UNIVERSIDADE FEDERAL DE PERNAMBUCO – UFPE
CENTRO DE ARTES E COMUNICAÇÃO – CAC
DEPARTAMENTO DE CIÊNCIA DA INFORMAÇÃO – DCI
GRADUAÇÃO EM GESTÃO DA INFORMAÇÃO – GI

TIAGO JOSÉ DA SILVA

AVALIAÇÃO DE FERRAMENTAS DE EXTRAÇÃO DE SINTAGMAS NOMINAIS: uma análise comparativa entre o OGMA e o OGMA Web

Recife
2014

TIAGO JOSÉ DA SILVA

AVALIAÇÃO DE FERRAMENTAS DE EXTRAÇÃO DE SINTAGMAS NOMINAIS: uma análise comparativa entre o OGMA e o OGMA Web

Trabalho de Conclusão de Curso apresentado ao Curso de Gestão da Informação da Universidade Federal de Pernambuco, como requisito parcial à obtenção do grau de Bacharelado em Gestão da Informação.

Orientador: Prof. Dr. Renato Fernandes Corrêa

Recife

2014

Catalogação na fonte
Andréa Marinho, CRB4-1667

S586a Silva, Tiago José da.
Avaliação de ferramentas de extração de sintagmas nominais: uma análise comparativa entre o OGMA e o OGMA Web / Tiago José da Silva. – Recife: O autor, 2014.
61 p.; Il.: fig., tab., graf. e quadros; 30 cm.

Orientador: Renato Fernandes Corrêa.
TCC (Graduação) – Universidade Federal de Pernambuco, CAC. Gestão da Informação, 2014.
Inclui referências e apêndices.

1. Recuperação da Informação. 2. Indexação Automática. 3. Processamento de Linguagem Natural. 4. Gramática Gerativa. I. Corrêa, Renato Fernandes. (Orientador). II. Titulo.

020 CDD (22.ed.) UFPE (CAC 2014-106)

Serviço Público Federal
Universidade Federal de Pernambuco
Centro de Artes e Comunicação
Departamento de Ciência da Informação

FOLHA DE APROVAÇÃO

Avaliação de ferramentas de extração de sintagmas nominais: uma análise comparativa entre o OGMA entre OGMA Web
(Título do TCC)

Tiago José da Silva
(Autor)

Trabalho de Conclusão de Curso submetido à Banca Examinadora, apresentado no Curso de Gestão da Informação, do Departamento de Ciência da Informação, da Universidade Federal de Pernambuco, como requisito parcial para obtenção do título de Bacharel em Gestão da Informação.

TCC aprovado em 11 de agosto de 2014.

Banca Examinadora:

Prof. Dr. Renato Fernandes Corrêa - Orientador
Universidade Federal de Pernambuco

Prof. Dr. Fábio de Assis Pinho - Examinador 1
Universidade Federal de Pernambuco

Prof. Me. André Anderson Cavalcante Felipe - Examinador 2
Universidade Federal de Pernambuco

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 Departamento de Ciência da Informação da UFPE pela oportunidade de aprender e desenvolver novas habilidades com o Curso de Bacharelado em Gestão da Informação, isso me fez crescer como profissional. Sou grato ao Corpo Docente do departamento pelo amadurecimento que tive ao longo desses quatro anos, em especial a Fábio Pinho, Fábio Mascarenhas, Cristina Oliveira, Nadi Presser, Raimundo Nonato, Sandra Siebra, Májory Miranda, André Fell, Marcos Galindo, Celly Brito, Edilene Silva e Vildeane Borba. Agradeço também ao coordenador do curso, professor Alexander Azevedo.

Sou agradecido ao meu orientador, Professor Dr. Renato Fernandes Corrêa, que sempre esteve comigo desde o começo do curso, dando orientações valiosíssimas que me fizeram atingir partes dos meus objetivos profissionais. – Renato, sempre serei grato por tudo, pelos puxões de orelha, pelos elogios e pela confiança. Muito Obrigado!

Agradeço aos professores Fábio Pinho e André Anderson pelas suas valiosas contribuições para o aperfeiçoamento desse trabalho, dadas na Banca de Avaliação.

Aos meus colegas do curso de Gestão da Informação que dividiam as mesmas dúvidas e conseguiam o mesmo sucesso, principalmente a Juliana Cysneiros (Minha irmã).

À Gabriela Ortega pelo companheirismo de sempre, dando sugestões sempre na hora certa. Aos meus amigos Liana e Erinaldo pelo apoio incondicional.

Agradeço aos meus colegas de trabalho Edilene Barreto, Ocidélia, Flávia Lima, Giovanna Weyne, Andrea Araújo, Laís Lira, Rafaely Maira, Ivson Marques, Vladirene Lima.

Agradeço a todos que de maneira direta ou indiretamente contribuíram para a minha formação em Gestão da Informação.

Os meus sonhos, o vento não pode levar,
A esperança encontrei no Teu olhar
Os meus sonhos, a areia não vai enterrar
Porque a vida recebi ao Te encontrar...
(JUNINHO AFRAM)

RESUMO

Levanta e sintetiza alguns fundamentos teóricos da indexação automática por meio da identificação, extração e seleção automáticas de sintagmas nominais em textos escritos em Língua Portuguesa. Conceitua Sintagmas nominais, Processamento de Linguagem Natural, Indexação automática, Identificação e extração automáticas de sintagmas nominais, assim como as ferramentas que desenvolvam essas últimas atividades e, por fim, seleção de sintagmas nominais como descritores documentais. Avalia e compara ferramentas de extração automática de sintagmas nominais como o OGMA tradicional e OGMA Web, usando como referência a extração manual de sintagmas nominais. Na comparação entre as referidas ferramentas automáticas, percebe-se que apesar do OGMA Web ter tido melhor desempenho em alguns aspectos como extrair um maior número de sintagmas nominais do que o OGMA tradicional esse ainda consegue ser melhor pelo número menor de erros. Sugere, então, algumas possíveis soluções para os problemas de extração de sintagmas nominais enfrentados pelas ferramentas automáticas. Conclui que o OGMA Web possui uma melhor interface, a qual permite que o usuário desenvolva várias atividades, mas ainda apresenta muitas falhas na extração de sintagmas nominais.

Palavras-chave: Sintagmas Nominais. Recuperação de Informação. Indexação Automática. Extração Automática de Sintagmas Nominais. OGMA e OGMA Web.

ABSTRACT

This work summarizes the theoretical foundations of automatic indexing through the identification, extraction and automatic selection of noun phrases in texts written in Portuguese. It conceptualizes noun phrases, Natural Language Processing, automatic indexing, automatic identification and automatic extraction of noun phrases, as well as the tools to develop these activities and, finally, selection of noun phrases as document descriptors. It evaluates and compares the automatic extraction of noun phrases in the tools OGMA and OGMA Web, using as reference the manual extraction of noun phrases. In comparing these automatic tools, it realizes that despite the OGMA Web have had better performance in some aspects such as extracting a larger number of noun phrases than OGMA, this still manages to be better at fewer errors. Then it suggests some possible solutions for the problems of extracting noun phrases faced by automatic tools. It concludes that the OGMA Web has a better interface, which allows the user to develop various activities, but still has many flaws in the extraction of noun phrases.

Keywords: Noun Phrases. Information Retrieval. Automatic Indexing. Automatic Extraction of Noun Phrases. OGMA e OGMA Web.

