em questão
E-ISSN 1808-5245

# Método de normalização de sintagmas nominais na indexação automática

**Renato Fernandes Corrêa**

Doutor; Universidade Federal de Pernambuco, Recife, PE, Brasil;
renato.correa@ufpe.br

**Victor Galvão Celerino**

Mestre; Universidade Federal de Pernambuco, Recife, PE, Brasil;
victorgalvao@gmail.com

**Resumo:** Propõe e avalia um método de normalização de sintagmas nominais em termos canônicos, que visa contribuir para a melhora qualitativa da indexação automática, evitando a dispersão terminológica e preservando as palavras-chave dos autores, presentes no interior dos sintagmas nominais. A pesquisa é exploratória e empírica, pautada em pesquisa bibliográfica e realização de um experimento em um *corpus* de artigos científicos da área de Ciência da Informação. O método proposto é constituído por regras e critérios de normalização que obedecem às restrições de preservação da estrutura válida do sintagma nominal e das palavras-chave. O método proposto é avaliado através da presença de termos do Tesauro Brasileiro em Ciência da Informação (TBCI) nos sintagmas nominais resultantes da aplicação das regras e critérios. O método consiste em duas etapas: a primeira é composta por 85 regras para minimizar os sintagmas nominais extensos; a segunda etapa é composta por sete critérios responsáveis por eliminar dos sintagmas nominais elementos gramaticais desnecessários em sua estrutura. Os resultados da avaliação indicam que o método de normalização permite o alcance de resultados positivos, mesmo com dois critérios da segunda etapa não apresentando resultados para o *corpus* utilizado. Conclui-se que a aplicação do método de normalização em sistema de indexação automática é viável e traz bons resultados.

**Palavras-chave:** Indexação automática. Sintagmas nominais. Normalização de sintagmas nominais. Palavras-chave. Tesauro.

## 1 Introdução

Atualmente, o crescimento exponencial do volume de informações publicadas tem dificultado a indexação e catalogação de documentos por profissionais da informação. Visando apresentar uma solução para esse problema, pesquisadores

têm investigado formas de indexar e catalogar mais rapidamente essas informações. Entre as soluções propostas está a aplicação da indexação automática.

Os estudos relacionados à indexação automática tiveram início na década de 70 e envolviam o uso das palavras isoladas como termos de indexação.

Segundo Kuramoto (1995) e Souza (2005), o uso das palavras isoladas para a representação do conteúdo do documento não é adequado, pois elas não constituem uma unidade do discurso. Conforme destacado por Kuramoto (1995), as palavras isoladas, por não possuírem valor semântico e descritivo, não podem ser qualificadas como descritores de documentos.

Em 1995, Kuramoto propôs que as palavras isoladas fossem substituídas pelos Sintagmas Nominais (SNs) na indexação automática, pois esses agregam valor semântico à descrição do documento, constituem unidades de discurso e são melhores descritores para os documentos.

O sintagma é uma unidade sintática composta por um conjunto de palavras organizadas hierarquicamente em torno de um núcleo sintático. A classificação dos sintagmas é definida com base na função do seu núcleo. Por exemplo, os sintagmas nominais possuem um núcleo que desempenha a função de nome, podendo ser um substantivo, pronome substantivado, numeral ou palavra substantivada.

Para exemplificar os SNs e seu valor semântico em um documento, o Quadro 1 apresenta SNs retirados do título e resumo do artigo 1 do *corpus* utilizado neste trabalho.

Ao observar o Quadro 1, é possível identificar alguns SNs que podem se tornar os descritores do documento. Esses foram destacados em negrito por serem relevantes e terem valor descritivo com relação aos assuntos do documento.

Entretanto, dentre os SNs listados, alguns não possuem valor descritivo para o artigo, como os SNs destacados em itálico e em preto. Os SNs em itálico são SNs sem valor como descritores ou irrelevantes para a descrição do documento. Os SNs em preto também são irrelevantes para a descrição de

documentos por tratarem-se de expressões comuns ou expressões anafóricas, sendo considerados SNs vazios de significado.

