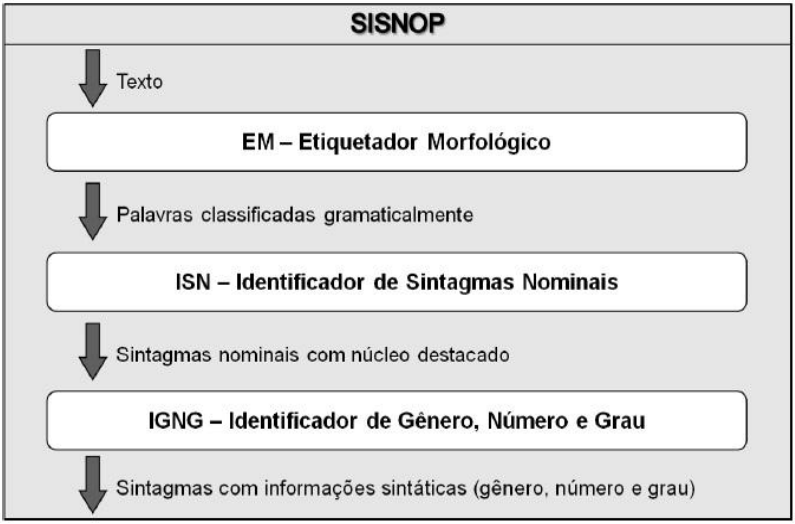
A linguagem utilizada para fazer a programação foi o PROLOG, usando como formalismo a gramática livre de contexto (DCG) associado ao analisador *top-down*. Para obter os resultados, foram utilizados diversos textos em variados gêneros discursivos, encontrados em *sites* de busca, assim o *corpus* da pesquisa era constituído de notícias jornalísticas, *blogs*, artigos científicos, teses de doutorados e livros literários. Apesar de atingir um resultado satisfatório na identificação dos SNs, a autora sugere que trabalhos futuros reduzam o tempo de processamento do algoritmo, assim como o aprimoramento das regras de identificação dos SNs (MORELLATO, 2007, p. 26-34).

Seguindo a mesma linha de pesquisa da sua monografia, Morellato (2010), em sua dissertação, propõe

uma metodologia de identificação automática de SNs. Para tanto, a autora tem como metas: analisar metodologias apresentadas em trabalhos correlatos na identificação de SNs em língua portuguesa, levando em consideração a representação linguística dos sintagmas e os métodos de implementação; implementar um sistema que recupere SNs em textos escritos na referida língua natural e determine as informações sintáticas dos elementos constituintes dos sintagmas; além de validar o sistema através da eficiência da identificação de SNs em corpora compostos por grandes quantidades de dados, observando o tempo de processamento, a precisão dos resultados.

Diferentemente do seu trabalho anterior, em que os resultados foram melhores com grupos de textos de uma mesma área, agora a autora quer que essa eficiência se dê em análise de grandes corpora constituídos de diversos gêneros discursivos sem restrição a área de domínio. Dessa maneira, é criado do SISNOP – Sistema Identificador de Sintagmas Nominais do Português, constituído por um conjunto de três módulos: etiquetador morfológico, identificador de sintagmas nominais e identificador de gênero, número e grau, como podem ser visto na Figura 15.

Figura 15 – Estrutura em módulos do sistema SISNOP



Fonte: Morellato (2010, p. 50)

No módulo Etiquetador Morfológico, é feita a etiquetagem e classificação de cada palavra, usando a categorização gramatical do programa FORMA[33].

Já no módulo Identificador de Sintagmas Nominais, são identificados os SNs das sentenças. Esse módulo delimita as frases nos textos, assim como implementa uma gramática e analisa sintaticamente se cada frase está de acordo com essa gramática. Nesse módulo, foi utilizada a ferramenta Bison que é um *parser* (do tipo *bottom-up*) conversor de gramática livre de contexto em analisador sintático.

[33]TR+. FORMA. Disponível em: <http://www.inf.pucrs.br/~gonzalez/tr+/forma/>. Acesso em 16 jul. 2013.

O Identificador de Sintagmas Nominais percebe a terminação da frase quando encontra com um sinal de pontuação. A gramática utilizada nesse modo foi uma livre de contexto, que se comporta recebendo como entrada as palavras classificadas gramaticalmente e verificando se existe uma regra pela qual seja possível gerar frase, de modo que informa na saída os SNs. Nesse módulo, foi interessante a aplicação de dois métodos devido à complexidade das regras gramaticais, a janela deslizante e a janela deslizante recursiva.

No módulo Identificador de Gênero, Número e Grau são adicionadas aos SNs extraídos informações relacionadas às flexões como gênero (masculino e feminino), número (singular e plural) e grau (aumentativo e diminutivo).

O corpus utilizado para os testes foram 186 frases que foram usadas nos testes do LXGram de Costa (2007). Os resultados foram considerados pela autora de boa precisão e abrangência e aponta que metodologia pode ser melhorada.

4.12 Contribuições de Maia (2008)

Maia (2008) investiga a utilização de SNs pontuados como atributos na classificação por similaridade e agrupamentos de documentos digitais. Para tanto, o autor usa técnicas e algoritmos de análise de texto para compor o OGMA que tem como objetivo automatizar a extração de SNs e o cálculo do peso de cada termo na indexação dos documentos, o que será usado como entrada para cada

método proposto pelo autor. Para classificar e agrupar documentos foram utilizados o OGMA e a ferramenta WEKA (*Waikato environment for knowledge analysis*).

O OGMA foi desenvolvido na ferramenta Visual Studio.NET em linguagem C# e tem um léxico adaptado dos arquivos utilizados pelo BR/ISPELL[34]. Para a identificação dos verbos, foi usada a ferramenta Conjugue[35] mais uma base de dados de 5000 verbos, o que permitiu uma tabela de 292.720 formas verbais. Maia (2008) se baseou na gramática de Tufano (1990) para compor a lista de *stopwords*.

O léxico utilizado permite que o OGMA etiquete cada palavra do texto com as possíveis categorias gramaticais correspondentes. O problema da ambiguidade é resolvido a partir de uma lista com todas as combinações de etiquetas encontradas para palavras de uma sentença, tendo cada combinação submetida às regras de extração de SNs. Por exemplo, na frase: "o mato estava grande", em que a palavra "mato" morfologicamente pode ser um verbo ou um substantivo dependendo do contexto e da posição, o OGMA etiquetaria da seguinte forma "O/**AD** mato/**VBSU** estava/**VB** grande/**AJ**". Depois, o OGMA submete às regras de extração as duas versões da frase etiquetada com diferentes combinações de etiquetas:

[34]Dicionário BR/ISPELL. Disponível em: < http://www.ime.usp.br/~ueda/br.ispell/>. Acesso em: 02 abr. 2013.

[35]Ferramenta que conjuga verbos da língua portuguesa, a partir de um banco de paradigmas. Conjugue. Disponível em: < http://www.ime.usp.br/~ueda/br.ispell/conjugue.html>. Acesso em: 02 abr. 2013.

"O/**AD** mato/**VB** estava/**VB** grande/**AJ**"; e "O/**AD** mato/**SU** estava/**VB** grande/**AJ**".

O OGMA usa um conjunto de regras para extrair SNs, conforme o Quadro 4.

Quadro 4 – Regras de extração de SN do OGMA

AR ← AD	de ← AR	AV ← AV ad	re ← SU	NS ← MD NS
AR ← AI	de ← PD	MD ← AV MD	de ← PP	NS ← NS pr NS
AJ ← VP	de ← PI	MD ← MD co MD	re ← NP	NS ← NS pr de NS
NU ← NR	qu ← AJ	NS ← NS MD	NS ← re	NS ← NS co NS
NU ← NC	qu ← NU	co ← CO	MD ← qu	NS ← NS co de NS
CO ← VG	qu ← PS	pr ← PR	SN ← NS	NS ← AV NS
CO ← CJ	ad ← AV		AV ← ad	SN ← de SN

Fonte: Maia (2008)

A ferramenta OGMA aplica regra por regra na ordem de leitura até obter um SN, atuando sobre as etiquetas, no intuito de marcar o início e fim do SN em cada frase do texto. Essas regras são baseadas no ED-CER (MIORELLI, 2001), mas com uma alteração: "nova regra: ***de*** ← ***PP*** substituindo a regra do ED-CER ***re*** ← ***PP*****".** A modificação da regra foi feita no intuito de melhorar a extração de SN com base em diversos testes executados (MAIA, 2008).

O OGMA, então, foi projetado para atender as metodologias prospectiva e consolidada do autor, as quais são compostas pelos seguintes objetivos: extrair os SNs; atribuir pesos a estes de acordo com a frequência em que aparecem no texto e dentro de outros SNs; identificar a categoria gramatical do SN extraído a partir da metodologia proposta por Souza (2005); calcular a pontuação de cada SN

extraído de acordo com a mesma metodologia de Souza (2005), extrair termos, exceto os constantes na lista de *stopwords*, atribuindo pesos de acordo com a frequência no texto; e calcular a similaridade entre duas listas de termos (extraídas do documento), utilizando o coseno. Vale ressaltar que Maia (2008), além de usar a metodologia de Souza (2008), também usa as percepções da metodologia do ED-CER (MIORELLI, 2001).

Maia (2008) usou dois corpora para aplicação da ferramenta. O primeiro corpus é composto por 50 artigos selecionados do Encontro Nacional de Pesquisa em Ciência da Informação (ENANCIB 2005). O segundo corpus é constituído de 160 textos extraídos do site do Jornal Hoje em Dia, Jornal do Brasil e Folha de São Paulo (esses dois últimos fazem parte do corpus do Temário) compondo quatro temas: Informática, Turismo, Veículos e Mundo.

Com os resultados, o autor conclui que os métodos de descoberta de conglomerados se mostram dependentes de técnicas que façam o pré-processamento dos textos, com o objetivo de padronizá-los, minimizando os problemas quanto ao vocabulário que, em muitos casos, traz ambiguidades nas análises feitas pelas ferramentas.

Salienta-se que os resultados obtidos por Maia (2008) são muito importantes para o estudo sobre a extração de SN, pois permitiu a criação de uma ferramenta de extração de SNs em português brasileiro, a automação do método de pontuação da metodologia de Souza (2005) que é usado como entrada para o método de seleção de SNs descritores.

4.13 Contribuições de Pinheiro (2009)

Com o objetivo de possibilitar uma solução flexível e eficiente para o problema de usar SNs como descritores na etapa de pré-processamento em Mineração de Textos[36], Pinheiro (2009) faz um estudo de caso em que coloca em prática o uso do algoritmo K-Means[37], tendo SNs como características dos vetores documentos, obtendo bons resultados em comparação com o K-Means ao utilizar palavras isoladas (Bag-of- Word, BOW), tendo como corpus para testes 855 documentos relacionados a esporte, imóveis, informática, política e turismo. Já no estudo de caso, o autor faz a mineração de opinião (de texto), usando a metodologia proposta por si, em uma base com 2000 arquivos no formato de "*e-mail*". Esses e-mails são de clientes corporativos da empresa Light S.A. no período entre 2007 e 2008.

O processo proposto por Pinheiro (2009) identifica e cria as palavras indesejadas (as *stopwords*) no primeiro momento. Diferentemente da abordagem tradicional em que as *stopwords* são os artigos, preposições, interjeições, conjunções, pronomes, sinais de pontuação e os caracteres indesejáveis, a abordagem proposta pelo autor entende como lista de *stopwords* as palavras que dependem da natureza e objetivos dos textos, observando se elas estão com boa ou

[36] Mineração de textos é um conjunto de métodos usados para navegar, organizar, achar e descobrir informação em bases textuais. (ARANHA; PASSOS, 2006, p. 03)

[37] Permite o agrupamento de informações segundo as próprias informações.

má formação. Fazem partes da lista de *stopwords* somente abreviações, datas, números, caracteres especiais e qualquer tipo de palavra indesejada, sendo as demais categorias gramaticais termos que ajudam na formação do SN.

Na etiquetação das palavras, o autor usa o *tagger* MXPost[38]. Logo após, há substituição do *stemming* por algoritmo de *matching*. O objetivo do *stemming* é diminuir o número de palavras que entram na BOW (Bag-of-Words), para tanto, cada palavra é reduzida a parte raiz, eliminando-se os sufixos, formas verbais e plurais. Mesmo com o aumento de precisão, o *stemming* descaracteriza totalmente os termos que representam ideias e assuntos do texto. Para resolver esse problema, o autor usa o algoritmo *matching Levenshtein Distance* que mede a similaridade entre duas palavras (raízes).

Em seguida, há uma adequação do TFIDF (*Term Frequency Inverse Document Frequency*) no intuito de que a computação da frequência de um termo em documento seja viável a proposta feita pelo o autor. TFIDF está relacionado à extração e criação de índices para descritores no processo de busca pela informação em um SRI.

A autora parte da seguinte fórmula TFIDF:

$$tfidf(j) = tf(j) \times idf(j)$$

[38] Faz parte do projeto Lácio-Web. NÚCLEO INTERINSTITUCIONAL DE LINGUÍSTICA COMPUTACIONAL. Disponível em: <http://www.nilc.icmc.usp.br/nilc/tools/nilctaggers.html>. Acesso em: 13 abr. 2013

> Onde **tf** é o número de vezes que um termo/descriptor **j** aparece no documento e **idf** é:
>
> $$idf(j) = \log\left(\frac{N}{df(j)}\right)$$
>
> Resultante do logaritmo natural de **N** dividido por **df(j)**. onde **N** é o tamanho do corpus (número de documentos) e **df(j)** é a quantidade de documentos onde o termo/descritor **j** está presente. (PINHEIRO, 2009, p. 53)

Na etapa seguinte, o autor usa o Aplicativo Gerador de Descritores para padronizar as palavras, retirar as *stopwords*, etiquetar as palavras usando o MXPost Tiger, recuperar descritores de acordo com os padrões configurados e gerar BOW usando algumas fórmulas.

Por fim, Pinheiro (2009) usa o *Software Rapid Miner* para fazer o processo de Mineração de Opiniões, pois possui todos recursos necessários para o tratamento e criação de tarefas de mineração de textos, tendo, ainda, uma interface que permite que se estude e aperfeiçoe os resultados gerados pelos modelos.