LISTA DE FIGURAS

LISTA DE GRÁFICOS

LISTA DE QUADROS

LISTA DE TABELAS

LISTA DE SIGLAS

BDTD-UFPE – Biblioteca Digital de Teses e Dissertações da Universidade Federal de Pernambuco
BOW – *Bag-of-Words*
CI – Ciência da Informação
DET – Determinante
MOD – Modificador
N – Nome
PHP – Hypertext Processor
PLN – Processamento de Linguagem Natural
RI – Recuperação da Informação (*Information Retrieval)*
SA – Sintagma Adjetival
SAdv – Sintagma Adverbial
SISNOP - Sistema Identificador de Sintagmas Nominais do Português
SN – Sintagmas Nominais
SP – Sintagma Preposicional
SRI – Sistema de Recuperação da Informação
SV – Sintagma Verbal
TBL – Aprendizado Baseado em Transformações
TCC – Trabalho de Conclusão de Curso
VISL – *Visual Interactive Syntax Learning*

SUMÁRIO

1 INTRODUÇÃO

Os estudos sobre sintagmas nominais (SNs) têm propiciado o desenvolvimento de novas ferramentas de análise de texto, voltadas para a extração de descritores de documentos digitais a fim de que haja a recuperação satisfatória dos documentos em Sistemas de Recuperação de Informação. Os SNs são considerados melhores descritores que as palavras isoladas, pois a recuperação da informação (RI) através das palavras isoladas está sujeita aos fenômenos da linguagem natural como, por exemplo, a sinonímia e polissemia das mesmas (KURAMOTO, 2002).

A sociedade vive um apogeu de grandes transformações tecnológicas e científicas. O usuário tem uma vasta extensão de documentos a sua disposição, mas recuperar e selecionar os relevantes a uma necessidade de informação não é fácil, pois, segundo Silva e Tomaél (2007), as habilidades de capitalização das informações não desenvolvidas pelas as instituições que detêm ou têm acesso às informações têm como causa a organização da informação como um recurso ainda inacessível.

Assim, a atividade de gestão de informação visa estabelecer, segundo as autoras, uma estratégia que maximiza recursos que permitam um melhor compartilhamento de informações entre as pessoas por meio de suas atividades e produção. Dentro do exercício de trabalho do gestor da informação no plano organizacional, têm-se, de acordo com Berbe (2005)[1]: 1 - identificação das necessidades de informação; 2 - coleta/entrada de informação; 3 - classificação e armazenamento da informação; 4 - tratamento e apresentação da informação; 5 - desenvolvimento de produtos e serviços de informação.

Dessa forma, o presente Trabalho de Conclusão de Curso (TCC) dá ênfase na atividade de tratamento e apresentação da informação, pois os documentos para serem recuperados, precisam passar por um processo de tratamento que é a organização feita por meio da representação, que por sua vez é constituída pela indexação, que se forem feitos adequadamente, aumentam as chances do usuário se sentir satisfeito com o resultado da busca. Tradicionalmente, a indexação era feita com palavras isoladas (SOUZA, 2005), o que poderia ser afetada pelos fenômenos já citados (sinonímia e polissemia), sendo assim, o referido TCC vem acrescentar à ideia de que os SNs são melhores descritores documentais do que uso das palavras isoladas.

[1] Apesar de o autor focar na gestão da informação feita por bibliotecários, a opção por referenciar os elementos por ele elencados se faz pela visão generalista que essas atividades são conceituadas, podendo ser aplicadas ao contexto do gestor da informação de organizações diferentes das bibliotecas.

Silva (2014), objetivando fazer o levantamento do estado da arte sobre a indexação automática por meio de SNs em língua portuguesa, descreve os trabalhos de quinze pesquisadores no intervalo de tempo que vai de 1995 com Kuramoto a 2011 com o trabalho de Corrêa et al (2011) e compara os trabalhos desses quinze pesquisadores quanto à: metodologia de extração de SN; metodologia de seleção de SN; e metodologia de avaliação de extração de SN.

Silva (2014) também faz a comparação desses quanto aos sistemas propostos pelos autores e os sistemas usados como etiquetadores, *parsers* e extratores de SNs. O autor ainda 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. Na comparação dessas ferramentas, Silva (2014) percebe 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.

Seguindo a linha de avaliar *softwares* de extração de SNs, Corrêa et al (2011) também submeteram resumos de teses e dissertações da BDTD-UFPE (Biblioteca Digital de Teses e Dissertações da Universidade Federal de Pernambuco) ao OGMA no intuito de verificar 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.

O OGMA é a única ferramenta disponível que faz a extração automática de SNs em textos escritos em língua portuguesa, pontuando-os a fim de se ter uma lista de melhores descritores de um documento. Ela foi desenvolvida por Maia (2008) em seu doutoramento na Escola de Ciência da Informação da Universidade Federal de Minas Gerais.

Acreditando no potencial do OGMA, Chaves (2013), sob a orientação de Luiz Maia, faz uma adaptação dessa ferramenta para plataforma web, sendo nomeada de OGMA Web. Objetivo de Chaves (2013) é solucionar a seguinte questão por meio de realização de experimentos “De que maneiras o agrupamento de documentos eletrônicos utilizando os sintagmas nominais pode ser aplicado de forma eficiente em um conjunto

de documentos?" (CHAVES, 2013, p. 7). Para tanto, esse autor aprimorou os recursos envolvidos com o agrupamento de documentos disponíveis na ferramenta OGMA.

Diante das modificações feitas no OGMA e agora disponíveis no OGMA Web, o presente TCC tem como **objetivo geral** avaliar a evolução do *software* OGMA para o OGMA Web, comparando-os e apontando as melhorias feitas na ferramenta por Chaves (2013) e os problemas ainda existentes no OGMA Web para a indexação automática por meio de SNs em textos escritos em Língua Portuguesa.

Os **objetivos específicos** são:

- Arrolar os processos de identificação, extração e seleção automática de SNs em textos escritos em Língua Portuguesa;
- Analisar a extração automática de SNs em textos escritos em Língua Portuguesa feita pelo o OGMA e o OGMA Web tendo como referência a extração manual.

A **justificativa** dessa pesquisa se baseia na ideia de que ao analisar as ferramentas citadas, busca-se contribui no aperfeiçoamento das mesmas, pois se apontam as falhas e os acertos desses softwares. Nas suas melhores performances, essas ferramentas podem possibilitar uma reflexão sobre a prática de descrever os documentos de uma BDTD, por exemplo, mostrando que o uso dos SNs na descrição dos documentos evita os problemas como sinonímia, polissemia e ambiguidade decorrentes do uso de palavras isoladas, como foram apontadas no início dessa introdução.

A estrutura desse trabalho é constituída de cinco capítulos. Sendo o primeiro a **introdução**, onde há apresentação do assunto, o contexto da temática, os objetivos desse TCC, a justificativa de trabalhar a indexação automática por meio dos SNs e a organização do trabalho.

O segundo capítulo é composto pela **revisão de literatura**, em que são trabalhados os conceitos de: SNs; Processamento de Linguagem Natural; indexação automática; identificação e extração automáticas de SNs; ferramentas de identificação e/ou extração de SNs; e seleção de SNs como descritores documentais. No terceiro capítulo, é descrito **os aspectos metodológicos** no levantamento bibliográfico e na realização do experimento. No quarto capítulo, são apresentados os **resultados** do experimento seguido de **uma análise comparativa**.