**Quadro 1 -** Sintagmas nominais do artigo 1 do *corpus*

| sintagmas nominais | | |
|---|---|---|
| **transferência da Informação** | *o valor de conhecimento* | **os estudos sobre a gestão do conhecimento** |
| **análise para valoração de unidades de conhecimento** | *o conhecimento* | **a gestão do conhecimento** |
| **valoração de unidades de conhecimento** | as mais discutidas | *conhecimento disponível* |
| **unidades de conhecimento** | menos compreendidas questões | esta dificuldade |
| *conhecimento* | *o conjunto de conhecimento de uma organização* | uma organização |
| *o valor do conhecimento* | *o conhecimento de uma organização* | *o mercado* |
| *o conhecimento* | *parâmetros* | algum processo |

Fonte: Elaborado pelos autores.

Independente da classificação dos SNs quanto à relevância, a grande maioria dos SNs precisa passar por um processo de normalização ou canonização para ser utilizada como descritor ou palavra-chave dos documentos.

É nesse cenário que o presente trabalho está inserido, pois pretende propor e avaliar um método de normalização de SNs em termos canônicos através de critérios e diretrizes que proporcionem controle de vocabulário e menor dispersão terminológica, a fim de melhor descrever o conteúdo informacional presente nos documentos indexados automaticamente.

A par disso, o objetivo deste artigo consiste na proposição e avaliação de método de normalização de sintagmas nominais, objetivando que os SNs normalizados, além do valor semântico e discursivo, qualifiquem-se como descritores dos documentos em processos de indexação automática.

## 2 Indexação automática por sintagmas nominais

De acordo com Gil Leiva (1997, p. 53-54), a indexação automática é o processo no qual o programa de computador extrai ou atribui termos para a indexação de um documento.

Vieira (1988) define a indexação automática como uma tarefa realizada por um computador que é responsável pela análise de textos e a construção de índices de assunto, com o objetivo de possibilitar a recuperação do documento.

Borges (2009) define o processo de indexação automática como um modelo de extração que possui características estatísticas e probabilísticas. Essas características devem-se à utilização de técnicas em que são considerados fatores como a ocorrência e a repetição de palavras.

Com base nas definições de indexação automática, é possível afirmar que ela tem como principal característica a utilização do computador para indexar o documento (CORRÊA; LAPA, 2013).

Entre as décadas de 60 e 70, alguns autores, como Edmundson (1969), Garvin e outros (1969) e Salton e McGill (1983, sugeriram que os estudos sobre processamento de informação e linguística computacional deveriam focar as propriedades estruturais e semânticas da linguagem natural, evidenciando que a relação semântica é muito importante, pois permite identificar estruturas de formação de conceitos que possibilitam a escolha de termos representativos com significado.

A partir da década de 90, diversos estudiosos buscaram aplicar o processamento sintático e semântico dos textos na indexação automática, e isso foi feito principalmente através de estudos voltados à aplicação de SNs na indexação automática. Para que a indexação automática por SNs seja mais bem compreendida, é necessário primeiro conhecer os que são os sintagmas.

Dentro dos textos, existem as orações que são organizadas de acordo com leis sintagmáticas, e a elas são atribuídos grupos de unidades de significado chamado sintagmas. Segundo Perini (2005), sintagmas são grupos de significado que compõem sequências maiores de forma coesa e são subdivisões presentes naturalmente nas orações.

Na semântica, os sintagmas são unidades com significados únicos e coerentes e possuem classificação com base na função que desempenham. Quando sua função é de predicado, são classificados como sintagma verbal (SV); se possuem função de substantivo, são classificados como sintagma nominal (SN) (PERINI, 2005).

Os sintagmas possuem uma estrutura sintática que segue determinadas regras responsáveis por não permitir a dispersão das palavras. Na estrutura sintática, a posição de cada palavra é importante, pois é essa organização que dá sentido e forma ao sintagma.

Os sintagmas possuem diversas relações de dependência na oração, pois estabelecem ordens de subordinação para os elementos presentes na frase e podem ser classificados, com base na sua função, em:

a) sintagma nominal;
b) sintagma adjetival;
c) sintagma verbal;
d) sintagma adverbial;
e) sintagma preposicional.

No Quadro 2, é possível observar que os SNs possuem o núcleo composto por um nome (substantivo, pronome substantivo, numeral ou palavra substantivada), que pode ser acompanhado de duas estruturas pré ou pós-nucleares.