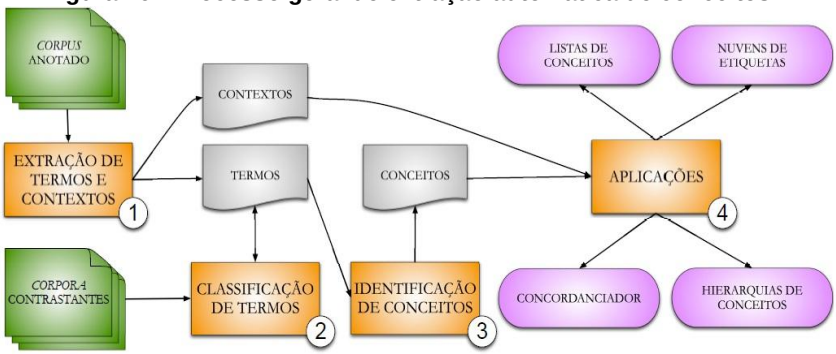
Com os bons resultados, o autor desenvolve um processo considerado inovador na etapa de pré-processamento de textos, com um melhor desempenho e interpretação do que a abordagem tradicional de mineração de textos. Além do que o autor comprova que a sua abordagem com o uso de SNs em Mineração de Opiniões

produz agrupamentos de melhor interpretação e desempenho em relação à abordagem tradicional.

4.14 Contribuições de Lopes (2011)

Lopes (2011), com o objetivo de extrair automaticamente conceitos para um domínio caracterizado por um corpus em língua portuguesa, no Programa de Pós-Graduação em Ciência da Computação da Pontifícia Universidade Católica do Rio Grande do Sul, define um método de extração de termos candidatos a conceitos a partir de um corpus marcado linguisticamente, ordena os termos extraídos segundo sua relevância, identifica, entre os termos extraídos, quais devem ser considerados conceitos do domínio e constrói um conjunto de recursos linguísticos a partir dos conceitos extraídos, facilitando a sua compreensão, manipulação e visualização. A Figura 16 representa as atividades da metodologia da autora.

Figura 16 - Processo geral de extração automática de conceitos



Fonte: Lopes (2011, p. 22)

No primeiro momento, acontece a extração de termos e contextos que está relacionada à detecção de termos candidatos a conceitos de uma área de domínio em um corpus marcado. A marcação linguística e a identificação SNs se dão pelo *parser* PALAVRAS. Logo em seguida, os SNs são extraídos e passam por um tratamento baseado em um conjunto de regras heurísticas para que sejam refinados no intuito de selecionar os candidatos a conceitos de um domínio.

As regras heurísticas propostas por Lopes (2011) são constituídas por três grupos: de ajuste - que tem como função adaptar os SNs extraídos; de descarte - que recusam SNs inadequados; e a de inclusão - que identificam SNs que não podem ser encontrados pela simples leitura sequencial dos textos do corpus pelo fato de estarem implícitos.

As regras da heurística de ajuste estão relacionadas à remoção dos artigos no começo de um SN, dos artigos em qualquer posição de um SN, dos pronomes no começo de um SN e dos pronomes em qualquer posição de um SN. A autora ressalta que as regras de remoção de palavras feita a partir de uma anotação linguística prévia tendem a ser mais precisas do que as abordagens estatísticas.

As regras da heurística de descarte rejeitam os SNs que contenham numerais ou outros símbolos além de letras como os hífens, que tenham como núcleo um pronome e que comecem com um advérbio. Diferentemente da heurística de ajuste que alteram o número de palavras dos SNs, a heurística de descarte diminui significativamente o número total de SNs extraídos.

Já nas regras de heurística de inclusão, são identificados SNs implícitos por meio de inferências permitidas pela marcação linguística. Estas regras são a remoção sucessiva de adjetivos, uso de predicado múltiplo[39] e conjunção de adjetivos[40]. Com as regras heurísticas de inclusão há um aumento no número total de SNs.

Dessa maneira, tem-se um conjunto de SNs extraídos com informações conceituais, o que possibilita seu uso em várias aplicações.

> 1. o termo na sua forma original;
>
> 2. o termo na sua forma canônica;
>
> 3. o número de palavras que compõem o termo [...];
>
> 4. a palavra indicada como núcleo do termo na sua forma canônica;
>
> 5. a etiqueta sintática do núcleo (substantivo, adjetivo, etc.);
>
> 6. a(s) etiqueta(s) semântica(s) do núcleo [...];
>
> 7. a etiqueta morfológica do núcleo [...];
>
> 8. a função gramatical do termo na oração [...];
>
> 9. a posição ocupada pelo termo na frase [...];
>
> 10. o número total de palavras da frase;
>
> 11. o predicado ao qual o termo exerce sua função gramatical na forma original;
>
> 12. o predicado ao qual o termo exerce sua função gramatical na forma canônica;

---

[39] Predicado com mais de um verbo.

[40] Quando um substantivo é caracterizado por dois ou mais adjetivos.

> 13. a etiqueta sintática do predicado ao qual o termo exerce a sua função gramatical;
>
> 14. a etiqueta morfológica do predicado ao qual o termo exerce a sua função gramatical;
>
> 15. a posição ocupada pelo predicado ao qual o termo exerce sua função gramatical na frase (onde situam-se as palavras que compõem o predicado);
>
> 16. um identificador da frase de onde o termo foi extraído;
>
> 17. um identificador do documento de onde o termo foi extraído.
>
> (LOPES, p.52)

Com esse produto da extração dos termos (SNs), tem-se a segunda etapa do processo geral de extração automática de conceitos que ordena os termos de acordo com sua relevância por meio de estatística que conta um conjunto de corpora usados como contraste de domínio, resultando na atribuição de um valor numérico a cada termo, o que permite a ordenação dos termos segundo sua relevância para o domínio. Na terceira etapa, há a escolha e aplicação de um ponto de corte à lista ordenada de conceitos.

Já na quarta etapa, são utilizados os conceitos e seus respectivos contextos com o objetivo de gerar recursos linguísticos sofisticados como: lista de conceitos com informações contextuais, nuvens de conceitos com informações estruturadas visualmente, hierarquia de conceitos que permitem detectar relações taxonômicas entre os conceitos, entre outras aplicações.

Pode-se dizer ainda que Lopes (2011), na metodologia de desenvolvimento de sua tese de doutorado, implementou todas as etapas do processo geral de extração automática de conceitos no EχATOLP (Extrator Automático de Termos para Ontologias em Língua Portuguesa) que é um programa computacional, ainda em desenvolvimento, que oferece diversos modos de saída de termos e conceitos na forma de listas, nuvens de conceitos, hierarquia de conceitos, entre outras.

Para os testes, a autora utilizou cinco corpora de domínios específicos: o de Pediatria constituída de 183 textos do Jornal de Pediatria; o de Modelagem Estocástica com 88 textos; Mineração de Dados com 53 textos; Processamento Paralelo com 62 textos; e Geologia com 234 textos. O corpus de Pediatria foi desenvolvido por Robert James Coulthard, em 2005, na sua tese "*The application of Corpus Methodology to Translation: the JPED parallel corpus and the Pediatrics comparable corpus*", defendida na Universidade Federal de Santa Catarina. Já os outros quatro corpora foram elaborados pela autora com o auxílio de especialista do Programa de Pós-Graduação em Ciência da Computação Departamento de Informática da Pontifícia Universidade Católica do Rio Grande do Sul.

Os resultados alcançados pela autora permitiram o desenvolvimento do processo de extração de conceitos a partir de corpora de domínios, além de contribuir com o desenvolvimento do EχATOLP com a junção dos cinco corpora que servem como um conjunto homogêneo, caracterizando-se como um recurso linguístico para o trabalho

com a PLN e a construção de uma lista de conceitos dos corpora de domínio que pode ser usada por outros pesquisadores. A avaliação dos resultados foi feita a partir de um processo comparativo entre corpora contrastantes e cálculo de índice de frequência dos termos.

4.15 Contribuições de Corrêa et al (2011)

Com o objetivo de analisar a extração automática de SNs e a utilização dos mesmos como descritores, Corrêa et al (2011) fazem um estudo de caso em que se analisa tais descritores como pontos de acesso a teses e dissertações de três programas de pós-graduação da Universidade Federal de Pernambuco (UFPE).

O estudo de caso se fez por meio da extração de SNs pela ferramenta OGMA. O corpus utilizado foi um conjunto de 30 resumos extraídos da Biblioteca Digital de Teses e Dissertações (BDTD) da UFPE. Os resumos são dos programas de Direito, Ciência da Computação e Nutrição, sendo 10 resumos de cada.

Os resumos submetidos ao OGMA eram constituídos pelos seguintes campos de metadados: titulação, resumo, assunto (contendo as palavras-chave), nome do programa e grande área do programa. Foi incluído nos resumos o caractere ponto como separador de campos, assim como múltiplos valores de um campo como no caso das palavras-chave.

Na extração de SNs por meio do OGMA, foram seguidas três fases: abertura do texto de cada resumo no *software*, a etiquetagem dos termos do resumo aberto, e extração de SNs pontuados do resumo etiquetado.

Para a avaliação desse processo, aplicou-se cálculo e análise dos percentuais de precisão em extrair SNs relevantes como descritores, também foram observados a taxa de erro ao extrair cadeias de caracteres que não configuram SNs e o percentual de SNs extraídos não relevantes como descritores, bem como precisão e abrangência na extração de bons descritores.

Para a realização do cálculo dos percentuais, há a classificação dos possíveis SNs extraídos como verdadeiros ou falsos. Esta classificação se dá baseada nas definições de SNs encontradas na revisão de literatura dos autores. Há também a classificação dos SNs extraídos em relevantes ou não como descritores do resumo de onde foram extraídos, analisando o assunto do resumo e as palavras-chave da respectiva dissertação ou tese.

Os resultados alcançados por Corrêa et al (2011) permitem afirmar que nem todo SN extraído pelo OGMA é um SN de fato, essa ferramenta não extrai todos os tipos de SNs como o caso de nomes próprios, nem sempre a lista de SNs extraídos ordenada em forma decrescente de pontuação aponta no topo os SNs mais relevantes. Contudo, na maioria das vezes, os SN extraídos e classificados como relevantes constituem bons descritores pra os resumos, permitindo tê-los como bons pontos de acesso a teses e dissertações no processo de RI.

Os autores concluem que os processos de extração de SNs e ordenação desses por relevância devem sofrer uma investigação mais profunda, no intuito de haver melhoras na extração desses SNs relevantes como descritores documentais.

4.16 Discussão Acerca do Estado da Arte

As pesquisas sobre a extração automática de SNs em língua portuguesa ainda não atingiram um grau de amadurecimento elevado, mesmo com os desenvolvimentos de algumas ferramentas como o *parser* PALAVRAS, LX-Parser e OGMA, pois elas não alcançam uma taxa de precisão elevada para todos os tipos de corpus.

Isso pode ser explicado pelo fato da língua portuguesa ser muito complexa gramaticalmente, diferente da língua inglesa que tem sua estrutura gramatical mais simplificada. Por exemplo, o adjetivo no inglês sempre antecederá o substantivo, já na língua portuguesa ele pode anteceder ou vir após o nome. Quando se depara com a conjugação verbal, os problemas são mais aparentes, pois na língua de Camões para cada pessoa do discurso há uma desinência verbal, tomando-se como exemplo o tempo presente do modo indicativo, têm-se seis desinências, enquanto no tempo equivalente no inglês se têm duas desinências nos verbos regulares. Essas diferenças de estruturas gramaticais acabam por requerer outras metodologias para extração de SNs.

Mas não se pode usar esse argumento da língua para desestimular os estudiosos, deve-se conhecer as regras gramaticais que regem a estrutura de uma sentença em português, percebendo as maneiras que elas podem ser inseridas em um programa de computador, por exemplo. A extração automática de informação de um documento por meio de SNs deve ser estimulada para que a sua aplicabilidade se dê efetivamente nos SRIs. Para tanto, é necessário um relacionamento interdisciplinar das ciências como a Linguística, Ciência da Computação, CI e Terminologia, e até mesmo as Ciências Cognitivas. É a junção dos conhecimentos dessas disciplinas que permite melhores ferramentas automáticas para extrair dos textos termos para descrever documentos, auxiliando no processo de tratamento da informação, no intuito de que, posteriormente, o usuário de um determinado sistema faça a recuperação desses documentos.

Morellato (2010) argumenta que os trabalhos produzidos em língua portuguesa relacionados à identificação de sintagmas, na maioria das vezes, são desenvolvidos a partir de aprendizado de máquina, o que exige uma base de dados marcada, gerando um sistema dependente do domínio tratado. Se os trabalhos são desenvolvidos com a implementação de regras no sistema, há uma limitação de processar conjuntos de dados pequenos, ou dependentes de *software*, não sendo indicado o uso em múltiplos domínios.

Os estudos sobre a extração automática de SNs em língua portuguesa começaram na final da década de 1980 e início da década de 1990. Brito (1992) sugere a indexação por

SNs de maneira não muito detalhista, baseando-se nos estudos de Michel Le Guern que aponta para a indexação automática por meio das análises morfossintáticas (BRITO, 1992; Kuramoto, 1995; 1999). Já Kuramoto (1995; 1999; 2002) sugere a navegação em um SRI por meio dos níveis do SN, o que permitira que o usuário atingisse seus objetivos na busca, já que cada nível do sintagma traz uma seleção de informações. Contudo, como argumenta Souza (2005), não se encontra na literatura científica trabalhos que dessem continuação ao trabalho de Kuramoto.

Bick (2000) desenvolveu o seu trabalho à mesma época de Kuramoto (1995; 1999). Com o parser PALAVRAS permitiu que análises morfossintáticas fossem feitas com maior precisão. Essa ferramenta foi utilizada por vários autores como Miorelli (2001), Santos (2005), Souza (2005), Morellato (2007; 2007), entre outros, ou para fazer parte de umas das etapas das metodologias ou para fazer as validações das metodologias propostas por esses autores.

Morellato (2010) faz uma boa observação quando percebe que os trabalhos de Kuramoto (2005), Miorelli (2001), Souza (2005), Santos (2005) e Costa (2007) almejavam a obtenção de altos valores de precisão e de abrangência para que pudessem ser aplicados nos desenvolvimentos de sistemas em diversas áreas.