As considerações finais estão no quinto capítulo, no qual são retomados os principais pontos da pesquisa com indicações para a realização de trabalhos futuros.

Logo em seguida, estão **as referências** que foram consultadas para a confecção desse TCC e **os apêndices** produtos do experimento.

2 REVISÃO DE LITERATURA

Essa revisão de literatura é feita a partir de visitas feitas a conceitos como **Sintagmas nominais** que são os objetos trabalhados pelas ferramentas que aqui serão avaliadas e comparadas, **Processamento de Linguagem Natural** que é uma área da linguística computacional que, por sua vez, é composta pela junção da Ciência das Linguagens e Ciência da Computação, **Indexação automática** que é uma atividade desenvolvida tanto pelos cientistas da computação quanto os cientistas da Ciência da Informação, **Identificação e extração automáticas de sintagmas nominais**, assim como as **ferramentas** que desenvolvam essas últimas atividades e, por fim, **seleção de sintagmas nominais como descritores documentais**.

2.1 Sintagmas Nominais

Silva e Koch (2009) definem sintagma como um grupo de elementos que constituem entre si uma unidade significativa dentro da oração em consonância com relações de dependência e ordem estabelecidas dentro desse conjunto. Os sintagmas têm uma organização em que suas partes se estruturam em torno de um elemento fundamental, que é chamado núcleo.

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.

Sintagmas nominais é uma parte da teoria linguística denominada Gramática Gerativa proposta por Chomsky (1965) e por estudiosos de Massachusetts nos Estados Unidos, na década de sessenta. Em 1957, Noam Chomsky publica o livro *Estruturas Sintáticas*[2] 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”.

---

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

Essa teoria dá conta das novas frases que não eram analisadas pelo modelo linguístico tradicional, pois como a língua é um sistema que está sempre em evolução, é impossível para a abordagem tradicional com suas regras compreender e absorver o sentido das frases inéditas. 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).

Segundo Kenedy (2008, p.04):

> Os gerativistas perceberam que as infinitas sentenças de uma língua eram formadas a partir da aplicação de um finito 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.

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

A Gramática Gerativa é um conjunto de teorias que se preocupam com o dispositivo mental inato, esse é responsável pela produção linguística do falante. Dentre os modelos gerativos mais conhecidos estão os dos estados finitos, o transformacional e o sintagmático (MAIA, 2008). Esse último é objeto dessa pesquisa aplicado à indexação automática.

Os sintagmas são classificados como essenciais e facultativos. Os essenciais são tidos como elementos básicos de uma oração[3], 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 preposicional (SP) que é composto por um núcleo chamado preposição e o sintagma adverbial (SAdv) cujo núcleo é um advérbio ou uma locução adverbial.

Perini et al (1996) definem o SN como classe gramatical, que tem o comportamento de sujeito, de objeto direto e, quando precedido de preposição, de adjunto adnominal ou de objeto indireto. Liberato (1997) diz que o SN é a parte do enunciado que representa conceitos ou referentes.

Na Recuperação da Informação, os sintagmas nominais são vistos como descritores no processo de classificação e recuperação da informação, tendo sua importância acentuada por Perini et al (1996) quando dizem ser os sintagmas nominais mais eficientes no processo classificar e recuperar.

O uso do SN na representação, no armazenamento, na organização e no acesso aos itens informativos se faz por meio de expressões nominais, não só com palavras isoladas (que por meio desse método se depara com as dificuldades impostas pela sinonímia e polissemia).

O SN geralmente traz o tema do enunciado, daí Kumaroto (1995) afirmar que os SNs são indexadores mais promissores de um texto. E a importância dos SNs na recuperação da informação é quando eles são extraídos dos textos, eles mantêm o mesmo significado. Há duas alternativas possíveis ‘de implementação em termos de indexação automática e de interfaces de busca’, utilizando os SNS, segundo 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.

Percebem-se os SNs como elementos identificadores de unidades de informação com alto poder discriminatório.

[3] 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.

O SN é definido como a menor unidade de sentido do discurso e que possui uma estrutura sintática e lógico-semântica (KURAMOTO, 1999).

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 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).

Morellato (2007) conceitua o SN como a unidade sintático-semântica cujo núcleo é constituído por um nome ou pronome.

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 documentos

- **Nível 3**: A extração de sintagma nominais em documentos
- **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 1, 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 1 – 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 um 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.

2.2 Processamento de Linguagem Natural

O processamento de linguagem natural (PLN) pode ser definido como a habilidade de um computador em processar a linguagem humana. O PLN visa, também, produzir ferramentas que compreendam a língua.

Para Barros e Robin (2001), o PLN tem por objetivo interpretar e gerar textos em linguagens naturais. É uma área de pesquisa 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, em um ou mais níveis de entendimento e/ou geração de sons, palavras, sentenças e discurso”.

2.2.1 Etiquetador Morfossintático (*Tagger*)

Dentro das estratégias do PLN para enfrentar a ambiguidade de termos e buscar os significados dos léxicos, 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".

2.2.2 Analisador Sintático (*Parser*)

O analisador sintático avalia o agrupamento das palavras, trabalhando a estrutura da frase e é denominado de *parser*. 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) aponta 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. 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".

2.3 Indexação automática

Para tornar a RI de um documento um processo mais simples e eficaz, é necessário representar esse documento por termos que sejam 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 do conteúdo de um determinado documento, levando-se em conta as necessidades informacionais da clientela específica.

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 reduzir os termos incluídos na indexação; e a especificidade que se refere à utilização de 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[4] e a diminuição da precisão[5] na RI, já se for aumentada a especificidade, haverá o aumento da precisão e a diminuição da revocação.

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.

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. Enquanto Hjørland (2008) percebe a indexação automática como um procedimento feito por meio de algoritmos. Já 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.

---

[4] 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.

[5] 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.

2.4 Identificação e Extração automáticas de SNs

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.

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 Sistemas de Recuperação de Informação (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.

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.

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.

2.5 Ferramentas de Identificação e/ou Extração

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.

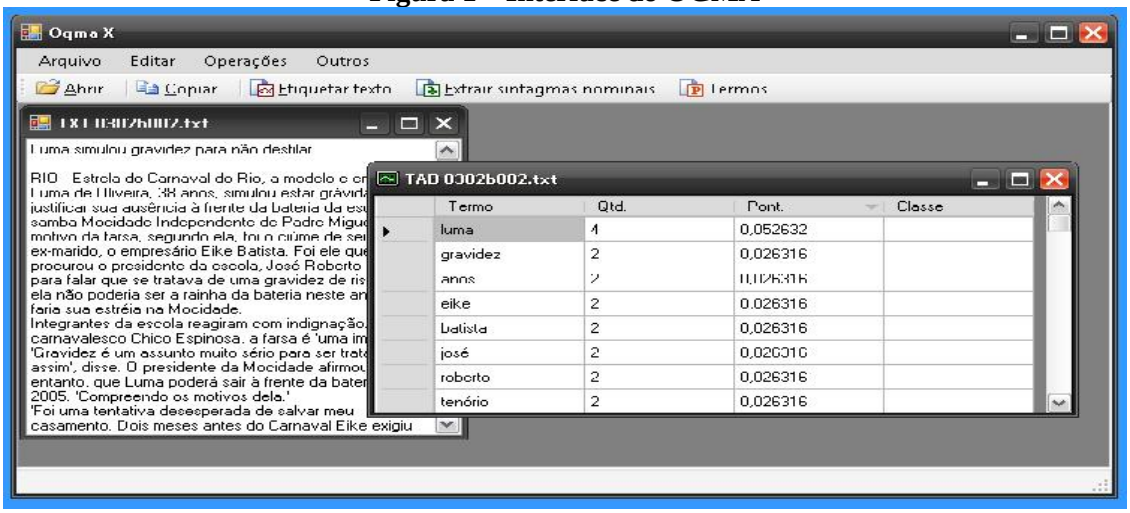
Um exemplo de identificador de SNs é o Sistema Identificador de Sintagmas Nominais do Português (SISNOP) de Morellato (2010). Essa ferramenta é compreendida como um extrator de SN com um conjunto de módulos ou de programas que interpretam textos através de análises morfológicas e sintáticas, permitindo 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.