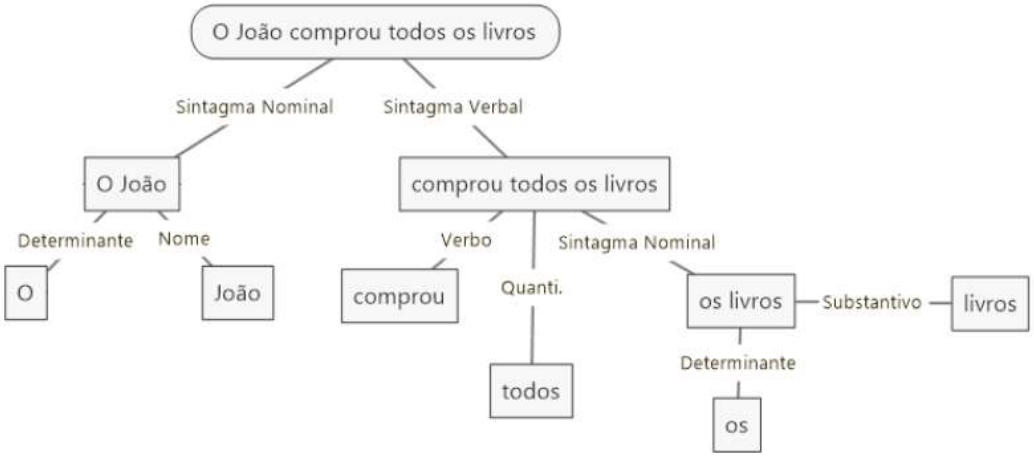
**Quadro 2 -** Elementos dos sintagmas nominais

| **Elementos que compõem os sintagmas nominais** | | |
|---|---|---|
| **Elementos pré-nucleares** | **Núcleo** | **Elementos pós-nucleares** |
| Predeterminantes, determinantes, quantificadores, possessivos sintéticos, numeral. | Nome (substantivo, pronome substantivo, numeral ou palavra substantivada). | Modificadores (palavra ou conjunto de palavras que qualificam o núcleo e restringem o seu sentido, inclusive outros nomes que também podem ser núcleos). |

Fonte: PERINI (2005).

Com o objetivo de exemplificar melhor um SN, é possível observar, na Figura 1, um exemplo de estrutura sintagmática de uma oração composta por sintagmas do tipo nominal e do tipo verbal.

**Figura 1 -** Exemplo de sintagma nominal e verbal




Fonte: Elaborado pelos autores.

Analisando a Figura 1, é possível identificar a presença de diversos sintagmas em uma mesma oração, como o SN "os livros", que está dentro do sintagma verbal "comprou todos os livros".

Assim, como os sintagmas obedecem a determinadas regras estruturais, os SNs possuem um conjunto de regras que determinam a sua formação. Silva (2014), baseado em Miorelli (2001) e Santos (2005), elaborou um quadro com um conjunto de regras de formação dos SNs (Quadro 3):

**Quadro 3 -** Elementos dos sintagmas nominais

| 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 3: DET + N | A informação |
| Regra 4: N | Informação |
| Regra 5: DET + N + DET + N + MOD | A filosofia e a ciência juntas |
| Regra 6: DET + DET + N + MOD | A minha recuperação da informação |
| Regra 7: MOD + N + MOD | Grande área da informação |
| Regra 8: DET+ DET + N | Uma certa área |

Fonte: Silva (2014, p. 50), baseado em Miorelli (2001) e Santos (2005).

Em 1995, Kuramoto apresentou, em sua pesquisa, uma classificação de níveis para os SNs. Segundo o autor, os SNs que possuíam outros SNs poderiam ter níveis atribuídos a eles de acordo com a quantidade de SNs presentes na oração. Portanto, o SN que não possui outro SN em sua estrutura é

considerado de nível um, o SN que possui um SN de nível um em sua estrutura é considerado nível dois e assim sucessivamente. Para exemplificar, temos: "Sistema de Classificação de Documentos" como um SN de nível três; "Classificação de Documentos" como um SN de nível dois; e "Documentos" como um SN de nível um.

Tendo a compreensão do que são os SNs, torna-se necessário entender por que os pesquisadores têm proposto a substituição das palavras isoladas pelos SNs na indexação automática de documentos. O principal motivo que justifica essa substituição é que as palavras isoladas não são suficientes para representar e recuperar a informação.