Quando os pesquisadores fazem o estudo do estado da arte, permite-se que ferramentas ou metodologias sejam desenvolvidas ou aperfeiçoadas a partir de outras, é o caso do OGMA (MAIA, 2008) que foi confeccionado baseado nas metodologias de Miorelli (2001) e Souza (2005).

Com a característica de se aperfeiçoar as ferramentas a partir da colaboração de vários estudos, alguns grupos como NILC (desde 1993) e o NLX-Group (desde 2002) passam a existir, com o objetivo de fazer parcerias entre disciplinas, instituições e estudiosos. Os corpora produzidos pelo NILC são submetidos por diversos pesquisadores às metodologias e ferramentas de extração de informação. O NLX-Group também disponibiliza recursos, serviços e ferramentas linguísticas desenvolvidas totalmente ou em parte por esse grupo. Esses serviços são oferecidos gratuitamente na *Web*.

Para a seleção de SNs, alguns autores como Santos (2005), Souza (2005), Maia (2008), Corrêa et al (2011) fazem uso de abordagens estatísticas para extrair termos como maior relevância. Lopes (2011) salienta que a seleção desses termos feita por regras de remoção de palavras é feita a partir de uma marcação linguística previamente estabelecida, sendo mais precisa do que as abordagens estatísticas.

Alguns autores como Souza (2005), Maia (2008) e Pinheiro (2009) propõem em suas metodologias a criação de uma lista de *stopwords,* de palavras ou de SNs indesejados. Contudo, percebe que se devem levar em consideração as características de cada texto, do qual tenham sido extraídas as palavras que compõe essa lista, pois com a exceção de abreviações, datas, números, caracteres especiais e qualquer tipo de palavra indesejada (PINHEIRO, 2009), podem-se excluir termos que ajudam na formação do SN.

Dos trabalhos desenvolvidos por diferentes autores, os que seguem uma linearidade são Kuramoto (1995) com a

proposição de oferecer regras que possibilitem a identificação de sintagmas nominais, Souza (2005) que cria taxas de relevância para os sintagmas nominais, Maia (2008) que cria uma ferramenta que implementa essas regras e metodologias de pontuação de relevância em um *software* e Corrêa et al (2011) que testa esse *software* em um outro ambiente diferente do idealizador da ferramenta.

Os resultados dos autores que trabalham com a extração e seleção automática de sintagmas nominais devem ser comparados entre si para que se possam detectar os problemas existentes quanto às metodologias e às ferramentas de extração dos SNs em língua portuguesa. No intuito de que ferramentas e metodologias sejam melhoradas e que efetivamente possam ser aplicadas em sistemas de recuperação de informação, satisfazendo as necessidades informacionais do usuário. Dessa forma, a Tabela 9 e a Tabela 10 trazem informações sobre os trabalhos de autores como Kuramoto (1995 – 2002), Bick (2000), Vieira et al (2000), Miorelli (2001), Souza (2005), Santos (2005), Arcoverde (2007), Costa (2007), Morellato (2007; 2010), Maia (2008), Pinheiro (2009), Lopes (2011) e Corrêa et al (2011), objetivando fazer uma síntese dos principais trabalhos da literatura sobre a extração e seleção automática de SNs. Não estão listados nas tabelas seguintes os trabalhos de Othero (2004) e David (2007) porque eles não fazem a extração de SNs diretamente, trabalham com a identificação dos mesmos por meio das regras da Teoria X-Barra.

Tabela 9 – Comparação entre alguns trabalhos que compõe a literatura de extração de SNs

Trabalho	Metodologia de extração de SN	Metodologia de seleção de SN	Metodologia de avaliação		
			Corpus	Métricas	Referência
Kuramoto (1995-2002)	Manual	-	15 artigos científicos da área de CI	Equivalência	Extração manual
Bick (2000)	Automática	Automática	"O Tesouro" (Eça de Queiroz) e Artigos da Revista Veja	Precisão Acurácia Abrangência	Análise da saída do *software* pelos autores
Vieira et al (2000)	Manual e automática	Classificação manual	15 artigos do Jornal Correio do Povo	Abrangência	Identificação ou extração de SN por outro *software* (PALAVRAS)
Miorelli (2001)	Automática	-	Resumos de dissertações do Programa de Pós-Graduação em Ciência da Computação da PUC-RS	Precisão	Palavras-chave
Souza (2005)	Automática	Atribuição de pesos aos SNs de acordo com a frequência e estrutura do SN	Artigos usados por Kuramoto (1995-2002) e Artigos dos periódicos DataGramaZero e Ciência da Informação	Precisão Abrangência	Palavras-chave
Santos (2005)	Automática	-	Mac-Morpho do NILC – 10 cadernos do Jornal Folha de São Paulo	Abrangência e Acurácia	Análise da saída do *software* pelos autores
Arcoverde (2007)	Automática	Atribuição de pesos aos SNs de acordo com a frequência	CLEF 2006	Equivalência Precisão Revocação MAP (*mean average precision)*	Identificação ou extração de SN por outro *software* (CLEF 2006)
Costa (2007)	Automática	-	186 itens – frases da própria dissertação	Equivalência Precisão	Análise da saída do *software* pelos autores
Morellato (2007)	Automática	-	Notícias jornalísticas, artigos científicos, *blogs*, teses de doutorado e livros literários encontrados pelos *sites* de busca.	Precisão	Análise da saída do *software* pelos autores
Maia (2008)	Automática	Atribuição de pesos aos SNs de acordo com a frequência e estrutura do SN (SOUZA, 2005)	Trabalhos do ENANCIB (2005) e de quatro temas: informática, turismo, veículos e mundo	Precisão Medidas de similaridade Taxa de erro Taxa de acerto	Identificação ou extração de SN por outro *software* (PALAVRAS) e Classificação de documentos (WEKA)

Trabalho	Metodologia de extração de SN	Metodologia de seleção de SN	Metodologia de avaliação		
			Corpus	Métricas	Referência
Pinheiro (2009)	Automática	Quantificação de frequência de um termo em um texto	2.000 e-mails de clientes corporativos da empresa Light S.A.	Equivalência	Análise da saída do *software* pelos autores
Morellato (2010)	Automática	-	36 expressões de corpus do NILC e as 186 expressões usadas por Costa (2007)	Precisão Abrangência	Identificação ou extração de SN por outro *software* (FORMA, PALAVRAS e LXGram)
Lopes (2011)	Automática	Aplicações de regras heurísticas e índice de relevância e frequência	Cinco corpora de domínios específicos: Pediatria; Modelagem Estocástica; Mineração de Dados; Processamento Paralelo; e Geologia.	Equivalência Precisão Abrangência	Análise da saída do *software* pelos autores e lista de termos que deveriam ser extraídos
Corrêa et al (2011)	Automática	Análise de relevância dos SNs	30 resumos da BDTD-UFPE	Precisão Abrangência Taxa de erro Taxa de acerto	Análise da saída do *software* pelos autores

Fonte: o autor

Tabela 10 - Comparação entre alguns trabalhos que compõe a literatura de extração de SNs quanto aos sistemas propostos e usados

Trabalho	Sistema proposto	Sistemas usados		
		Etiquetador	*Parser*	Extrator de SN
Kuramoto (1995-2002)	Protótipo	-	-	Manual
Bick (2000)	PALAVRAS	PALAVRAS	PALAVRAS	-
Vieira et al (2000)	Sistema baseado em Prolog	PALAVRAS	PALAVRAS	-
Miorelli (2001)	ED-CER	Manual	-	ED-CER
Souza (2005)	Metodologia de seleção automática de descritores	PALAVRAS	PALAVRAS	Xtractor
Santos (2005)	Sistema baseado em TBL	-	-	-
Arcoverde (2007)	Indução automática de filtro	MxPost	-	Sistema baseado em TBL (SANTOS, 2005)
Costa (2007)	LX-GRAM	LX-GRAM	LX-GRAM	-
Morellato (2007)	SIDSN	-	-	-
Maia (2008)	OGMA	OGMA	-	ED-CER
Pinheiro (2009)	Processo de Mineração de Opiniões	MxPost	-	Software Rapid Miner
Morellato (2010)	SISNOP	PALAVRAS e FORMA	PALAVRAS e BI-SON	SISNOP
Lopes (2011)	ExATOlp	PALAVRAS	PALAVRAS	ExATOlp
Corrêa et al (2011)	-	-	-	OGMA

Fonte: o autor

De acordo com a Tabela 9, os trabalhos usaram diferentes métodos para alcançar, um bom grau de precisão na extração de SNs. Alguns trabalhos não tinham como um dos objetivos a seleção de SNs, mas todos, de alguma maneira, almejavam realizar a extração automática.

O preenchimento da Tabela 9 visa à descrição dos trabalhos de acordo com as categorias nas colunas. Assim, quando se fala em metodologia de extração, ou seja, a

maneira como os autores fizeram a extração dos SNs, aponta-se se a extração foi feita de forma automática, manual ou híbrida. A mesma técnica foi tomada como referência para a coluna “metodologia de seleção de SN”.

Ao falar de métricas, classificam-se as fórmulas usadas pelos pesquisadores para obter os índices de validação das metodologias e/ou ferramentas, tendo como categorias acurácia, precisão, abrangência e equivalência.

A coluna referência está direcionada a indicar os resultados apontados por outras ferramentas e/ou metodologias que servem para comparar com os resultados dos sistemas propostos, validando esses. Nessa coluna, trabalha-se com as seguintes categorias: corpora etiquetados; extração manual de SNs; Palavras-chave de textos; identificação ou extração de SN por outro *software*; análise da saída do *software* pelos autores.

Diante da Tabela 9, percebe-se que as tendências quanto às métricas utilizadas são referentes ao uso das taxas de precisão, abrangência e equivalência. Quanto à metodologia de seleção, vê-se que Souza (2005) e Maia (2008) trabalham com a atribuição de pesos aos SNs de acordo com a frequência e a estrutura do SN, na qual Corrêa et al (2011) se baseiam para aplicar cálculo e análise dos percentuais de relevância. Já Lopes (2011) aplica regras heurísticas e índice de relevância e frequência, enquanto Pinheiro (2009) quantifica a frequência de um termo em um texto.

Em relação à referência para avaliação dos SNs extraídos, 7 trabalhos partem para a análise de saída do

software pelos autores, seguida de outras duas tendências que são a identificação ou extração de SN por outro *software* e a equivalência entre um grupo de palavras (conceitos, palavras-chave, corpus já etiquetados) com as obtidas no experimento do próprio autor.

Já na Tabela 10, são verificados os sistemas propostos pelos autores do trabalho, assim como os *taggers* (etiquetadores) e os *parsers* utilizados nos sistema propostos. Também é observado se foi usado algum extrator de SN.

Quanto aos programas mais utilizados, é evidente que o PALAVRAS (BICK, 2000) é o mais referenciado. Contudo, a metodologia de seleção automática de descritores (SOUZA, 2005), o ED-CER (MIORELLI, 2001), o OGMA (MAIA, 2008) e o TBL (SANTOS, 2005) são usados por outros pesquisadores, como foram vistos na Tabela 10.

Dessa forma, percebe-se a necessidade de se comparar os programas a partir de suas funcionalidades práticas, o que será exposto no próximo capítulo.

5 IDENTIFICADORES DE SINTAGMAS NOMINAIS: uma análise comparativa entre o OGMA, Parser PALAVRAS, LX-Parser e a extração manual

Esse capítulo tem como intuito fazer uma comparação entre os *softwares* que identificam SNs, apontando as possíveis falhas que cada programa apresenta. Para tanto, toma-se a extração manual como referência, já que essa, provavelmente, está em um nível de acerto maior, pois a leitura da linguagem natural feita pelo homem possibilita a interpretação e compreensão dos recursos linguísticos utilizados pelas autorias do texto. Há particularidades que são difíceis para os programas computacionais compreenderem, principalmente na Língua Portuguesa: o autor de um determinado texto verbal pode dizer a mesma coisa de maneiras diferentes, mudando as posições lexicais, trocando a pontuação, usando figuras de linguagem, figuras de pensamento, entre outros recursos. O olhar humano é mais perceptível a essas particularidades, contudo os projetos de PLN avançam visando alcançar esse nível através do computador.

O processo de construção da Tabela 11 foi feito a partir da extração dos SNs pelo *software* OGMA, identificação dos SNs pelos *softwares* Parser PALAVRAS e LX-Parser, bem como da extração manual. Da saída dos *softwares* PALAVRAS e LX-Parser foram extraídos os SNs marcados, ou seja, esses programas identificavam os SNs por meio de representações arbóreas em que cada palavra era etiquetada de acordo com seu valor semântico e logo em seguida era

feita a transcrição dos SNs (manualmente) em um editor de texto e editor de planilhas. O OGMA já faz o processo de extração de SNs, permitindo, assim, rapidez nas transferências dos sintagmas para um editor de texto e de planilha. O processo de extração manual foi feito a partir da leitura especializada do autor dessa dissertação.

A Tabela 11 apresenta o número total de todas as ocorrências de SNs extraídos pelos referidos programas e pela extração manual. O que identificou mais SNs foi LX-Parser, contudo os SNs extraídos por essa ferramenta são muito vulneráveis a erros que serão expostos logo mais. O que obteve o menor número na extração foi o OGMA, mas quando se observa os “verdadeiros” SNs esse número cai ainda mais. O PALAVRAS foi o que mais se aproximou da extração manual, porém apresentou alguns erros nas estruturas dos SNs identificados.