2.5.1 OGMA tradicional

O OGMA, ferramenta elaborada e criada por Maia (2008), faz 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 1 traz a representação da interface dessa ferramenta.

Figura 1 – Interface do OGMA



Fonte: OGMA. Disponível em: <http://www.luizmaia.com.br/ogma/>. Acesso em: 15 jul. 2014.

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.

Na construção do léxico de língua portuguesa para o OGMA, foram adaptados arquivos com vocabulário utilizado pelo BR/ISPELL. Dessa forma, tem-se um arquivo de dados com uma tabela de 41978 substantivos e adjetivos, 292720 verbos. Para compor a lista de *stopwords*, foram digitalizadas 475 palavras de diferentes classes gramaticais da gramática de Tufano (1990, apud MAIA, 2008). O léxico é utilizado para etiquetar cada palavra do texto com as possíveis classes gramaticais correspondentes.

No intuito de evitar as ambiguidades, o OGMA cria uma lista com todas as combinações de etiquetas identificadas para vocábulos de uma frase e submete cada combinação às regras para extração dos sintagmas nominais (CORREA et al 2011).

No Quadro 1, está o conjunto de regras das quais o OGMA lança mão para extrair os SNs,

> Aplicando regra por regra na ordem de leitura até obter um sintagma nominal cujo símbolo é SN. Estas regras atuam sobre as etiquetas (que representam as classes gramaticais) atribuídas às palavras, visando marcar o início e fim do sintagma em cada sentença do texto. (CORREA et al, 2011, p.17)

Quadro 1 - Regras de extração de Sintagmas Nominais do software 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).

Para cada sentença do texto, com diferentes possíveis combinações de etiquetas para as palavras que a compõem, os SNs são identificados levando em conta cada combinação, todos os SNs identificados entram em uma lista geral na qual os sintagmas duplicados são eliminados. Dessa forma a ambiguidade é contornada resolvida eficientemente.

2.5.2 OGMA Web

O OGMA Web (CHAVES, 2013) [6] é uma adaptação da ferramenta OGMA tradicional (MAIA, 2008) para a plataforma *web*, mas com algumas modificações na funcionalidade. O OGMA Web foi desenvolvido em linguagem PHP (Hypertext Processor). Diferentemente do OGMA tradicional, o da plataforma *web* tem a funcionalidade de identificação dos usuários, no intuito de que acervos personalizados sejam disponibilizados para cada usuário cadastrado na ferramenta, o que possibilita a livre construção de bases de documentos eletrônicos para análises automatizadas. A Figura 2 mostra o funcionamento do OGMA Web.

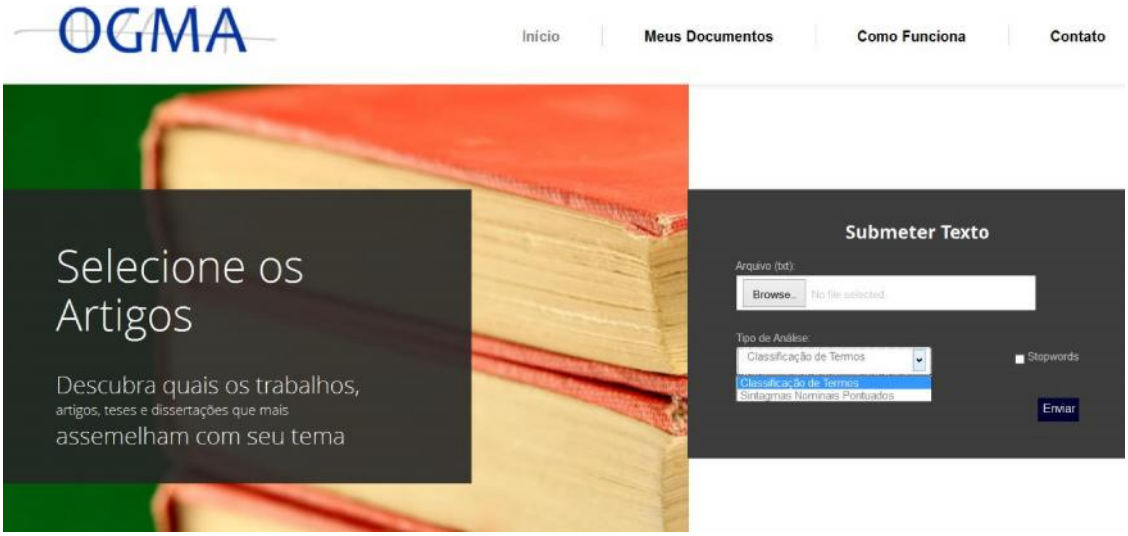
[6] Disponível em: <http://ogmaweb.com.br/ogma/?p=Inicio> Acesso em 10 jul. 2014.

Figura 2 - Funcionamento do OGMA Web

Fonte: Disponível em: <http://ogmaweb.com.br/ogma/?p=Inicio> Acesso em 10 jul. 2014.

As funções do OGMA que foram mantidas no OGMA Web consistem na extração de termos e na atribuição de pesos vinculadas à frequência de aparição no texto; na consideração à lista de *stopwords*; na extração dos sintagmas nominais, sintagmas nominais únicos e pontuados, no cálculo de similaridade e no método de etiquetação.

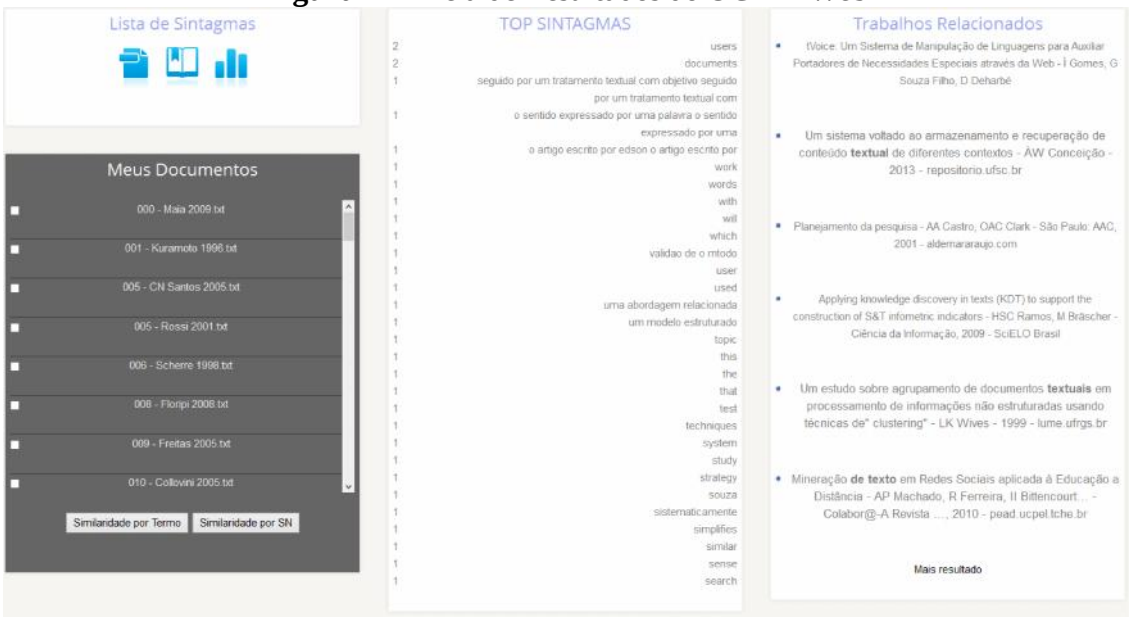
Figura 3 - Tela de Inserção de Documentos do OGMA Web