Um dos pioneiros nas pesquisas sobre a utilização de SNs na indexação automática foi Michel Le Guern (1991). Em sua proposta, ele justifica a substituição das palavras isoladas por SNs como descritores da informação, pois os SNs são portadores de significado, e, para a indexação e recuperação da informação, isso é bastante relevante. Ademais, ele enfatizou a diferença entre descritores e palavras, sendo os descritores unidades do discurso e as palavras, unidades da língua. As unidades da língua têm significado definido somente no texto do documento, enquanto as unidades do discurso preservam seu significado mesmo retiradas do seu contexto. Assim, as palavras isoladas (como unidades da língua) têm menos valor que os sintagmas nominais (como unidades do discurso) na indexação e recuperação da informação.

No contexto de documentos escritos em língua portuguesa, pesquisas como a de Kuramoto (1995) apresentaram resultados de representação e recuperação da informação mais satisfatórios quando utilizada a indexação por SNs. Portanto, as palavras isoladas (símbolos sem referências) são pouco adequadas como termos de indexação (KURAMOTO, 2002).

Tal qual a indexação por palavras isoladas, a indexação automática por SNs pode ser subdividida em etapas para a sua aplicação. Com base nas etapas definidas por Nascimento (2015) para a indexação automática por SNs, o presente artigo propõe a adição de mais uma etapa a esse processo: a normalização dos SNs em termos canonizados. Logo, o novo quadro de etapas da indexação automática por meio de SNs pode ser observado a seguir.

**Quadro 4 -** Etapas da indexação automática por sintagmas nominais

| **Processo de indexação automática por meio de SNs** | |
|---|---|
| **1ª Etapa** | **Identificação** dos SNs através das subetapas de "etiquetagem" e de "cotejamento dos léxicos etiquetados com as regras de formação dos SNs" |
| **2ª Etapa** | **Extração** dos SNs do **texto**, mostrando-os em listas. |
| **3ª Etapa** | **Normalização** dos SNs em termos canonizados |
| **4ª Etapa** | **Seleção** dos SNs normalizados com base em critérios que os classifiquem como "bons descritores" |

Fonte: Elaborado pelos autores.

No contexto da indexação automática por SNs, os seguintes autores investigaram as etapas de identificação, extração e seleção de SNs para textos escritos em português: Souza (2005), Souza e Raghavan (2006, 2014), Maia (2008), Maia e Souza (2010), Corrêa et. al.(2011), Silva (2014), Silva e Corrêa (2015), Corrêa e Bazílio (2017) e Nascimento e Correa (2018). Os autores Lopes (2012) e Martins (2014) investigaram indiretamente a etapa de normalização de SNs (3ª etapa), sendo considerados como trabalhos relacionados ao presente artigo.

Lopes (2012) e Martins (2014) trataram indiretamente da normalização de SNs, e a discussão de seus trabalhos tem como foco apresentar os critérios utilizados por eles para a normalização dos SNs. Tais critérios foram tomados como base para o desenvolvimento do método proposto de normalização de SNs.

A pesquisa de Lopes (2012) destaca-se por apresentar diversos critérios de extração de conceitos que foram adaptados para a normalização de SNs no presente artigo. Seu objetivo era propor um processo de extração de termos de coleções de documentos que seriam conceitualmente relevantes para construção de vocabulário controlado, tendo como base os SNs extraídos dos documentos.

Um método de extração de conceitos que possui quatro etapas foi proposto por Lopes (2012): (1) Extração de termos; (2) Ordenação de termos; (3) Identificação de conceitos; e (4) Aplicação dos termos e conceitos extraídos. A ferramenta utilizada para a extração dos termos foi o Extrator Automático de Termos para Ontologias em Língua Portuguesa (ExATOLP).

Lopes (2012) apresentou um conjunto de heurísticas baseadas em análises linguísticas que visam à extração de termos dos documentos. As heurísticas têm como propósito ajustar, excluir, reestruturar e incluir os SNs extraídos e são agrupadas em três categorias de acordo com a finalidade: heurísticas de ajuste, que modificam os SNs removendo artigos e pronomes; heurísticas de descarte, que excluem SNs que contêm numerais, caracteres especiais, símbolos, expressões anafóricas envolvendo pronomes ou que iniciam com advérbios; heurísticas de inclusão, que modificam os SNs, extraindo SNs implícitos através da remoção sucessiva de adjetivos e do tratamento das expressões adjetivas envolvendo conjunção, bem como a contabilização de frequência dupla de SNs adjacentes a predicados múltiplos.