Tabela 11 - Total de SNs identificados nos resumos por cada programa e pela extração manual

Resumos	OGMA	PALAVRAS	LX-Parser	Extração Manual
CC 1	55	88	112	90
CC 2	66	145	186	163
CC 3	56	102	136	120
CC 4	44	98	126	100
CC 5	97	165	202	181
CC 6	37	47	55	54
CC 7	24	53	87	56
CC 8	37	64	92	76
CC 9	57	123	152	131
CC 10	34	73	83	65
Total CC	**507**	**958**	**1.231**	**1.036**
DI 1	29	38	60	48
DI 2	25	39	52	39
DI 3	41	67	99	75
DI 4	41	63	98	67
DI 5	31	52	64	51
DI 6	27	64	80	69
DI 7	35	75	105	82
DI 8	24	61	80	68
DI 9	14	34	49	43
DI 10	43	65	82	67
Total DI	**310**	**558**	**769**	**609**
NU 1	39	74	165	74
NU 2	47	77	121	95
NU 3	44	87	129	87
NU 4	42	68	115	77
NU 5	48	100	156	116
NU 6	45	73	93	77
NU 7	72	106	145	98
NU 8	54	94	164	109
NU 9	35	74	116	84
NU 10	55	89	136	107
Total NU	**481**	**842**	**1.340**	**924**
TOTAL	**1.298**	**2.358**	**3.340**	**2.569**
TOTAL GERAL	**9565**			
CC – DI – NU –	Ciência da Computação Direito Nutrição			

Fonte: o autor

Na seleção das expressões que, de fato, constituem SNs, foram descartados os que começavam com preposição ou conjunção, que continham verbos em suas estruturas e numerais que não exerciam função de determinantes dos nomes. Desse modo, apresentam-se os resultados para cada identificador de SN.

Das 1.298 expressões extraídas pelo OGMA, 357 não constituem SN, tendo um total de 941 de expressões que constituem SNs, de fato. Assim, a taxa de acerto[41] desse *software* foi de 72,5% em relação aos seus próprios resultados, porém se for feita uma comparação com a extração manual (2.569 SNs extraídos), a taxa de revocação[42] é de 36,5%.

Os problemas apresentados pelo referido programa estão relacionados à marcação de adjetivos e verbos como SN, por exemplo, a expressão "*bilateral de negociação de serviços do controlador*" que começa com um sintagma adjetival, mas é extraído pelo OGMA como um SN. Alguns dos SNs extraídos começavam com preposição ou conjunção. Outro problema identificado é quanto ao quebramento do SN, deixando-o incompleto.

O Parser PALAVRAS, considerado pela literatura como o melhor identificador de SNs, identificou 2.358

[41] Essa taxa de acerto é a razão do total de expressões que de fato constituem SN sobre o total de expressões extraídas pelos programas.

[42] A taxa de revocação é a razão do número de expressões que de fato são SNs extraídos pelos *softwares* sobre o número total de SNs extraído pela técnica manual.

expressões, sendo 2.220 como expressões que constituem SNs, de fato, representando, assim, um total de 94% de acerto em relação aos seus próprios resultados, o que reafirma a taxa obtida pelo seu idealizador, Bick (2000), que foi de 98% de acerto na identificação de SNs (lembrando que o método de análise é diferente do usado pelo referido autor), contudo se for levada em consideração a extração manual, a taxa de revocação é de 86,5%.

Os problemas apresentados pelo PALAVRAS estão relacionados a várias circunstâncias como algumas palavras que deixaram de ser identificadas como SN, sendo apenas marcadas como nomes, outros vocábulos que pertenciam a uma classe gramatical eram marcados como outra classe diferente (por exemplo, adjetivos e verbos marcados como nomes e, por conseguinte como SN), algumas palavras foram excluídas de alguns SNs, numerais foram identificados isoladamente como SNs, e o artigo “os” foi confundido algumas vezes com o pronome oblíquo, sendo marcado, dessa maneira, como um SN.

O LX-Parser, como já foi mencionado anteriormente, faz a leitura de frases simples e sua análise sempre acaba no primeiro ponto final, tendo ainda dificuldades de etiquetagem quando há a falta de verbo. Sendo assim, a submissão dos resumos foi feita período por período (cada sentença até o ponto final). Um dos pontos positivos desse *software* e do PALAVRAS, é que o usuário visualiza os SNs em seus diversos níveis, ou seja, o LX-Parser consegue identificar diferentes sentenças e analisá-las em uma representação arbórea.

Esse programa identificou 3.340 expressões nominais, dentre os *softwares* utilizados foi o que mais identificou SNs, sendo que desse total, 2.509 constituíam SNs de fato, atingindo 75% de acerto na identificação de SN de acordo com os próprios resultados, mas em comparação com o resultado da extração manual, a taxa de revocação é de 97,5%.

De modo geral, o LX-Parser enfrenta alguns problemas de etiquetagem. Algumas palavras de outras classes gramaticais foram marcadas como verbos, por exemplo, adjetivos, advérbios, numerais, siglas e até letras isoladas. Em algumas situações, palavras como adjetivos, verbos, numerais, artigos e até símbolos e letras isoladas foram identificados como SNs.

Também havia problemas quando LX-Parser identificava o SN começando por sinal de pontuação (vírgula “,”, dois pontos “:”), conjunção e preposição. Mas um ponto positivo é que esse programa raramente partia o SN ao meio.

A Tabela 12 aponta a quantidade de expressões identificadas pelos três *softwares*, mas que não constituem SNs.

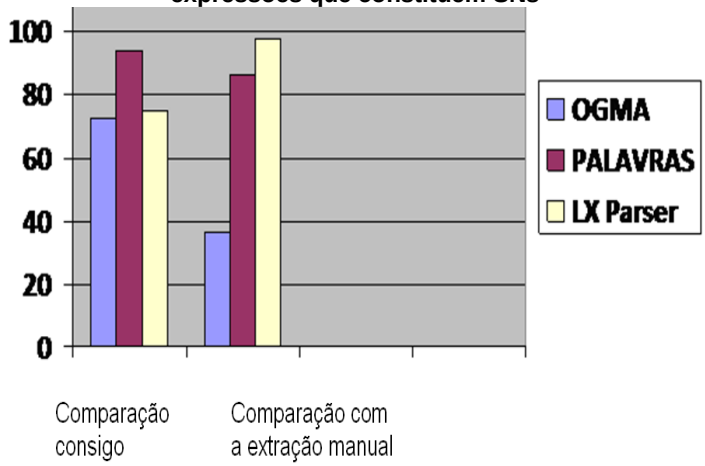
Tabela 12 – Quantidade de expressões identificadas, mas que não constituem sintagmas

Resumos	OGMA	PALAVRAS	LX-Parser
CC 1	14	3	26
CC 2	19	10	32
CC 3	13	5	28
CC 4	15	3	31
CC 5	21	7	35
CC 6	12	1	9
CC 7	5	0	24
CC 8	5	3	32
CC 9	12	4	22
CC 10	5	6	21
Total CC	121	42	260
DI 1	11	2	18
DI 2	5	2	17
DI 3	11	6	31
DI 4	17	7	29
DI 5	10	6	23
DI 6	6	3	14
DI 7	11	2	34
DI 8	8	3	15
DI 9	5	0	4
DI 10	17	2	18
Total DI	101	33	203
NU 1	9	19	83
NU 2	12	4	28
NU 3	8	4	31
NU 4	13	4	22
NU 5	11	7	36
NU 6	17	1	14
NU 7	22	9	44
NU 8	19	1	34
NU 9	9	3	32
NU 10	15	11	44
Total NU	135	63	368
TOTAL	357	138	831
TOTAL GERAL	1.326		
CC – DI – NU –	Ciência da Computação Direito Nutrição		

Fonte: o autor

O Gráfico 1 apresenta uma comparação entre os softwares em relação à porcentagem de acerto, em que o primeiro conjunto de colunas representa a taxa de acerto em comparação com os próprios resultados (*comparação consigo*) e o segundo conjunto de colunas representa a taxa de acerto quando comparada com os resultados da extração manual.

Gráfico 1 – Níveis de acertos e revocação na identificação de expressões que constituem SNs



Fonte: o autor

Os percentuais de acerto atingidos pelo OGMA e o LX-Parser estão muito aproximados, tendo suas próprias saídas como referências. A vantagem do LX-Parser é que ele consegue identificar um número bem maior de SNs. Mesmo tendo uma taxa de erro aproximado do OGMA, a diferença na quantidade de SNs é destoante, pois o LX consegue identificar quase 2,5 vezes mais SNs que aquele. Já a vantagem do OGMA é que não precisa fazer a submissão

sentença por sentença, de maneira que o usuário faz a submissão com o texto completo. Além de identificar os SNs, o OGMA os extrai, o que facilitaria o trabalho do indexador.

Enquanto OGMA já pontua os SNs, sinalizando quais são, hipoteticamente, mais relevantes, o LX consegue apresentar os SNs por níveis, percebidos pelo usuário por meio da representação arbórea. Essa estrutura por níveis também é encontrada no Parser PALAVRAS, isso remonta a Kuramoto (1995) que formulou um protótipo em que o usuário navegaria pelos níveis do SN até satisfazer a sua necessidade informacional.

O PALAVRAS apresentou uma taxa de erro de 6% em comparação com seus próprios resultados, o que demonstra que a leitura do software já pode ser usada no processo de indexação, pois, além dessa dissertação, outros trabalhos já apontavam para a eficiência dessa ferramenta (MAIA, 2008; LOPES, 2009; MORELLATO, 2010), contudo ela não pontua os SNs que poderiam ser usados como descritores de um documento. Em alguns momentos o LX-Parser se mostrou mais eficiente do que o PALAVRAS conforme alguns resultados que serão apresentados a seguir e a quantidade de SNs de fato identificados conforme já exposto.

Outros aspectos observados nessa análise foram a quantidade de vezes em que os programas e as extrações manuais recuperaram os SNs que compunham o corpo das palavras-chave utilizadas pelas autorias dos resumos e a quantidade de SNs que continham essas palavras-chave.

A extração de SNs que fossem semelhantes às palavras-chave dos resumos trás uma maior confiabilidade

aos programas identificadores e/ou extratores dessas expressões nominais. Assim, na Tabela 13, há a quantidade de vezes que cada programa e a extração manual fizeram a extração de SNs que se assemelhavam às palavras-chave dos resumos, tendo em vista que algumas palavras-chave estavam no singular, enquanto nos resumos se encontravam flexionadas no plural. Desse modo, esse grupo é composto pelas palavras iguais às palavras-chave encontradas nos resumos e pelas suas flexões no plural.

Tabela 13 – Quantidade de vezes que cada programa e a extração manual recuperam SNs semelhantes às palavras-chave dos resumos

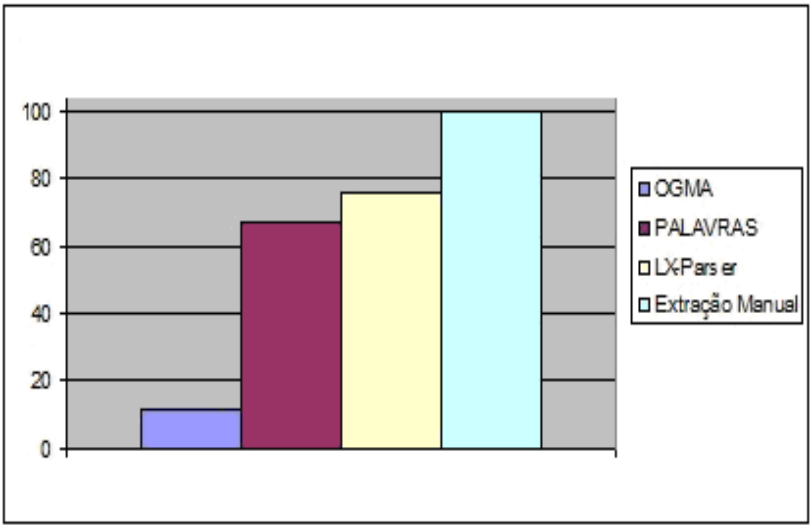
Resumos	OGMA	PALAVRAS	LX-Parser	Extração Manual
CC 1	2	2	2	2
CC 2	-	5	5	5
CC 3	-	3	3	7
CC 4	-	1	1	1
CC 5	1	4	4	12
CC 6	-	3	3	4
CC 7	1	7	8	8
CC 8	1	-	-	1
CC 9	1	8	7	8
CC 10	-	7	5	7
Total CC	6	40	38	55
DI 1	1	5	3	6
DI 2	-	2	1	3
DI 3	-	-	1	-
DI 4	1	5	1	5
DI 5	1	4	3	4
DI 6	-	2	2	2
DI 7	-	-	-	-
DI 8	1	3	3	4
DI 9	1	4	6	6
DI 10	-	2	3	3
Total DI	5	27	23	33
NU 1	1	2	4	5
NU 2	1	7	6	8
NU 3	-	3	1	2
NU 4	-	5	5	5
NU 5	-	5	8	8
NU 6	-	4	7	8
NU 7	2	3	3	5
NU 8	3	8	12	12
NU 9	-	-	14	15
NU 10	1	8	6	11
Total NU	8	45	66	79
TOTAL	19	112	127	167
TOTAL GERAL	425			
CC – DI – NU –	Ciência da Computação Direito Nutrição			

Fonte: o autor

A comparação entre as ferramentas, diante desse aspecto, é cabida, já que aqui não se leva em consideração o total de SNs extraídos em relação aos próprios resultados dos *softwares*, podendo, então, considerar os resultados da extração manual.

Dessa maneira, tendo a extração manual com o total de 100%, apresenta-se o Gráfico 2, representando a taxa de revocação na recuperação de SNs semelhantes às palavras-chave dos resumos.

Gráfico 2 – Taxa de revocação na recuperação de SNs semelhantes às palavras-chave dos resumos



Fonte: o autor

A ferramenta que mais se aproximou da extração manual (167 SNs semelhantes às palavras-chave) foi o LX-Parser que conseguiu atingir a taxa de 76% de revocação, seguido do Parser PALAVRAS com 67% e, por fim, o OGMA com 11,5%.

Os SNs que continham as palavras-chave dentro de suas partes constituintes foram também observados, tendo assim a Tabela 14.