Fonte: Disponível em: <http://ogmaweb.com.br/ogma/?p=Inicio> Acesso em 10 jul. 2014.

Depois que usuário faz a submissão de um documento de texto verbal à ferramenta, o OGMA Web apresenta três colunas de resultados: a primeira coluna é composta por três ícones que levam a três documentos resultantes da análise do texto submetido, tendo o texto etiquetado, os sintagmas nominais, ou termos de acordo com a seleção do tipo de análise realizada na tela de inserção e a classificação destes sintagmas ou termos levando em consideração a ponderação adotada e o método de extração selecionado pelo usuário. Também são exibidos os documentos já inseridos pelo usuário, para a realização de cálculos de similaridade entre os documentos (CHAVES, 2013); a segunda coluna exibe os 30 sintagmas nominais ou termos que apresentaram maior relevância no texto, já em ordem classificatória, com o número de ocorrências ao lado; já a terceira coluna faz a relação dos resultados encontrados com uma consulta no Google Scholar por meio de uma busca utilizando os três primeiros sintagmas ou termos encontrados, tendo um link chamado “Mais Resultados” que direciona a página do Google que disponibilizou os dados. A Figura 4 apresenta a interface de resultados do OGMA Web.

Figura 4 - - Tela de Resultados do OGMA Web



Fonte: Disponível em: <http://ogmaweb.com.br/ogma/?p=Inicio> Acesso em 10 jul. 2014.

Segundo Chaves (2013), com a seleção de dois documentos já submetidos pelo usuário, o OGMA Web compara o cosseno de similaridade, que representa o ângulo entre dois vetores em um espaço vetorial e retorna um valor entre 0 e 1. Em que 0 significa que nenhuma relação foi encontrada entre estes documentos, e 1 para documentos com total similaridade.

> O banco de dados foi convertido de Microsoft Access para MySQL e novas mudanças ocorreram na estruturação dos dados. Foram criadas associações entre os usuários e seus documentos para que os acervos sejam personalizados. Também foram armazenados os termos do documento processado para uma futura análise de similaridade, o que reduz o tempo de resposta da ferramenta (CHAVES, 2013, p. 16)

Para validar o OGMA Web, Chaves (2013) comparou os dados retornados tanto pelo o OGMA tradicional como o OGMA web. O autor se baseou nos experimentos de Maia (2008) em que foram verificados os resultados alcançados com a extração automática de SN, realizada pelo OGMA, em comparação com os SN extraídos por outra ferramenta textual, o VISL (*visual interactive syntax learning*).

> (...) 6 textos encontrados no ANEXO I, do estudo de Maia (2008), foram submetidos ao OGMA Web para assegurar a conformidade entre os resultados encontrados pelas ferramentas. Quatro destes textos obtiveram a mesma listagem de SN nas duas plataformas, OGMA e OGMA Web. (CHAVES, 2013, p. 17)

Vale ressaltar que em alguns casos a ordem dos sintagmas encontrados não foi igual, o que não representa risco. O autor argumenta que em um dos arquivos comparados houve uma inconsistência entre os resultados, o que justifica o presente trabalho.

> Na última inconsistência encontrada, ambas ferramentas identificaram o sintagma: “artigos ainda não publicados”. As palavras “ainda” e “não” se juntaram nos SN encontrados nas duas ferramentas. O que pode representar um problema. Em seguida, apenas o OGMA Web identificou os SN: “ainda não publicados”, “projeto internacional”. Como o OGMA Web apresentou resultados mais convincentes, a incoerência entre os resultados aqui descritas não serão levadas a diante. [sic] (CHAVES, 2013, p. 17)

Após a validação da extração dos SNs, o autor realizou o teste de similaridade para assegurar o correto funcionamento da ferramenta web a partir das regras de correlação de Pearson, nas quais os valores obtidos na similaridade entre dois documentos não podem ser maior que 1, nem menor que 0. O cálculo do cosseno de similaridade acontece para medir o ângulo formado entre dois elementos em um espaço

vetorial, com isso documentos semelhantes tendem possuir um fator de similaridade próximo de 1 (CHAVES, 2013).

Quanto à similaridade entre os documentos, os resultados foram de acordo com o esperado pelo o autor, os documentos iguais apresentam 0 como cosseno de similaridade. O autor conclui que documentos semelhantes acerca de um determinado tema, sintagmas nominais pontuados podem não ser a melhor opção para medir a similaridade entre documentos, apesar de apresentarem enorme eficiência na coleta de descritores portadores de informações relevantes.

2.6 Seleção de Sintagmas Nominais como Descritores Documentais

De acordo com LE GUERN (1991, apud KURAMOTO, 1995, p. 03) existe uma diferença entre a **palavra** e os descritores, conforme mostra a citação a seguir:

> Não constitui finalidade do descritor a sua visualização mediante a abstração do valor referencial de suas ocorrências no acervo de documentos. As palavras da língua, enquanto palavras da língua possuem apenas atributos sem qualquer substância, até que elas façam parte do discurso. Quanto ao descritor, ele representa uma entidade segundo a filosofia de Aristóteles. Assim, o descritor não pode ser considerado, a exemplo das palavras da língua, como um símbolo sem referência.

Kuramoto (2002) observa que os SNs são grupos nominais organizados por uma hierárquica em forma de árvore. O autor propõe uma estrutura arbórea de SNs, para que o usuário possa filtrar com mais fidelidade os resultados de sua consulta, e cada vez que o usuário especifica (sobe de nível na estrutura arbórea), mais precisa é a recuperação de documentos, ou seja, os níveis da estrutura vão do mais generalizado ao mais específico (nível 1 ao nível 5), resultando em maior interação do usuário com o sistema e maior retorno de documentos relevantes. Observe o seguinte exemplo:

a) A biblioteca digital como canal de comunicação (nível 3)

   Trata-se de um SN complexo, pois este possui dois outros SNs embutidos:

b) Canal de comunicação (nível 2)
c) Comunicação (nível 1)
d) Biblioteca digital (nível 1)

Os SNs podem ser organizados em formato de árvore, como exposto anteriormente. Assim, observando as sugestões de KURAMOTO (2002), a enumeração do nível é feita atribuindo ao SN mais simples (Comunicação) o nível 1, ao SN que contém “canal de comunicação” o nível 2 e ao SN que contém “A biblioteca digital como canal de comunicação” o nível 3. KURAMOTO (2002) apresenta os SNs como

abordagem alternativa na recuperação da informação. Por meio do protótipo elaborado por Kuramoto (1995), o sistema de recuperação é capaz de navegar em uma estrutura hierárquica formada por árvores de sintagmas nominais.