Os resultados apresentados por Lopes (2012) através da aplicação dessas heurísticas nos SNs extraídos possibilitaram identificar quais eram mais eficazes. Além disso, percebe-se que algumas heurísticas têm a função de normalização dos SNs e outras podem ser ajustadas para esse propósito.

Martins (2014) desenvolveu um trabalho com o objetivo de avaliar o uso de SNs como característica para um sistema automático de classificação de documentos textuais em formato digital. A metodologia utilizada pelo autor foi dividida em duas etapas: na primeira, foi feito um teste qualitativo de comparação entre as representações dos documentos do *corpus*; e, na segunda etapa, foi utilizado o *software* SVMLight para classificar automaticamente as representações dos documentos.

Os resultados da pesquisa de Martins (2014) indicaram que, no momento de treinar o classificador automático, a utilização do processo de remoção de sufixos das palavras (*stemming*) permitiu o alcance de resultados mais satisfatórios, se comparados com a utilização do próprio SN extraído. Segundo o autor, alcançou-se 80% de classificação correta dos documentos utilizando apenas SNs puros, enquanto que utilizando os SNs após a remoção de sufixos, esse percentual foi de 100%. Foi também avaliada a remoção de certos quantificadores e qualificadores dos SNs como: “a”, “o”, “as”, “os”, “um”, “uma”, “uns”, “umas”, “aquele”, “aquela”, “aqueles”, “aquelas”, “este”, “esta”, “essa”, “esse”, “essas”, “esses”. Porém, optou-se pelo uso da ferramenta de remoção de sufixos, que já realizava automaticamente a remoção dessas

palavras vazias de significado dos SNs. Ambas as etapas têm a função de remover palavras pouco significativas e permitir a conflação das palavras restantes dos SNs, visando à posterior comparação léxica de equivalência dos mesmos. Tais etapas buscam aproximar o efeito da normalização dos SNs.

## 3 Metodologia

O presente artigo tem como objetivo inicial propor um método para normalização de SNs em termos canônicos, a fim de contribuir para a melhora qualitativa da indexação automática por SNs. O objetivo secundário consiste na avaliação do método proposto.

Trata-se de uma pesquisa exploratória com base bibliográfica e experimental que, de forma empírica, propõe e avalia o método de normalização de SNs em termos canônicos, tendo como dados da pesquisa os SNs extraídos de um *corpus* de 60 artigos de periódicos em Ciência da Informação escritos em português do Brasil e selecionados por Souza (2005).

Os SNs foram extraídos do título e resumo dos artigos, visando garantir uma amostra pequena e significativa dos SNs mais relevantes de cada artigo e minimizar o esforço humano na realização do experimento de normalização. Para extrair os SNs do título e resumo de cada artigo do *corpus* de Souza (2005), foi utilizado o *software* PyPLN (COELHO et al., 2013).

O PyPLN é capaz de realizar a etiquetagem, identificação e extração de SNs (COELHO et al., 2013) e foi criado a partir de um projeto de pesquisa da Escola de Matemática Aplicada da Fundação Getúlio Vargas. Trata-se de uma plataforma de Processamento de Linguagem Natural desenvolvida na linguagem de programação Python. Esse *software* utiliza o parser PALAVRAS como analisador sintático-semântico e permite realizar diversas análises linguísticas em documentos no formato texto.

Após a extração dos SNs, foram amostrados somente os que continham as palavras-chaves dos autores, visando normalizar os SNs considerados mais relevantes e mais próximos de descritores documentais. Estando estabelecida a amostra de SNs relevantes a serem normalizados, tendo como base os critérios

de normalização utilizados nos trabalhos relacionados e a análise da estrutura dos SNs mais relevantes extraídos do *corpus*, é desenvolvida e apresentada a proposta de método de normalização de SNs. Após a proposição do método, esse é avaliado na normalização dos SNs mais relevantes que foram extraídos do *corpus*.