Tabela 14 – Quantidade de SNs que continham palavras-chave dos resumos

Resumos	OGMA	PALAVRAS	LX-Parser	Extração Manual
CC 1	7	17	24	24
CC 2	3	12	12	15
CC 3	9	16	13	16
CC 4	10	8	15	15
CC 5	8	17	14	22
CC 6	-	6	5	7
CC 7	6	15	11	15
CC 8	4	5	-	5
CC 9	3	6	6	12
CC 10	3	5	11	11
Total CC	53	107	111	142
DI 1	1	6	2	6
DI 2	2	-	-	-
DI 3	1	2	2	2
DI 4	2	6	6	6
DI 5	1	1	2	3
DI 6	9	11	21	23
DI 7	2	10	1	10
DI 8	5	9	11	12
DI 9	1	14	12	14
DI 10	5	2	3	3
Total DI	29	61	60	79
NU 1	2	5	11	11
NU 2	1	10	18	18
NU 3	8	18	14	18
NU 4	6	7	4	7
NU 5	2	6	4	7
NU 6	5	15	17	17
NU 7	6	11	6	14
NU 8	7	18	24	24
NU 9	9	25	24	26
NU 10	3	13	19	24
Total NU	49	128	141	166
TOTAL	131	296	312	387
TOTAL GERAL	1.126			
CC – DI – NU –	Ciência da Computação Direito Nutrição			

Fonte: o autor

Somando-se a quantidade de SNs semelhantes às palavras-chave dos resumos com a quantidade de SNs constituídos por palavras-chave dos resumos, tem-se a quantidade de SNs relevantes para a descrição dos resumos. Desse modo, a Tabela 15 apresenta a quantidade de SNs relevantes para a descrição dos resumos.

Tabela 15 – Quantidade SNs relevantes para a descrição dos resumos

Tipos de SNs extraídos	OGMA	PALAVRAS	LX-Parser	Extração Manual
Semelhantes às palavras-chave dos resumos	19	112	127	167
Contém palavras-chave dos resumos	131	296	312	387
TOTAL	**150**	**408**	**439**	**554**

Fonte: o autor

A quantidade de SNs relevantes identificados pela extração manual atinge uma taxa de precisão de 21,5% do total de SNs extraídos. Na comparação entre as ferramentas automáticas, tomam-se os 554 SNs relevantes como referência para calcular a revocação dos SNs relevantes, isto é, a razão de extração dos SNs relevantes por cada programa sobre o total de SNs relevantes extraídos pela técnica manual.

O OGMA tem uma taxa de precisão de 16% de SNs relevantes dos 941 SNs de fato extraídos, mas em comparação com a quantidade de SNs relevantes extraídos

pela extração manual, aquele programa tem 27% na taxa de revocação. O Parser PALAVRAS atinge uma taxa de precisão de 18,5%, se forem tomados como referência os seus 2.220 SNs de fato identificados, mas tomando como referência a extração manual, o PALAVRAS alcança 73,5% de revocação, uma situação confortável e muito próximo do LX-Parser que, por sua vez, consegue atingir a marca de 79% de revocação quando se remete a quantidade SNs relevantes apresentados pela extração manual, mas em relação aos seus resultados, alcança a taxa de precisão de 17,5%.

Diante desses resultados, cabe agora apontar a ferramenta que teve melhor desempenho levando em consideração todos os aspectos apresentados nessa análise.

No primeiro momento, em que se observou o total de expressões que, de fato, constituíam SNs, o LX conseguiu 1,13 vezes a mais do que o PALAVRAS, assim como 6 vezes mais na identificação de expressões que não constituíam SNs. Esse último aspecto foi um ponto muito negativo na identificação de SNs feita pelo o LX-Parser, assim como atingir uma taxa de 25% de erro na identificação de SNs, enquanto o PALAVRAS teve apenas 6% de erro. Já o OGMA atingiu a taxa de 27% de erro na identificação de SNs, um resultado próximo ao do LX-Parser.

De acordo com os resultados apresentados, o LX-Parser teve um desempenho melhor, todavia algumas questões devem ser retomadas como a submissão dos resumos que nessa ferramenta foi feita sentença por sentença, o que de certa forma facilita o trabalho do

programa, mesmo assim, ainda identificou um número de falsos SNs muito alto em relação ao OGMA e ao PALAVRAS.

O PALAVRAS obteve um bom resultado na identificação de SNs, pois o objetivo das ferramentas automáticas identificadoras de SNs é apontar um maior número de SNs de fato e as expressões nominais apontadas pelo PALAVRAS, como já dito, tiveram um taxa de 6% de erro na identificação de SNs.

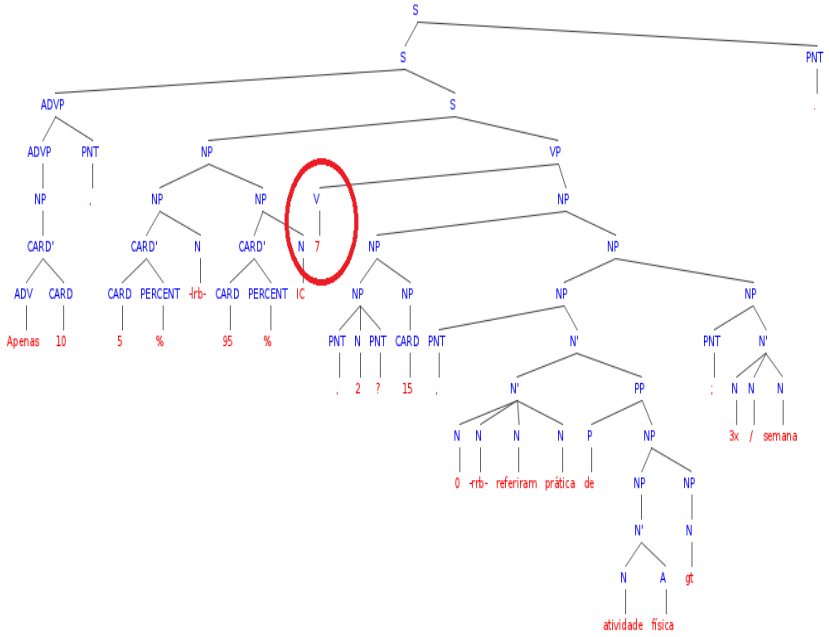
Considera-se, assim, reafirmando o que a literatura diz, o PALAVRAS como a ferramenta que diante de todo contexto tem maior potencial para identificação de SNs, facilitando o trabalho do indexador, diferentemente do LX-Parser, o usuário pode submeter todo o texto de uma vez só, o que demanda maior trabalho do *software.*

Frente aos problemas apresentados pelas ferramentas automáticas, algumas sugestões podem ser dadas no intuito que os programas identificadores ou extratores de SNs consigam melhorar os seus desempenhos.

Os valores semânticos de algumas palavras ainda são confundidos na etapa de etiquetagem, parte essencial para identificação dos SNs posteriormente. Quando na etapa de marcar as palavras de acordo com suas categorias gramaticais ocorrem equívocos, as expressões podem ser identificadas como SN, sendo outro tipo de sintagma como preposicionado, verbal, adjetival ou adverbial. Isso pode ser resolvido quando identificados os fenômenos sinonímicos e polissêmicos, categorizando as palavras de acordo com o contexto linguístico. Esse problema foi apresentado por todos os *softwares* em uma quantidade proporcional aos SNs

identificados por cada um. O LX-Parser cometeu um erro muito inesperado ao etiquetar alguns números como verbo, conforme o círculo vermelho na Figura 17.

Figura 17 – Erro de etiquetagem do LX-Parser



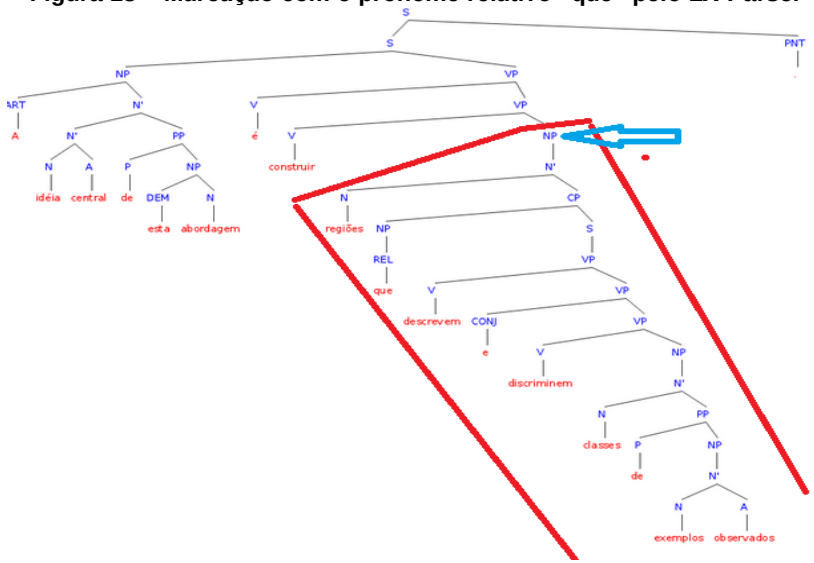
Fonte: LX Group. Disponível em: <http://lxcenter.di.fc.ul.pt/services/pt/LXParserPT.html>. Acesso em: 6 jan. 2014.

Outro erro cometido pelas ferramentas é em relação ao uso do pronome relativo “que”, que segundo a gramática normativa fica entre um substantivo e um verbo, assim a palavra que vem logo depois do pronome relativo “que” é marcado como parte do SN. O programa que cometeu mais

esse erro foi o LX-Parser, contudo as outras duas ferramentas também cometeram esse equívoco, para exemplificar tem-se a expressão "*a idéia central desta abordagem é construir* ***regiões que descrevem*** *e discriminem classes de exemplos observados*" analisada pelo LX-Parser e pelo PALAVRAS e apresentada na Figura 18 e Figura 19.

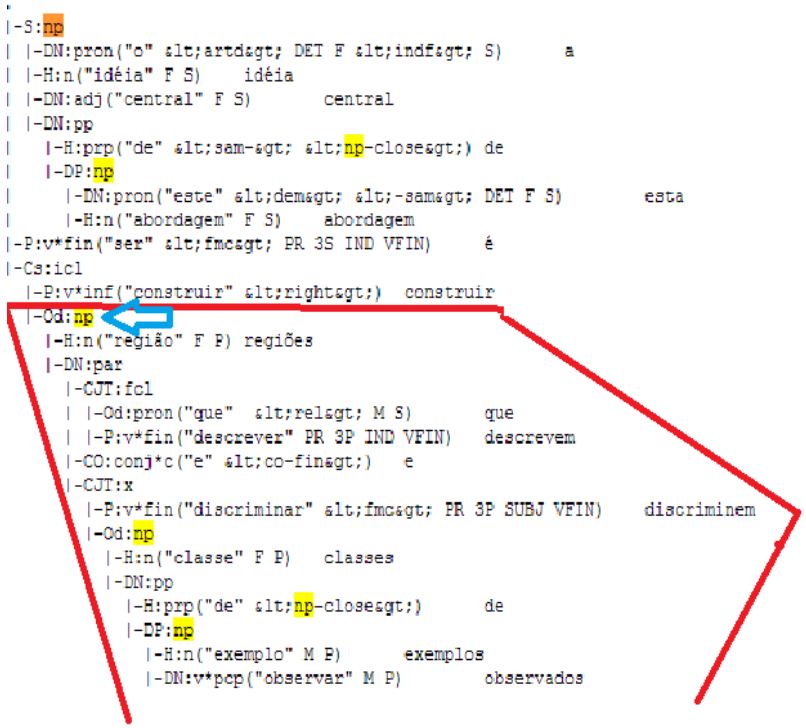
Quando houver esse tipo de situação, os programas deveriam marcar o SN até o substantivo preposicionado ao pronome relativo "que", logo em seguida identificar o "que" (pronome relativo) como SN, marcando as outras palavras subsequentes como outro tipo de sintagma, no caso do exemplo como sintagma verbal.

Figura 18 – Marcação com o pronome relativo "que" pelo LX-Parser



Fonte: LX Group. Disponível em: <http://lxcenter.di.fc.ul.pt/services/pt/LXParserPT.html>. Acesso em: 6 jan. 2014.

Figura 19 – Marcação com o pronome relativo “que” pelo Parser PALAVRAS



Fonte: VISL. Disponível em: <http://beta.visl.sdu.dk/visl/pt/parsing/automatic/trees.php>. Acesso em: 06 jan. 2014

Há outra dificuldade enfrentada pelos *softwares*, para a qual existe uma maneira bem simples para superá-la, gramaticalmente falando. Quando a regência nominal tem uma preposição que é seguida de um verbo, esse passa a ser considerado como parte do SN, por exemplo, “o direito de despedir”. Vale ressaltar que se tem o verbo, no exemplo, na forma nominal, ou seja, quando o comportamento dele se assemelha a um substantivo, adjetivo ou advérbio. As formas

nominais (infinitivo – caso do exemplo-, gerúndio e o particípio) atrapalharam muito o desempenho do PALAVRAS, principalmente do LX-Parser, fazendo com que eles marcassem o “despedir” como SN, quando desmembrado do “o direito de”, o fato é que o referido verbo faz parte de um SN, não constituindo, assim, isoladamente tal expressão nominal, como se vê quando o referido SN é visto em seu nível 1:

Nível 2: *O direito de despedir*

Nível 1: *despedir*

As preposições mais presentes nessas situações são “de” e “para”. A solução para esse problema seria a construção de um banco de palavras para cada radical de um léxico, assim, ao se deparar com essa dificuldade, o programa buscaria um substantivo que remetesse ao verbo, tendo para o “despedir” do exemplo a substituição por “dispensa” ou “demissão”.

Para problemas como os sintagmas preposicionados sendo identificados como SN, fica a sugestão da eliminação da preposição inicial, podendo reavaliar a organização das palavras subsequentes se elas se enquadram em uma estrutura de SN.

Quanto ao problema do LX-Parser começar alguns SNs com conjunção e sinais de pontuação, tem-se como solução a eliminação desses na hora em que é apresentada a estrutura arbórea da sentença para o usuário.

Todavia, todas as ferramentas têm seus méritos e devem ter um maior número de estudos que atentem para

seus objetivos, aperfeiçoando suas funcionalidades no intuito de que facilite o trabalho do indexador que, por sua vez, ajude o usuário a suprir suas necessidades informacionais.

6 CONSIDERAÇÕES FINAIS

Quanto aos objetivos dessa pesquisa, acredita-se que eles tenham sido atingidos, permitindo que sugestões fossem dadas para a melhoria das análises morfossintáticas das ferramentas automáticas de identificação e extração de SNs. No levantamento do estado arte, percebeu a existência de várias metodologias e ferramentas de identificação e/ou extração automática de SNs para textos escritos em língua portuguesa.