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).

Mas 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.

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. O processamento da linguagem natural envolve as operações sobre o texto como análises léxicas, 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).

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 (KURAMOTO, 1995).

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 descartar os com ocorrência inferior a um patamar de 1% sobre o total de SN únicos do documento. 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) sugerem que 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.

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.

3 ASPECTO METODOLÓGICO

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.

Nos procedimentos, foi utilizada a pesquisa bibliográfica para trazer alguns conceitos necessários ao entendimento da indexação automática por meio da extração e seleção de SNs.

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. Os mesmos resumos foram usados por Silva (2014). Dessa forma, a escolha desses se deu com o intuito de comparar 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 e pelo OGMA Web, a fim de perceber alguma evolução das ferramentas e apresentar as falhas ainda existentes nessas ferramentas, possibilitando 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. Após a extração feita por essas ferramentas, os SNs foram copiados e colocados em editores de texto e de 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 e de precisão 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.

4 OGMA x OGMA WEB: uma análise comparativa

Esse capítulo tem como intuito fazer uma comparação entre os *softwares* OGMA e o OGMA Web, 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, como já apontado por Silva (2014): 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.

A Tabela 2 foi construída a partir da extração dos SNs pelas referidas ferramentas, bem como pela extração manual. Os dois sistemas fazem a extração de SNs automaticamente, permitindo rapidez nas transferências dos sintagmas para um editor de texto e de planilha. Já o processo de extração manual foi feito a partir da leitura especializada dos resumos.

Sendo assim, o número total de todas as ocorrências de SNs extraídos pelos referidos programas e pela extração manual é apresentado na Tabela 2. O OGMA Web extraiu um número maior de expressões, contudo os SNs extraídos por essa ferramenta são muito vulneráveis a erros, os quais serão expostos logo mais.

Tabela 2 - Total de expressões extraídas dos resumos da BDTD-UFPE pelo OGMA, pelo OGMA Web e pela extração manual

Resumos	OGMA	OGMA WEB	Extração Manual
CC 1	55	70	90
CC 2	66	85	163
CC 3	56	72	120
CC 4	44	69	100
CC 5	97	125	181
CC 6	37	46	54
CC 7	24	38	56
CC 8	37	53	76
CC 9	57	80	131
CC 10	34	38	65
Total CC	507	676	1.036
DI 1	29	49	48
DI 2	25	36	39
DI 3	41	60	75
DI 4	41	56	67
DI 5	31	43	51
DI 6	27	35	69
DI 7	35	65	82
DI 8	24	33	68
DI 9	14	22	43
DI 10	43	53	67
Total DI	310	452	609
NU 1	39	51	74
NU 2	47	50	95
NU 3	44	61	87
NU 4	42	58	77
NU 5	48	71	116
NU 6	45	60	77
NU 7	72	83	98
NU 8	54	79	109
NU 9	35	49	84
NU 10	55	70	107
Total NU	481	632	924
TOTAL	1.298	1.760	2.569
TOTAL GERAL		5.627	
CC –	Ciência da Computação		
DI –	Direito		
NU –	Nutrição		

Fonte: o autor

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[7] 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[8] 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 OGMA Web extraiu 1.760 expressões, desse total, 803 não eram SNs, tendo um total de 957 expressões que de fato constituem SNs. Desse modo, a taxa de acerto dessa ferramenta foi de 54,5% em relação aos seus próprios resultados, enquanto que a taxa de revocação foi de 37%, em comparação com a extração manual.

As dificuldades enfrentadas pelo OGMA Web são em relação à identificação de SA (sintagma adjetival) como SN, a colocação de verbos no meio dos SNs, identifica preposições, artigos, verbos e advérbios isolados como SN. Esse programa extrai expressões que seriam bons SNs como descritores, mas acaba partindo-os ao meio ou iniciando com classes de palavras que os caracterizam como outros tipos de sintagmas. Esses problemas também foram observados no OGMA, como exposto anteriormente.

A Tabela 3 aponta a quantidade de expressões extraídas pelos dois *softwares*, mas que não constituem SNs.

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

[8] 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.

Tabela 3 – Quantidade de expressões identificadas pelo OGMA e pelo OGMA Web, mas que não constituem sintagmas

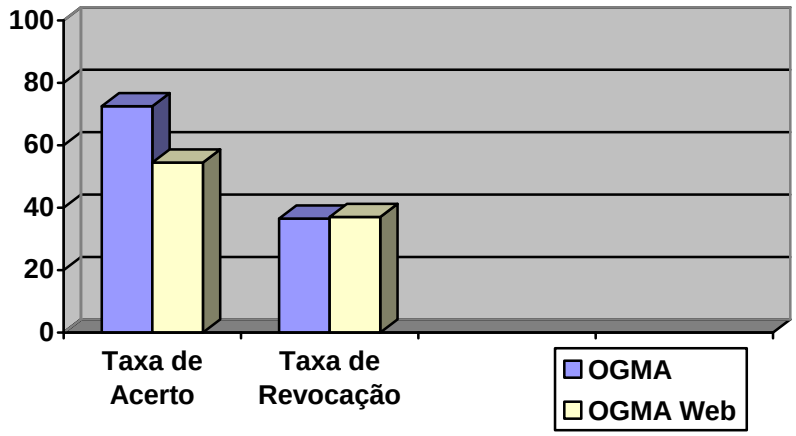
Resumos	OGMA	OGMA WEB
CC 1	14	36
CC 2	19	37
CC 3	13	22
CC 4	15	33
CC 5	21	45
CC 6	12	11
CC 7	5	16
CC 8	5	27
CC 9	12	39
CC 10	5	16
Total CC	121	282
DI 1	11	27
DI 2	5	20
DI 3	11	35
DI 4	17	34
DI 5	10	29
DI 6	6	17
DI 7	11	31
DI 8	8	18
DI 9	5	14
DI 10	17	28
Total DI	101	253
NU 1	9	23
NU 2	12	12
NU 3	8	30
NU 4	13	27
NU 5	11	38
NU 6	17	31
NU 7	22	20
NU 8	19	34
NU 9	9	24
NU 10	15	29
Total NU	135	268
TOTAL	357	803
TOTAL GERAL	1.160	
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 revocação quando comparada com os resultados da extração manual.

Gráfico 1 – Níveis de acerto e revocação na extração de expressões que constituem SNs



Fonte: o autor

Os percentuais de acerto atingidos pelo OGMA demonstram certa vantagem em relação ao OGMA Web. Já em relação à taxa de revocação, o OGMA Web tem apenas uma taxa de 0,5% a mais que o OGMA tradicional. Uma vantagem que as duas ferramentas apresentam é que não é necessária a submissão sentença por sentença, de maneira que o usuário faz a submissão com o texto completo. Além de identificar os SNs, o OGMA e o OGMA Web os extrai, já pontuando de acordo com um índice de relevância, o que facilitaria o trabalho do indexador, diferentemente de outras ferramentas existentes como o parser LX-Parser, que a submissão é feita por sentenças e não por texto completo de única vez.