A presente pesquisa consiste em sete etapas gerais, que podem ser observadas a seguir:

a) escolha dos documentos - seleção dos 60 documentos que compõem o *corpus* de Souza (2005);
b) coleta e organização dos documentos - o título e resumo de cada artigo são copiados para um arquivo em formato texto e submetido à etapa seguinte;
c) extração dos SNs - processo automatizado feito pelo *software* PyPLN;
d) formatação dos SNs - formatação dos SNs extraídos, quando as marcas do PyPLN nos SNs são removidas;
e) seleção dos SNs relevantes - teste de revocação realizado para selecionar os SNs mais relevantes para a normalização;
f) proposição e aplicação dos critérios de normalização - momento em que os SNs são submetidos à normalização;
g) análise dos SNs normalizados - análise da semelhança entre SNs normalizados e termos canonizados.

Para a primeira etapa, foram selecionados os 60 documentos presentes no *corpus* de Souza (2005). A justificativa para a sua escolha foi por já terem sido utilizados em outros trabalhos como os de: Kuramoto (1995; 2002), Souza e Raghavan (2006; 2014). Tal utilização agrega qualidade ao *corpus* como um material experimental. Na segunda etapa, foi feita a extração e cópia do título e resumo dos documentos do *corpus*, que se encontram em formato PDF e HTML, para arquivos separados no formato de texto simples (TXT).

Na terceira etapa, todos os 60 arquivos em formato de texto simples foram submetidos à extração dos SNs na plataforma do PyPLN. Cada título continha, em média, 11 palavras, sendo o menor com quatro palavras, o maior com 21 palavras e a soma total, 698 palavras. Cada resumo continha, em média,

126 palavras, sendo o menor com 32 palavras, o maior com 314 palavras e a soma, 7582 palavras.

A quarta etapa consistiu em formatar os SNs extraídos pelo PyPLN através da filtragem de caracteres. Foram eliminados dos SNs os caracteres especiais retornados pela plataforma, como: "_" (sublinhado), "*" (asterisco), aspas, chaves e sinais de pontuação (? ! , . : ;). A plataforma PyPLN, ao retornar os SNs, destaca os referentes e os quantificadores através desses caracteres. Por exemplo, no SN extraído "o *valor de_ _o conhecimento", o caractere sublinha é utilizado para denotar o quantificador do SN, e o asterisco, para denotar o referente do SN.

Foram extraídos pelo PyPLN 2787 SNs, resultando em uma média de 46 SNs por documento, sendo 100 o maior número de SNs extraídos por documento (referente ao documento 21) e 16, o menor (referente ao documento 58). No Quadro 5, é apresentada a quantidade de SNs extraídos de cada documento.

Na quinta etapa, foi realizada uma análise da revocação das palavras-chave, tendo como base os SNs extraídos pelo PyPLN e as palavras-chave atribuídas pelos autores dos documentos. O cálculo de revocação foi feito através do número de palavras-chave recuperadas nos SNs dividido pelo número total de palavras-chave do documento.

Através do teste de revocação, foi possível selecionar os SNs que contêm as palavras-chave dos autores, pois são considerados os SNs mais relevantes e, por conter os descritores, são passíveis de normalização.

A sexta etapa é a mais importante do artigo, pois trata da proposição e aplicação do método de normalização. Através dessa etapa, os SNs são submetidos a critérios do método de normalização que causarão alterações em sua estrutura, a fim de transformá-los em termos canonizados.

A sétima etapa consiste na avaliação da normalização dos SNs. A avaliação foi feita com base no número de SNs que foram considerados normalizados, sem alterar a estrutura válida do SN e nem a palavra-chave.