O levantamento e as sínteses feitos a partir dos trabalhos existentes permitiram perceber que os fundamentos teóricos da indexação automática por meio da extração de SNs se restringem a três áreas do conhecimento: a CI, Linguística e Ciência da Computação. A coleta e análises dos trabalhos e pesquisas sobre a referida temática permitiram identificar, na literatura, critérios de extração e seleção automáticas de SNs com valor de descritores.

Com o levantamento de *softwares* extratores e identificadores de SNs para textos em Português, foram feitas avaliações e comparações de tais *softwares* com base na representação arbórea de sentenças estudada em Linguística, especialmente pela Gramática Gerativa.

As contribuições para a Ciência da Informação dada por essa pesquisa são a retomada e apresentação de ferramentas que sendo aperfeiçoadas podem auxiliar no processo de escolha automática de melhores termos descritores de documentos.

O estado da arte levantado permitiu perceber que as ferramentas ou metodologias são desenvolvidas e aperfeiçoadas a partir de outras, por exemplo, o OGMA (MAIA, 2008), desenvolvido de acordo com as metodologias sugeridas por Miorelli (2001) e Souza (2005). Alguns grupos se formaram para desenvolver pesquisas acerca do Processamento de Linguagem Natural para diferentes fins como NILC e NLX-Group.

Nos trabalhos desenvolvidos no âmbito da Linguística acerca de SNs, a ênfase está nas regras de formação de SNs, já na Ciência da Informação os estudos objetivam encontrar dentre os SNs extraídos de um texto os que mais se enquadram como descritores de um documento, enquanto os trabalhos desenvolvidos na área da Ciência da Computação estão focados na maneira em que o computador possa fazer a extração e seleção de SNs automaticamente.

O parser PALAVRAS é a ferramenta mais usada na identificação de SNs, foi desenvolvida por Bick (2000) e permite que análises morfossintáticas sejam feitas com maior precisão. Dessa forma, autores como Miorelli (2001), Santos (2005), Souza (2005), Morellato (2010), entre outros, fazem uso do PALAVRAS para compor uma das fases das metodologias ou para fazer as validações das metodologias propostas pelos referidos pesquisadores.

Acerca dos trabalhos desenvolvidos sobre a extração e/ou seleção de SNs, pode-se traçar uma linha contínua com os trabalhos de Kuramoto (1995; 1997; 1999; 2002; 2006), que trabalha com a proposição de regras que possibilitem a identificação de sintagmas nominais, de Souza (2005) que

cria taxas de relevância para os sintagmas nominais, de Maia (2008) que cria uma ferramenta que implementa essas regras e metodologias de pontuação de relevância em um *software* e de Corrêa et al (2011) que avalia esse *software* na indexação automática de resumos de teses e dissertações.

Os trabalhos Kuramoto (2005), Miorelli (2001), Souza (2005), Santos (2005) e Costa (2007) almejavam a obtenção de altos valores de precisão e de abrangência na extração de SNs para que pudessem ser aplicados nos desenvolvimentos de sistemas de RI em diversas áreas (MORELLATO, 2010).

As principais métricas utilizadas na identificação, extração e seleção de SNs foram a acúracia, precisão, abrangência e equivalência, tendo a avaliação dos resultados a partir de referências como corpora etiquetados, extração manual de SNs, palavras-chave de textos, identificação ou extração de SN por outro *software*, análise da saída do *software* pelos autores.

Desse modo, percebeu-se, nessa pesquisa, que há várias ferramentas e métodos para fazer a identificação de SNs já apresentados na literatura, contudo a maioria dos trabalhos desenvolvidos fica restrita a um trabalho de conclusão de curso de graduação ou de pós-graduação. Muitos dos programas desenvolvidos são feitos em código fechado e não estão disponíveis na Internet. Há algumas ferramentas de identificação de SN que foram desenvolvidas no âmbito acadêmico, pesquisas financiadas com recursos públicos, mas que não estão disponibilizadas, o que compromete o andamento dos trabalhos por novos pesquisadores. Vale ressaltar que não se pretende adentrar

numa discussão política aqui nessas considerações finais, apenas sugerir uma reflexão.

A criação de novas ferramentas que permitam a satisfação do usuário em sua busca deve continuar sendo o objetivo dos estudiosos do objeto informação. Ter um bom processo de RI é ter os documentos descritos por termos contextualizados e que represente, de fato, o assunto do documento sem descaracterizá-lo. Dessa maneira, a presente pesquisa corrobora em dizer que os SNs são essenciais descritores dos documentos e que as pesquisas sobre a extração automática de SNs em língua portuguesa devem receber uma atenção maior para que elas atinjam um grau de amadurecimento elevado, aperfeiçoando as ferramentas automáticas existentes e/ou a criação de outras novas e eficientes.

Os resultados obtidos na comparação entre as ferramentas automáticas (em que o LX-Parser, em alguns casos, mostrou uma performance melhor que o PALAVRAS) permitem dizer que a extração automática de informação de um documento por meio de SNs deve ser estimulada para que aconteça a aplicabilidade nos SRIs efetivamente. O número de falsos SNs foi muito mais alto para o LX-Parser do que para o PALAVRAS e o OGMA. O PALAVRAS alcançou uma taxa de erro de 6%, o que corrobora a literatura quando se diz que essa ferramenta, diante de todo contexto, tem maior potencial para a identificação de SNs.

Quanto aos resultados de SNs relevantes, os quais foram tomados como referências os SNs que se assemelhavam as palavras-chave dos resumos ou as

continham em suas estruturas, fica a ressalva de que algumas palavras que não se enquadravam nesses dois grupos tinham grande potencial de serem descritores documentais. Às vezes, algumas palavras-chave eram menos relevantes que outras palavras que apareciam nos resumos. Desse modo, sugere-se, em **trabalhos futuros**, uma atenção maior a esse tipo de palavra, pois as palavras-chave são escolhidas pela autoria do texto, a qual pode escolhê-las inadequadamente.

O uso dos SNs no processo de indexação de documentos vai além de recuperar documentos relevantes para o usuário, é uma questão, também, de inclusão social, o que poderá ser visto em **trabalhos futuros** com a descrição de documentos para pessoas com deficiência auditiva, pois como são apontados por Crato e Cárnio (2009), os surdos apresentam dificuldades na escrita do português, principalmente no uso de verbos. Diante dessa constatação, os SNs seriam alternativas para que esses usuários pudessem navegar pelos níveis sintagmáticos nominais até encontrar a informação que procura, como sugeriu Kuramoto (1995) com o seu protótipo de interface de busca que se baseava no encadeamento hierárquico existente entre os SNs, fazendo com que haja interatividade entre a estrutura arbórea e o usuário.

Outras contribuições que os SNs podem dar para trabalhos futuros seriam em relação à confecção de mapas conceituais, que poderiam ser usados também como alternativa na representação de documentos, ao ler os descritores de um documento, o usuário poderiam também ter possibilidade de ver aquele documento por meio de um mapa

conceitual, o que possibilitaria confirmar se aquele documento é relevante para a sua busca.

Espera-se que esse trabalho tenha contribuído, inspirando novas ideias para os trabalhos sobre a extração automática de SNs. Pretende-se continuar os estudos comparativos entre os *softwares* extratores/identificadores de SNs levando em consideração a taxa de revocação e precisão de SNs relevantes para a descrição de documentos.

REFERÊNCIAS

ANDREEWSKI, Alexandre; RUAS, Vitoriano. Indexação automática baseada em métodos linguísticos e estatísticos e sua aplicabilidade à língua portuguesa. **Ciência da Informação**, Brasília, v. 12, n. 1, p. 61-73, 1983.

ARANHA, Christian; PASSOS, Emmanuel. A tecnologia de mineração de textos (Artigo Tutorial). **RESI-Revista Elerônica de Sistemas de Informação**, Curitiba, v. 5, n. 2, p. 01-08, 2006.

ARAÚJO JÚNIOR, Rogério Henrique de. **Precisão no processo de busca e recuperação da informação**. Brasília: Thesaurus, 2007.

ARCOVERDE, João Marcelo Azevedo. **Indução de filtros lingüisticamente motivados na recuperação de informação.** 2007. 107 f. Dissertação (Mestrado) - Curso de Ciências - Ciências de Computação e Matemática Computacional, Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo - USP, São Carlos, 2007.

ASSOCIAÇÃO BRASILEIRA DE NORMAS TÉCNICAS – ABNT. **NBR 12676**: Métodos para análise de documentos – determinação de seus assuntos e seleção de termos de indexação. Rio de Janeiro, 1992.

BAEZA-YATES, Ricardo; RIBEIRO-NETO, Berthier. **Modern information retrieval.** New York: Addison Wesley, 1999. 513 p. (ACM Press Series).

BARRETO, Aldo de Albuquerque. A condição da informação. **São Paulo em Perspectiva***,* São Paulo, v.16, n.3, p.1-12, jul./set. 2002.

BARROS, Flávia Almeida; ROBIN, Jacques. Processamento de Linguagem Natural. **SBC**: Revista Eletrônica de Iniciação Científica, Porto Alegre, v. 1, n.2, nov. 2001.

BERBER SARDINHA, Tony. Lingüística de Corpus: uma entrevista com Tony Berber Sardinha. **Revista Virtual de Estudos da Linguagem - ReVEL**. Vol. 2, n. 3, agosto de 2004.

BIBLIOTECA NACIONAL DE PORTUGAL. REPOSITÓRIO NACIONAL DE DISSERTAÇÕES E TESES DIGITAIS – DITED. Disponível em: <http://dited.bn.pt>. Acesso em: 05 fev. 2013.

BICK, Eckhard. Structural Lexical Heuristics in the Automatic Analysis of Portuguese. In: NORDIC CONFERENCE ON COMPUTATIONAL LINGUISTICS, 11. **Proceedings**… Copenhagen: Nodalida ‘98, 1998. p. 44 - 56.

_____. **The Parsing System Palavras:** Automatic Grammatical Analysis of Portuguese in a Constraint Grammar Framework. 2000. 505 f. Tese (Doutorado) - Curso de Linguística, Departamento de Department Of Linguistics, University Of Aarhus, Arthus, 2000.

BORGES NETO, José. O empreendimento gerativo. In: MUSSALIM, Fernanda; BENTES, Ana Christina (Org.). **Introdução à linguística:** fundamentos epistemológicos. 5. ed. São Paulo: Cortez, 2011. Cap. 3, p. 93-129. v.3.

BORGES, Graciane Silva Bruzinga. **Indexação automática de documentos textuais: proposta de critérios essenciais**. 2009. 111 f. Dissertação (Mestrado em Ciência

da Informação) – Programa de Pós-Graduação em Ciência da Informação, Escola de Ciência da Informação, Universidade Federal de Minas Gerais – UFMG, Belo Horizonte, 2009.

BRITO, Marcílio. Sistemas de informação em linguagem natural: em busca de uma indexação automática. **Ci. Inf.**, Brasília, v.21, n. 3, p. 223-232, set./dez. 1992.

BROOKSHEAR, J. Glenn. **Ciência da Computação**: uma visão abrangente. 7. ed. Porto Alegra: Bookman/ARTMED, 2003.

CAMARA JUNIOR, Auto Tavares da. **Indexação automática de acórdãos por meio de processamento de linguagem natural**. 2007. 141 f. Dissertação (Mestrado em Ciência da Informação) - Departamento de Ciência da Informação e Documentação, Universidade de Brasília, Brasília, 2007.

CAPURRO, Rafael. Epistemologia e Ciência da Informação. In: ENCONTRO NACIONAL DE PESQUISA EM CIÊNCIA DA INFORMAÇÃO – ENANCIB, 5. **Anais**... Belo Horizonte: ECI/UFMG, 2003.

CARDIE, Claire; PIERCE, David. Error-driven pruning of treebank grammars for base noun-phrase identification**.** In: Annual Meeting of the Association for Computational Linguistics, 36. **Proceedings**... Stroudsburg**:** Association for Computational Linguistics**,** 1998. p. 218-224.

CARDOSO, Ana Maria Pereira. Pós-modernidade e informação: conceitos complementares?. **Perspectivas em Ciência da Informação**, Belo Horizonte, v. 1, n. 1, p. 63-79, jan./jun. 1996.

CHOMSKY, Noam. **Syntact Structures**. Haia: Mouton, 1957.

_____. **Aspects of the theory os syntax**. Cambridge, Mass.: MIT Press, 1965.

Ciência da Informação. Disponível em: < http://revista.ibict.br/ciinf/index.php/ciinf>. Acesso em: 05 jul. 2013.

COMARELLA, Rafaela Lunard; CAFE, Lígia Maria Arruda. Chatterbot: conceito, características, tipologia e construção. **Informação & Sociedade: Estudos**, João Pessoa, v. 18, n. 2, p. 55-67, maio/ago. 2008.

Conjugue. Disponível em: < http://www.ime.usp.br/~ueda/br.ispell/conjugue.html>. Acesso em: 05 jul. 2013.

COREIXAS, Tatiane de Moraes; VIEIRA, Renata. Sobre a resolução de correferência. In: ENCONTRO DO CÍRCULO DE ESTUDOS LINGUÍSTICOS DO SUL, 8. **Anais...** Pelotas: Educat, 2008. p. 1-17.

Corpus TeMário. Disponível em: <http://www.linguateca.pt/Repositorio/TeMario/>. Acesso em: 03 abr. 2013.

CORRÊA, Renato Fernandes; MIRANDA, Darliane Goes de; LIMA, Camila Oliveira de Almeida; SILVA, Tiago José da. Indexação e recuperação de teses e dissertações por meio de sintagmas nominais. **AtoZ**: Novas Práticas em Informação e Conhecimento, Curitiba, v. 1, n. 1, p. 11-22, 2011.

COSTA, Francisco Nuno Quintiliano Mendonça Carapeto. **Deep Linguistic Processing of Portuguese Noun Phrases.** 2007. 214 f. Dissertação (Mestrado) - Curso de Mestrado em Informática, Departamento de Informática, Faculdade de Ciências da Universidade De Lisboa, Lisboa, 2007.

CRATO, Aline Nascimento; CARNIO, Maria Silvia. Análise da flexão verbal de tempo na escrita de surdos sinalizadores. **Rev. bras. educ. espec.**, Marília , v. 15, n. 2, Aug. 2009.