Outro aspecto observado nessa análise foi a quantidade de vezes em que os programas e a extração manual 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 extratores dessas expressões nominais. Assim, na Tabela 4, há a quantidade de vezes em que cada programa e a extração manual fizeram a extração de SNs semelhantes 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 4 – Quantidade de vezes que o OGMA, o OGMA Web e a extração manual recuperam SNs semelhantes às palavras-chave dos resumos da BDTD-UFPE

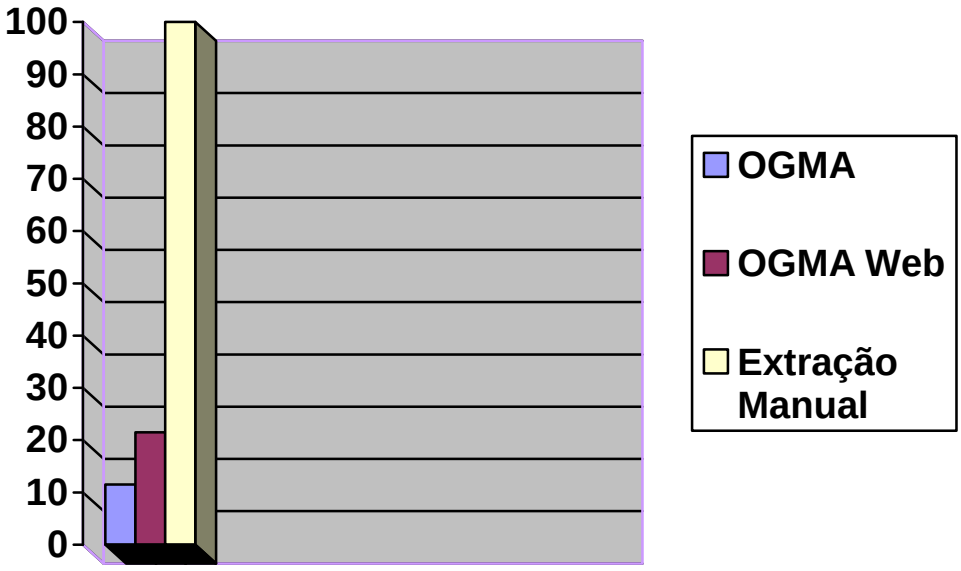
Resumos	OGMA	OGMA WEB	Extração Manual
CC 1	2	2	2
CC 2	-	1	5
CC 3	-	-	7
CC 4	-	1	1
CC 5	1	-	12
CC 6	-	-	4
CC 7	1	-	8
CC 8	1	2	1
CC 9	1	3	8
CC 10	-	2	7
Total CC	6	11	55
DI 1	1	2	6
DI 2	-	-	3
DI 3	-	-	-
DI 4	1	-	5
DI 5	1	1	4
DI 6	-	2	2
DI 7	-	1	-
DI 8	1	-	4
DI 9	1	-	6
DI 10	-	1	3
Total DI	5	7	33
NU 1	1	1	5
NU 2	1	1	8
NU 3	-	2	2
NU 4	-	3	5
NU 5	-	-	8
NU 6	-	-	8
NU 7	2	1	5
NU 8	3	7	12
NU 9	-	1	15
NU 10	1	2	11
Total NU	8	18	79
TOTAL	19	36	167
TOTAL GERAL	222		
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

Ambas as ferramentas ficaram distante da extração manual (167 SNs semelhantes às palavras-chave). O OGMA conseguiu atingir a taxa de 11,5% de revocação. Já o OGMA Web alcançou a taxa de 21,5% de revocação. Nesse aspecto, o OGMA Web se saiu melhor, com quase o dobro da taxa do OGMA tradicional.

Os SNs que continham as palavras-chave dentro de suas partes constituintes foram também observados, tendo assim a Tabela 5.

Tabela 5 – Quantidade de SNs que continham palavras-chave dos resumos da BDTD-UFPE

Resumos	OGMA	OGMA WEB	Extração Manual
CC 1	7	8	24
CC 2	3	6	15
CC 3	9	8	16
CC 4	10	5	15
CC 5	8	5	22
CC 6	-	2	7
CC 7	6	9	15
CC 8	4	1	5
CC 9	3	3	12
CC 10	3	3	11
Total CC	53	50	142
DI 1	1	3	6
DI 2	2	2	-
DI 3	1	-	2
DI 4	2	4	6
DI 5	1	2	3
DI 6	9	6	23
DI 7	2	4	10
DI 8	5	6	12
DI 9	1	3	14
DI 10	5	7	3
Total DI	29	37	79
NU 1	2	1	11
NU 2	1	4	18
NU 3	8	4	18
NU 4	6	4	7
NU 5	2	6	7
NU 6	5	4	17
NU 7	6	3	14
NU 8	7	7	24
NU 9	9	9	26
NU 10	3	2	24
Total NU	49	44	166
TOTAL	131	131	387
TOTAL GERAL	649		
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 6 apresenta a quantidade de SNs relevantes para a descrição dos resumos.

Tabela 6 – Quantidade de SNs relevantes para a descrição dos resumos da BDTD-UFPE

Tipos de SNs extraídos	OGMA	OGMA WEB	Extração Manual
Semelhantes às palavras-chave dos resumos	19	36	167
Contém palavras-chave dos resumos	131	131	387
TOTAL	150	167	554

Fonte: o autor

A quantidade de SNs relevantes identificados pela extração manual atinge uma taxa de precisão de 20% 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 29,5% na taxa de revocação.

O OGMA Web atinge uma taxa de precisão de 17,5%, se forem tomados como referência os seus 957 SNs de fato identificados, mas tomando como referência a extração manual, o OGMA Web alcança 33% de revocação. 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, apesar do OGMA Web conseguiu extrair 16 SNs a mais que o OGMA tradicional, a taxa de erro daquele foi de 45,5% na extração de SNs, enquanto que o OGMA tradicional atingiu a taxa de 27% de erro na extração de SNs.

Em relação à quantidade de SNs relevantes, a diferença entre as duas ferramentas é de apenas 17 SNs que estão na categoria de SNs semelhantes às palavras-chave do resumo. Todavia, na outra categoria em SNs contém palavras-chave do resumo, não há diferença na quantidade.

Diante dos problemas apresentados pelas ferramentas automáticas, algumas sugestões podem ser dadas no intuito de que os programas 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 pelos dois *softwares* em uma quantidade proporcional aos SNs identificados por cada um.

Silva (2014) apontou que as ferramentas de identificação e/ou extração de SNs cometem erros 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. Os verbos quando não fazem parte de um complemento nominal não podem ser considerados SN ou parte desse SN, já que esse acontece quando as palavras giram em torno de um substantivo e não em torno de um verbo, o que é um SV (sintagma verbal). Desse modo, percebeu-se que as duas ferramentas cometeram esses erros, mas com destaque para o OGMA Web que teve muitas ocorrências na identificação de verbos propostos ao pronome relativo “que” como SNs.

Segundo Silva (2014), 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. Também foram identificados problemas relacionados às preposições “para” e “de”, em que as expressões extraídas começavam com uma preposição, o que caracteriza um SP (sintagma preposicionado). A ocorrência com a preposição “de” teve maior frequência nas duas ferramentas, principalmente quando estavam diante de um pronome demonstrativo como “esse”, ocorrendo também, mas com menor frequência, quando estavam diante de um verbo.

A solução para o problema de se ter um verbo proposto a uma preposição, exercendo a função de complemento nominal, 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. Por exemplo: “necessidade de extrair sintagmas nominais” ficaria “necessidade de extração de sintagmas nominais”. Essa modificação ocorreria quando os SNs fossem extraídos do texto e estivessem prontos para serem termos de indexação.

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.

5 CONSIDERAÇÕES FINAIS

Quanto aos objetivos desse TCC, acredita-se que eles tenham sido atingidos, pois com a comparação entre as ferramentas, viu-se que o OGMA tradicional tem certa vantagem sobre o OGMA Web. Mesmo extraindo 462 expressões a mais que o OGMA (o que representa 16% a mais), o OGMA Web tem uma taxa de acerto de 54,5%, enquanto que a outra ferramenta tem 72,5%. Na taxa de revocação na extração de SNs, a diferença entre os dois sistemas é de apenas 0,5 ponto percentual, o OGMA com 36,5%, enquanto o OGMA Web com 37%. Quanto à extração de SNs relevantes (SNs semelhantes às palavras-chave dos resumos mais SNs que continham as palavras-chave dos resumos em suas estruturas), o OGMA Web teve o desempenho melhor, pois sua taxa de precisão foi 17,5% enquanto que a do OGMA tradicional foi de 16% e em relação à taxa de revocação desses SNs relevantes, O OGMA Web teve a taxa de 33% enquanto o OGMA tradicional teve 29,5%.

Os problemas existentes nas expressões extraídas pelas ferramentas estão relacionados à estrutura, tendo uma grande incidência de verbos nessas expressões, o que descaracteriza ser um SN. Muitas vezes os SP (sintagmas preposicionados) eram tidos como SN.

Houve uma frustração por parte da autoria desse TCC em relação aos resultados alcançados pelo OGMA Web, pois com as adaptações que Chaves (2013) fez no OGMA tradicional, pensou-se que os resultados na extração de SNs seriam melhor em relação à taxa de acerto e revocação, o que não aconteceu. Pode-se dizer que os resultados do OGMA tradicional foram melhores do que o da plataforma *web*. Mas vale ressaltar que em alguns momentos o OGMA Web conseguiu extrair SNs complexos, que são aqueles que têm uma estrutura composta por muitas preposições, modificadores e quantificadores.

Em relação à interface e as funcionalidades da ferramenta, acredita-se no potencial do OGMA Web que depois de alguns aperfeiçoamentos pode se tornar igual ou melhor que o Parser PALAVRAS, considerado por Silva (2014) e demais autores como melhor identificador de SNs em textos escritos em língua portuguesa.

Para trabalhos futuros, tem-se a descrição de documentos por meio de SNs para pessoas com deficiência auditiva, já que, segundo a literatura especializada, esses usuários têm dificuldades na escrita da língua portuguesa, principalmente com o uso de verbos. Os SNs também podem ser trabalhados na confecção de mapas conceituais, os

quais poderiam ser usados como alternativa na representação de documentos, já que o usuário, ao ler os descritores de um texto, poderia ter a possibilidade de ver informações sobre o conteúdo por meio de um mapa conceitual, o que possibilitaria confirmar se aquele documento é relevante para a sua busca. I

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, pois 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.

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.

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

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

BERBE, Alexandre Campos. **Gestão da Informação e do Conhecimento**: reflexão de conceitos e o papel da biblioteconomia. 2005. 102f. Trabalho de Conclusão de Curso – TCC (Graduação em Biblioteconomia) – Escola de Comunicação e Artes, Universidade de São Paulo – USP, São Paulo, 2005.

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.

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.

CHAVES, Rodrigo Soares. **Agrupamento de documentos eletrônicos por meio de Sintagmas Nominais**. 2013. Projeto de Dissertação (Mestrado Profissional em Sistemas da Informação e Gestão do Conhecimento) – Universidade Fundação Mineira de Educação e Cultura. Belo Horizonte, 2013.

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

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

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.

Dicionário BR/ISPELL. Disponível em: < http://www.ime.usp.br/~ueda/br.ispell/>. Acesso em: 02 abr. 2014.

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

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.

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. 2014.

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.

_____. **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.

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.

LX-Parser. Disponível em: <http://lxcenter.di.fc.ul.pt/tools/pt/conteudo/LX-Parser.html>. Acesso em 17 jun. 2014.

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.

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

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.

NÚCLEO INTERINSTITUCIONAL DE LINGUÍSTICA COMPUTACIONAL. **Curupira**. Disponível em: <http://www.nilc.icmc.usp.br/nilc/tools/curupira .html>. Acesso em: 12 jul. 2014.

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

OGMA Web. Disponível em: <http://ogmaweb.com.br/ogma/?p=Inicio> Acesso em 10 jul. 2014.

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.

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, 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.

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.

SILVA, Maria Cecília Pérez de Souza e; KOCH, Ingedore Grunfeld Villaça. **Linguística aplicada ao português**: sintaxe. 15.ed. São Paulo: Cortez, 2009.

SILVA, Terezinha Elisabeth da.; TOMAÉL, Maria Inês. Editorial: A gestão da informação nas organizações. **Inf . Inf .** , Londrina , v. 12, n. 2, j u l./d e z. 2007.

SILVA, Tiago José da Silva. Indexação automática por meio da extração e seleção de sintagmas nominais em textos em língua portuguesa 2014. 219 f. Dissertação (Mestrado) – Departamento de Ciência da Informação, Universidade Federal de Pernambuco – UFPE, Recife, 2014.

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.

UNIVERSIDADE FEDERAL DE PERNAMBUCO – UFPE. BIBLIOTECA DIGITAL

DE TESES E DISSERTAÇÕES – BDTD. Disponível em: <http://www.bdtd.ufpe.br>. Acesso em: 20 set. 2012

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. 2014.

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 OGMA e pelo OGMA Web.

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 OGMA

PALAVRAS 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

APÊNDICE C – Extração pelo OGMA Web

PALAVRAS DI1

1. capacidade
2. contributiva nas contribuições
3. à
4. previdência social direitos fundamentais do cidadão
5. contribuinte e justiça fiscal
6. a ser denominado
7. estado moderno
8. vivenciou
9. transformações
10. a fiscalidade
11. fenômeno essencial
12. existência do estado também tem
13. sofrido mudanças
14. dentre
15. se destaca
16. o grande aumento de contribuições
17. percentualmente em relação
18. receita estatal tributária que devido
19. razões
20. passam
21. a substituir
22. gradativamente os impostos diretos diante
23. dessa nova perspectiva do fenômeno
24. tributário torna
25. se fundamental adequar
26. o contribuinte
27. o direito à
28. existência materializado
29. um desses direitos fundamentais que deve
30. ser inexoravelmente respeitado
31. seu aspecto ampliado
32. justamente a condição humana
33. tal qual descreve
34. hannah
35. arendt
36. surge
37. assim a necessidade de harmonizar
38. as contribuições previdenciárias do cidadão
39. a o mínimo existencial
40. a aplicação do princípio da capacidade
41. contributiva
42. mediado
43. o seus subprincípios clássicos
44. como a progressividade
45. a proporcionalidade
46. o estado desde quando passou

47. denominado estado
48. devido a inúmeras
49. os direitos fundamentais do contribuinte