CUNHA, Angélica Furtado da; COSTA, Marcos Antonio; MARTELOTTA, Mário Eduardo. Linguística. In: MARTELOTTA, Mário Eduardo (Org.). **Manual de Linguística.** 2. ed. São Paulo: Contexto, 2012. Cap. 1, p. 15-30.

DataGramaZero. Disponível em: <http://www.dgz.org.br/>. Acesso em: 05 jul. 2013.

DAVID, Karine Alves. **Sintaxe das expressões nominais no português do Brasil:** uma abordagem computacional. 2007. 118 f. Dissertação (Mestrado) - Curso de Mestrado em Linguística, Centro de Humanidades, Universidade Federal do Ceará - UFC, Fortaleza, 2007.

DIAS, Eduardo Wense. O específico da ciência da informação. In: AQUINO, Mirian de Albuquerque. **O campo da ciência da informação***:* gênese, conexões e especificidades. João Pessoa: Editora Universitária/UFPB, 2002. p. 87-99.

Dicionário BR/ISPELL. Disponível em: < http://www.ime.usp.br/~ueda/br.ispell/>. Acesso em: 02 abr. 2013.

DUBOIS-CHARLIER, Françoise. **Base de análise linguística**. Coimbra: Almedina, 1977.

EVANS, David Andreoff. **A Summary of the CLARIT project**. 1991. Disponível em: <http://repository.cmu.edu/philosophy/457/ >. Acesso em: 10 jun. 2013.

FARMER, Linda. Automatic categorization: what's it all about?. **Serials Librarian**, v. 51, n. 2, p. 91-101, 2006.

GIL, Antonio Carlos. **Como elaborar projetos de pesquisa**. 4. ed. São Paulo: Atlas, 2002.

GONZALEZ, Marco.; LIMA, Vera Lúcia Strube. Recuperação de Informação e Processamento da Linguagem Natural. In: Congresso da Sociedade Brasileira de Computação, 23; Jornada de Mini-Cursos de Inteligência Artificial, 3. **Anais…** Campinas: SBC, 2003. p. 347-395.

GOOGLE ACADÊMICO. Disponível em: <http://scholar.google.com.br/>. Acesso em: 07 fev. 2013.

HJØRLAND, Birger. Automatic Indexing. In: _____. **Lifeboat for Knowledge Organization**. [s.l.]:[s.n.], 2008. Disponível em: <http://www.iva.dk/bh/lifeboat_ko/CONCEPTS/automatic_indexing.htm >. Acesso em: 23 abr. 2013.

HOLANDA, Cínthia; BRAZ, Márcia Ivo. INDEXAÇÃO AUTOMÁTICA DE CONTEÚDOS NA WEB: análise de sites de museus. **Biblionline**, João Pessoa, v. 8, n. 1, p. 42-59, 2012.

HOUAISS, Antônio. **Dicionário eletrônico Houaiss da língua portuguesa.** Rio de Janeiro: Objetiva, 2001. CD-ROM.

INSTITUTO BRASILEIRO DE INFORMAÇÃO EM CIÊNCIA EM TECNOLOGIA – IBICT. BIBLIOTECA DIGITAL BRASILEIRA DE TESES E DISSERTAÇÕES – BDTD. Disponível em: <http://bdtd.ibict.br/>. Acesso em: 05 fev. 2013.

KENEDY, Eduardo. Gerativismo. In: MARTELOTTA, Mário Eduardo (Org.). **Manual de Linguística.** 2. ed. São Paulo: Contexto, 2012. Cap. 8, p. 127-140.

KRIEGER, Maria da Graça; FINATTO, Maria José Bocorny. **Introdução à terminologia**: teoria e prática. São Paulo: Contexto, 2004.

KURAMOTO, Hélio. Uma abordagem alternativa para o tratamento e a recuperação de informação textual : os sintagmas nominais. **Ciência da Informação**, Brasília, v. 25, n. 2, p. 182-192, maio/ago. 1995.

_____. Proposta de um Sistema de Recuperação de Informação Assistido por Computador - SRIAC. **Revista de Biblioteconomia de Brasília**, Brasília, v. 21, n. 2, p. 211-228, jul./dez. 1997.

. **Proposition d'un système de recherche d'information assistée par ordinateur:** avec application au portugais. 1999. 1 v. Thèse (Doctorat) - Curso de Doctorat En Sciences de L'information Et de La Communication, Université Lumière–lyon 2, Lyon, França, 1999.

_____. Sintagmas nominais: uma nova proposta para a recuperação de informação. **DataGramaZero**, Rio de Janeiro, v. 3, n. 1, fev. 2002.

_____. Sintagmas nominais: uma nova abordagem no processo de indexação. In: NAVES, Madalena Martins Lopes; KURAMOTO, Hélio (Orgs.). **Organização da informação**: princípios e tendências. Brasília: Briquet de Lemos/Livros, 2006. p. 117-137.

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

LE COADIC, Yves-François. **A ciência da informação**. Brasília: Briquet de Lemos/Livros, 2004.

LIBERATO, Yara Goulart. **A estrutura do SN em português:** uma abordagem cognitiva. 1997. 2003 f. Tese (Doutorado) - Curso de Doutorado em Letras, Faculdade de Letras, Universidade Federal de Minas Gerais – UFMG, Belo Horizonte, 1997.

LOPES, Lucelene. **Extração automática de conceitos a partir de textos em língua portuguesa.** 2011. 156 f. Tese (Doutorado) - Curso de Ciência da Computação, Faculdade de Informática, Pontifícia Universidade Católica Do Rio Grande Do Sul - PUCRS, Porto Alegre, 2011.

LXGram. Disponível em: <http://nlxgroup.di.fc.ul.pt/lxgram/>. Acesso em: 15 jun. 2013.

LX-Parser. Disponível em: <http://lxcenter.di.fc.ul.pt/tools/pt/conteudo/LX-Parser.html>. Acesso em 17 jun. 2013.

MAIA, Luis Claudio Gomes. **Uso de sintagmas nominais na classificação automática de documentos eletrônicos**. 2008. 158 f. Tese (Doutorado) – Curso de Doutorado em Ciência da Informação, Escola de Ciência da Informação, Universidade Federal de Minas Gerais - UFMG, Belo Horizonte, 2008.

MAIA, Luiz Cláudio Gomes; SOUZA, Renato Rocha . Uso de sintagmas nominais na classificação automática de documentos eletrônicos. **Perspectivas em Ciência da Informação**, v. 15, p. 154-172, 2010.

MAIMONE, Giovana Deliberali; FRANCELIN, Marivalde Moacir. Indexação e resumos: produtos documentários na organização da informação. **Revista Científica do IMAPES**, v. 6, p. 75-82, 2008.

MARCONI, Marina de Andrade; LAKATOS, Eva Maria. **Fundamentos de metodologia científica**. 6. ed. São Paulo: Atlas, 2009.

MARTELOTTA, Mário Eduardo. Conceitos de Gramática. In: _____ (Org.). **Manual de Linguística**. 2. ed. São Paulo: Contexto, 2012.

MARTINS, Ronaldo Teixeira; HASEGAWA, Ricarod; NUNES, Maria das Graças Volpe. Curupira: a functional parser for Brazilian Portuguese. In: Computational Processing of the Portuguese Language (PROPOR). International Workshop, 6. **Proceedings…** Faro, Portugal: LNAI 2721, 2003. v.1. p. 179-183.

MEADOW, Charles T.; BOYCE, Bert R.; KRAFT, Donald H. **Text Information Retrieval Systems**. San Diego: Academic Press, 2000.

MINISTÉRIO DA EDUCAÇÃO E CIÊNCIA (PORTUGAL). REPOSITÓRIO CIENTÍFICO DE ACESSO ABERTO DE PORTUGAL. Disponível em: <http://www.rcaap.pt>. Acesso em: 06 fev. 2013.

MINISTÉRIO DA EDUCAÇÃO. COORDENAÇÃO DE APERFEIÇOAMENTO DE PESSOAL DE NÍVEL SUPERIOR – CAPES. PORTAL DE PERIÓDICOS DA CAPES. Disponível em: <http://www.periodicos.capes.gov.br>. Acesso em: 15 fev. 2013.

MIORELLI, Sandra Teresinha. **Extração do sintagma nominal em sentenças em português**. 2001. 98 f. Dissertação (Mestrado em Ciência da Computação) – Faculdade de Informática, Pontifícia Universidade Católica do Rio Grande do Sul, Porto Alegre, 2001.

MORELLATO, Luana Vieira. **SIDSN**: Sistema Identificador de Sintagmas Nominais. 2007. 58 f. Monografia (Bacharelado

em Ciência da Computação) – Departamento de Informática, Centro Tecnológico, Universidade Federal do Espírito Santo, Vitória, 2007.

_____. **Metodologia Computacional para Identificação de Sintagmas Nominais na Língua Portuguesa**. 2010. 112 f. Dissertação (Mestrado em Ciência da Computação) – Departamento de Informática, Centro Tecnológico, Universidade Federal do Espírito Santo, Vitória, 2010.

MUSSALIM, Fernanda; BENTES, Ana Christina (Org.). **Introdução à linguística**: fundamentos epistemológicos. 5. ed. São Paulo: Cortez, 2011. v. 3.

NASCIMENTO, Denise Morado. A abordagem sócio-cultural da informação. **Informação & Sociedade: Estudos**, João Pessoa, v. 16, n. 2, p. 25-35, jul./dez. 2006.

NAVARRO, Sandrelei. Interface entre linguística e indexação: uma revisão de literatura. **Rev. Bras. Biblio. Doc.**, São Paulo, v.21, n. 1/2, p. 46-62, jan./jun. 1988.

NÚCLEO INTERINSTITUCIONAL DE LINGUÍSTICA COMPUTACIONAL. Disponível em: <http://www.nilc.icmc.usp.br/nilc/tools/nilctaggers.html>. Acesso em: 13 abr. 2013.

_____. **Curupira**. Disponível em: <http://www.nilc.icmc.usp.br/nilc/tools/curupira .html>. Acesso em: 13 abr. 2013.

O CONSTRUCTOR. Disponível em: <http://www.reocities.com/Athens/crete/1546 /indexc.htm>. Acesso em: 20 jun. 2013.

OGMA. Disponível em: <http://www.luizmaia.com.br/ogma/>. Acesso em: 16 maio 2013.

ORTEGA, Cristina Dotta. Relações históricas entre Biblioteconomia, Documentação e Ciência da Informação. **DataGramaZero**, Rio de Janeiro, v. 5, n. 5, , out. 2004.

OTHERO, Gabriel de Ávila. **Grammar Play**: um parser sintático em Prolog para a língua portuguesa. 2004. 265 f. Dissertação (Mestrado em Letras) – Faculdade de Letras, Pontifícia Universidade Católica do Rio Grande do Sul - PUCRS, Porto Alegre, 2004.

_____. **A Gramática da Sentença em Português**: uma descrição formal com um "olho" na implementação computacional. 2008. 213 f. Tese (Doutorado) - Curso de Letras, Faculdade de Letras, Pontifícia Universidade Católica Do Rio Grande Do Sul - PUCRS, Porto Alegre, 2008.

PERINI, Mário Alberto. **Para uma nova gramática do português**. 8. Ed. São Paulo: Ática, 1995.

_____. **Gramática descritiva do português**. 4. ed. São Paulo: Ática, 2003.

PERINI, Mário Alberto et al.. O Sintagma nominal em português: estrutura, significado e função. **Revista de Estudos da Linguagem**. Belo Horizonte, v. 5, n. especial, p. 01-180, jul./dez. 1996.

PINHEIRO, Lena Vânia Ribeiro. Campo interdisciplinar da Ciência da Informação: fronteiras remotas e recentes. In: _____. **Ciência da Informação, ciências sociais e interdisciplinaridade.** Brasília: IBICT, 1999. p. 155-182.

PINHEIRO, Lêna Vania Ribeiro; LOUREIRO, José Mauro Matheus. Traçados e limites da ciência da informação. **Ciência da Informação**, Brasília, v. 24, n. 1, p. 42-53, jan./abr. 1995.

PINHEIRO, Marcello Sandi. **Uma abordagem usando sintagmas nominais como descritores no processo de mineração de opiniões**. 2009. 110 f. Tese (Doutorado em Engenharia Civil) – COPPE, Universidade Federal do Rio de Janeiro, Rio de Janeiro, 2009.

RAMSHAW, Lance A. ; MARCUS, Mitchell P. Text chunking using transformation-based learning. In: Workshop on Very Large Corpora, 3. **Proceedings**... New Jersey: Association for Computational Linguistics , 1995. p. 82–94.

RANGANATHAN, Shiyali Ramamrita. **As cinco leis da Biblioteconomia**. Brasília: Briquet de Lemos Livros, 2009.

ROBREDO, Jaime. **Da ciência da informação revisitada aos sistemas humanos de informação.** Brasília: Ed. Thesaurus; SSRR Informações, 2003.

SANTOS, Cícero Nogueira dos. **Aprendizado de máquina na identificação de sintagmas nominais**: o caso do português brasileiro. 2005. 104 f. Dissertação (Mestrado em Sistemas e Computação) – Instituto Militar de Engenharia, Rio de Janeiro, 2005.

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

_____. Information Science. **JASIS:** Journal of The American Society for Information Science, New York, v. 50, n. 12, p. 1051-1063, 1999.

SciELO BRASIL. Disponível em: <http://www.scielo.br>. Acesso em: 20 fev. 2013.

SciELO. Disponível em: <http://www.scielo.org>. Acesso em: 20 fev. 2013.

SILVA, Maria Cecília Pérez de Souza e; KOCH, Ingedore Grunfeld Villaça. **Linguística aplicada ao português**: sintaxe. 15.ed. São Paulo: Cortez, 2009.

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

_____. Uma proposta de metodologia para indexação automática utilizando sintagmas nominais. **Encontros Bibli: Revista Eletrônica de Biblioteconomia e Ciência da Informação**, Florianópolis, v. 11, n. esp., p. 42-59, 1º sem. 2006.

TR+. FORMA. Disponível em: <http://www.inf.pucrs.br/~gonzalez/tr+/forma/>. Acesso em 16 jul. 2013.

UNIVERIDADE FEDERAL DO PARANÁ – UFPR. BASE DE DADOS REFERENCIAIS DE ARTIGOS DE PERIÓDICOS EM CIÊNCIA DA INFORMAÇÃO – BRAPCI. Disponível em: < http://www.brapci.ufpr.br>. Acesso em: 15 jan. 2013

UNIVERSIDADE DE LISBOA. DEPARTAMENTO DE INFORMÁTICA. NATURAL LANGUAGE AND SPEECH GROUP – NLX-GROUP. Disponível em: <http://nlx.di.fc.ul.pt/index.html>. Acesso em: 15 jun. 2013.

UNIVERSIDADE FEDERAL DE PERNAMBUCO – UFPE. BIBLIOTECA DIGITAL DE TESES E DISSERTAÇÕES – BDTD. Disponível em: <http://www.bdtd.ufpe.br>. Acesso em: 20 set. 2012

VICKERY, Brian C.; VICKERY, Alina. **Information science in theory and practice.** London: Bowker-Saur, 1987.

VIEIRA, Renata et al. Extração de sintagmas nominais para o processamento de co-referência. In: Encontro para o Processamento Computacional do Português Escrito e Falado (PROPOR), 5. **Anais**... São Carlos: ICMC/USP, 2000. p. 165-173.

VIEIRA, Renata; LIMA, Vera Lúcia Strube. Linguística computacional: princípios e aplicações. In: CONGRESSO DA SOCIEDADE BRASILEIRA DE COMPUTAÇÃO, 21., 2001, Fortaleza. **Anais**... Fortaleza: SBC, 2001. v. 2, p. 47-88.

VIEIRA, Simone Bastos. Indexação automática e manual: revisão de literatura. **Ci. Inf.**, Brasília, v. 17, n. 1, p. 43-57, jan./jun. 1988a.

_____. Análise comparativa entre indexação automática e manual da literatura brasileira de Ciência da informação. **R. Bibliotecon**., Brasília, v. 16, n. 1, p. 83-94, jan./jun. 1988b.

VISUAL INTERACTIVE SYNTAX LEARNING – VISL. Disponível: <http://beta.visl.sdu.dk/> Acesso em: 10 jun. 2013.

VISUAL INTERACTIVE SYNTAX LEARNING – VISL. **Parsing**. Disponível em: <http://beta.visl.sdu.dk/visl/pt/parsing/automatic/dependency.php>. Acesso em: 30 maio 2013.

VISUAL INTERACTIVE SYNTAX LEARNING – VISL. **TREES**. Disponível em: <http://beta.visl.sdu.dk/visl/pt/parsing/automatic/trees.php>. Acesso em: 30 maio 2013.

VOUTILAINEN, Atro. **Nptool, a detector of English noun phrases.** 1995. Disponível em**:** http://arxiv.org/pdf/cmp-lg/9502010.pdf Acesso em: 03 mar. 2013.

WIVES, Leandro K. **Indexação de documentos textuais.** 1997. 19f. Trabalho Monográfico - Disciplina de Sistemas de Banco de Dados (Programa de Pós-Graduação em Ciência da Computação) - Instituto de Informática, Universidade Federal do Rio Grande do Sul, Porto Alegre, 1997. Orientadora: Prof. Lia Golendziner. Disponível em: < http://www.leandro.wives.nom.br/pt-br/publicacoes/IDT.pdf>. Acesso em 05 maio 2013.

WÜSTER, Eugen. **Introducción a la teoría general de la terminología y a la lexicografia terminológica**. Barcelona: Institut Universitari de Lingüístca Aplicada/Universitat Pompeu Fabra, 1998.

APÊNDICE

Extração de SNs dos resumos de teses e dissertações

Nos apêndices que se seguem, são trazidas, a título de exemplificação, as extrações de SNs feitas pela extração manual, pelo Parser PALAVRAS, pelo LX-Parser e pelo OGMA.

APÊNDICE A – Extração manual

Extração manual DI1

Expressão: Capacidade contributiva nas contribuições à Previdência Social: direitos fundamentais do cidadão-contribuinte e justiça fiscal.

1. Capacidade contributiva nas contribuições à Previdência Social
2. as contribuições à Previdência Social
3. a Previdência Social
4. direitos fundamentais do cidadão-contribuinte
5. o cidadão-contribuinte
6. justiça fiscal

Expressão: O Estado, desde quando passou a ser denominado Estado Moderno, vivenciou inúmeras transformações.

7. O Estado
8. Estado Moderno
9. Inúmeras transformações

Expressão: A fiscalidade, fenômeno essencial à existência do Estado, também tem sofrido mudanças, dentre as quais se destaca o grande aumento de contribuições percentualmente em relação à receita estatal tributária, que, devido a inúmeras razões, passam a substituir gradativamente os impostos diretos.

10. A fiscalidade
11. fenômeno essencial à existência do Estado
12. a existência do Estado
13. o Estado
14. mudanças
15. as quais

16. o grande aumento de contribuições percentualmente em relação à receita estatal tributária
17. contribuições percentualmente em relação à receita estatal tributária
18. relação à receita estatal tributária
19. a receita estatal tributária
20. que
21. inúmeras razões
22. os impostos diretos

Expressão: Diante dessa nova perspectiva do fenômeno tributário, torna-se fundamental adequar as contribuições aos direitos fundamentais do contribuinte.

23. Essa nova perspectiva do fenômeno tributário
24. o fenômeno tributário
25. as contribuições
26. os direitos fundamentais do contribuinte
27. o contribuinte

Expressão: O direito à existência, materializado pelo mínimo existencial, é um desses direitos fundamentais que deve ser inexoravelmente respeitado e preenchido, mas em seu aspecto ampliado que é justamente a condição humana, tal qual descreve Hannah Arendt.

28. O direito à existência
29. A existência
30. mínimo existencial
31. um desses direitos fundamentais
32. esses direitos fundamentais
33. que
34. seu aspecto ampliado
35. que
36. a condição humana
37. tal qual

38. Hannah Arendt

Expressão: Surge, assim, a necessidade de harmonizar as contribuições previdenciárias do cidadão ao mínimo existencial, através da aplicação do princípio da capacidade contributiva, mediado não pelos seus subprincípios clássicos, como a progressividade, mas pela proporcionalidade.

39. A necessidade
40. as contribuições previdenciárias do cidadão ao mínimo existencial
41. o cidadão ao mínimo existencial
42. o mínimo existencial
43. a aplicação do princípio da capacidade contributiva
44. o princípio da capacidade contributiva
45. a capacidade contributiva
46. seus subprincípios clássicos
47. a progressividade
48. proporcionalidade

APÊNDICE B – Extração pelo Parser PALAVRAS

PALAVRAS DI1

Expressão: capacidade contributiva nas contribuições à Previdência Social:

1. capacidade contributiva nas contribuições à Previdência Social
2. as contribuições à Previdência Social
3. a Previdência Social

Expressão: direitos fundamentais do cidadão-contribuinte e justiça fiscal.

4. direitos fundamentais do cidadão-contribuinte e justiça fiscal
5. o cidadão-contribuinte e justiça fiscal
6. justiça fiscal

Expressão: o estado, desde quando passou a ser denominado Estado Moderno, vivenciou inúmeras transformações.

7. O estado
8. Estado Moderno
9. Inúmeras transformações

Expressão: a fiscalidade, fenômeno essencial à existência do estado, também tem sofrido mudanças, de entre as quais se destaca o grande aumento de contribuições percentualmente em relação a a receita estatal tributária, que, devido a inúmeras razões, passam a substituir gradativamente os impostos diretos.

10. A fiscalidade
11. fenômeno essencial à existência do estado
12. a existência do estado
13. o estado
14. o grande aumento de contribuições
15. a receita estatal tributária
16. inúmeras razões
17. os impostos diretos

Expressão: diante de essa nova perspectiva do fenômeno tributário, torna-se fundamental adequar as contribuições aos direitos fundamentais do contribuinte.

18. essa nova perspectiva do fenômeno tributário
19. o fenômeno tributário
20. as contribuições
21. os direitos fundamentais do contribuinte
22. o contribuinte

Expressão: o direito à existência materializado por o mínimo existencial, é um desses direitos fundamentais que deve ser inexoravelmente respeitado e preenchido, mas em seu aspecto ampliado que é justamente a condição humana, tal qual descreve Hannah Arendt.

CAVE CIRCULARITY in sentence 6 at 39 (qual)

23. O direito
24. O mínimo existencial
25. Esses direitos fundamentais
26. Esses direitos
27. A condição humana
28. Hannah Arendt

Expressão: surge, assim, a necessidade de harmonizar as contribuições previdenciárias do cidadão ao mínimo existencial, através da aplicação do princípio da capacidade contributiva, mediado não por os seus subprincípios clássicos, como a progressividade, mas por a proporcionalidade.

29. a necessidade de harmonizar as contribuições previdenciárias do cidadão ao mínimo existencial, através da aplicação do princípio da capacidade contributiva, mediado não por os seus subprincípios clássicos, como a progressividade
30. a necessidade
31. as previdências do cidadão
32. o cidadão
33. o mínimo existencial
34. a aplicação do princípio da capacidade contributiva
35. o princípio da capacidade contributiva
36. a capacidade contributiva
37. a progressividade
38. a proporcionalidade

APÊNDICE C – Extração pelo LX-Parser

Extração de SNs LX DI1

Expressão: Capacidade contributiva nas contribuições à Previdência Social: direitos fundamentais do cidadão-contribuinte e justiça fiscal.

1. contributiva nas contribuições à Previdência Social: direitos fundamentais do cidadão-contribuinte e justiça fiscal
2. contributiva nas contribuições à Previdência Social
3. as contribuições à Previdência Social
4. a Previdência Social
5. direitos fundamentais do cidadão-contribuinte e justiça fiscal
6. o cidadão-contribuinte e justiça fiscal
7. o cidadão-contribuinte
8. e justiça fiscal

Expressão: O Estado, desde quando passou a ser denominado Estado Moderno, vivenciou inúmeras transformações.

9. O Estado
10. denominado Estado Moderno
11. inúmeras transformações

Expressão: A fiscalidade, fenômeno essencial à existência do Estado, também tem sofrido mudanças, dentre as quais se destaca o grande aumento de contribuições percentualmente em relação à receita estatal tributária, que, devido a inúmeras razões, passam a substituir gradativamente os impostos diretos.

12. A fiscalidade, fenômeno essencial à existência do Estado

13. A fiscalidade
14. fenômeno essencial à existência do Estado
15. a existência do Estado
16. o Estado
17. sofrido mudanças, dentre as quais se destaca o grande aumento de contribuições percentualmente em relação à receita estatal
18. as quais se destaca o grande aumento de contribuições percentualmente
19. quais
20. se (pronome)
21. o grande aumento de contribuições percentualmente
22. contribuições percentualmente
23. a receita estatal
24. tributária, que, devido a inúmeras razões
25. que
26. a inúmeras razões
27. os impostos diretos

Expressão: Diante dessa nova perspectiva do fenômeno tributário, torna-se fundamental adequar as contribuições aos direitos fundamentais do contribuinte.

28. Diante dessa nova perspectiva do fenômeno tributário
29. essa nova perspectiva do fenômeno tributário
30. o fenômeno tributário
31. se (pronome)
32. as contribuições aos direitos fundamentais do contribuinte
33. os direitos fundamentais do contribuinte
34. o contribuinte

Expressão: O direito à existência, materializado pelo mínimo existencial, é um desses direitos fundamentais que deve ser inexoravelmente

respeitado e preenchido, mas em seu aspecto ampliado que é justamente a condição humana, tal qual descreve Hannah Arendt.

35. O direito à existência, materializado pelo mínimo existencial
36. O direito à existência
37. A existência
38. materializado pelo mínimo existencial
39. o mínimo existencial
40. um desses direitos fundamentais que deve ser inexoravelmente respeitado e preenchido, mas em seu aspecto ampliado que é justamente a condição humana, tal qual descreve Hannah Arendt
41. esses direitos fundamentais que deve ser inexoravelmente respeitado e preenchido, mas em seu aspecto ampliado que é justamente a condição humana, tal
42. que
43. seu aspecto ampliado que é justamente a condição humana
44. que
45. a condição humana
46. qual
47. Hannah Arendt

Expressão: Surge, assim, a necessidade de harmonizar as contribuições previdenciárias do cidadão ao mínimo existencial, através da aplicação do princípio da capacidade contributiva, mediado não pelos seus subprincípios clássicos, como a progressividade, mas pela proporcionalidade.

48. A necessidade de harmonizar
49. Harmonizar
50. as contribuições previdenciárias do cidadão ao mínimo existencial, através da aplicação do princípio da capacidade contributiva, mediado não pelos seus subprincípios clássicos, como a progressividade, mas pela proporcionalidade

51. o cidadão ao mínimo existencial, através da aplicação do princípio da capacidade contributiva, mediado não pelos seus subprincípios clássicos, como a progressividade
52. o mínimo existencial, através da aplicação do princípio da capacidade contributiva, mediado não pelos seus subprincípios clássicos, como a progressividade
53. da aplicação do princípio da capacidade contributiva, mediado não pelos seus subprincípios clássicos, como a progressividade
54. a aplicação do princípio da capacidade contributiva
55. o princípio da capacidade contributiva
56. a capacidade contributiva
57. os seus subprincípios classicos
58. como a progressividade
59. a progressividade
60. a proporcionalidade

APÊNDICE D – Extração pelo OGMA

Extração de SNs OGMA DI1

1. o cidadão
2. contribuinte e justiça fiscal
3. denominado estado moderno
4. transformações
5. a fiscalidade
6. sofrido mudanças
7. dentre
8. percentualmente em relação
9. estatal tributária que devido a inúmeras
10. razões
11. gradativamente os impostos diretos diante
12. dessa nova perspectiva do fenômeno tributário
13. o contribuinte
14. existência materializado
15. seu aspecto ampliado
16. justamente a condição humana
17. hannah
18. arendt
19. assim a necessidade
20. as contribuições previdenciárias do cidadão
21. a o mínimo existencial
22. a aplicação do princípio da capacidade
23. contributiva mediado
24. o seus subprincípios clássicos
25. a progressividade
26. a proporcionalidade
27. o estado
28. fenômeno essencial à existência do estado
29. contributiva nas contribuições à previdência social direitos fundamentais do cidadão