**Quadro 5 -** Número de SNs por Documento

| NÚMERO DE SNs POR DOCUMENTO | | | | | | | |
|---|---|---|---|---|---|---|---|
| **Documento** | **SNs** | **Documento** | **SNs** | **Documento** | **SNs** | **Documento** | **SNs** |
| DOC. 1 | 37 | DOC. 16 | 42 | DOC. 31 | 51 | DOC. 46 | 61 |
| DOC. 2 | 49 | DOC. 17 | 41 | DOC. 32 | 36 | DOC. 47 | 20 |
| DOC. 3 | 45 | DOC. 18 | 28 | DOC. 33 | 55 | DOC. 48 | 43 |
| DOC. 4 | 31 | DOC. 19 | 36 | DOC. 34 | 49 | DOC. 49 | 46 |
| DOC. 5 | 70 | DOC. 20 | 47 | DOC. 35 | 76 | DOC. 50 | 45 |
| DOC. 6 | 30 | DOC. 21 | 00 | DOC. 36 | 44 | DOC. 51 | 27 |
| DOC. 7 | 40 | DOC. 22 | 95 | DOC. 37 | 65 | DOC. 52 | 52 |
| DOC. 8 | 32 | DOC. 23 | 42 | DOC. 38 | 35 | DOC. 53 | 70 |
| DOC. 9 | 60 | DOC. 24 | 33 | DOC. 39 | 43 | DOC. 54 | 76 |
| DOC. 10 | 17 | DOC. 25 | 65 | DOC. 40 | 59 | DOC. 55 | 50 |
| DOC. 11 | 36 | DOC. 26 | 43 | DOC. 41 | 82 | DOC. 56 | 49 |
| DOC. 12 | 25 | DOC. 27 | 49 | DOC. 42 | 58 | DOC. 57 | 29 |
| DOC. 13 | 44 | DOC. 28 | 48 | DOC. 43 | 41 | DOC. 58 | 16 |
| DOC. 14 | 53 | DOC. 29 | 36 | DOC. 44 | 48 | DOC. 59 | 37 |
| DOC. 15 | 26 | DOC. 30 | 37 | DOC. 45 | 41 | DOC. 60 | 45 |
| | | | | | | **MÉDIA** | **46,45** |
| | | | | | | **DESVIO** | **17,09** |

Fonte: Elaborado pelos autores.

Para um SN ser considerado normalizado, não é preciso que tenha sido alterado por algum dos critérios de normalização, pois alguns poucos SNs já se encontram normalizados na extração. Entretanto, o SN normalizado deve apresentar, em sua estrutura, algum termo descritor que esteja presente no Tesauro Brasileiro de Ciência da Informação (TBCI), sendo o tesauro considerado como uma fonte confiável de termos canonizados.

Para permitir o casamento entre os SNs finais com os descritores do TBCI, foi utilizada a remoção de sufixos (*stemming*). A remoção do sufixo consiste em remover o sufixo das palavras dos SNs, sendo para isso utilizado o *software* Luke (LUKE, 2018).

Os radicais das palavras dos SNs finais foram, em seguida, utilizados para buscas por termos do TBCI. O SN final, contendo um termo descritor do TBCI, é categorizado como SN normalizado, uma vez que o tesauro é considerado uma fonte de descritores canonizados, organizados e estruturados.

## 4 Análise dos resultados

Nesta seção, discutem-se os principais resultados alcançados no presente artigo. Na primeira subseção, é descrito o método proposto de normalização de sintagmas nominais. Na segunda subseção, são discutidos os resultados obtidos na avaliação do método proposto na normalização dos SNs mais relevantes do *corpus*.

### 4.1 Método de normalização de sintagmas nominais

O método proposto consiste em duas etapas, que levam em consideração duas restrições: a preservação da validade ou da estrutura correta do SN (garantindo que nenhum SN deixe de ser válido) e a preservação da palavra-chave contida no SN.

A primeira etapa é composta por 89 regras desenvolvidas com o objetivo de minimizar os SNs extensos, aproximando-os das palavras-chave dos autores ou de termos descritivos. O desenvolvimento das regras teve por base a análise dos SNs relevantes. As regras consistem na identificação de: expressões comuns (como “alguns pontos”, “as várias”, “aspectos de”, “através de”, “cada tipo de”, “definição de”, etc.); sequências de palavras de determinadas classes gramaticais utilizadas como conectores entre SNs (preposições, pronomes, verbos e conjunções); e determinados sinais de pontuação (como travessão e parênteses).

A ação tomada ao identificar a aplicação de uma regra em um SN foi a exclusão do elemento e, se possível, a divisão do SN. Por exemplo: em “os estudos sobre a gestão do conhecimento”, foi identificada a regra “preposição + artigo (sem contração)” nos elementos “sobre a”. Sendo assim, os elementos que compõem a regra são excluídos e o SN é dividido em dois SNs. Portanto, do SN “os estudos sobre a gestão do conhecimento”, surgiram dois SNs: “Os estudos” e “Gestão do conhecimento”.

A segunda etapa é composta por critérios aplicados sequencialmente com o objetivo de eliminar elementos que não são importantes para os SNs.

Esses critérios foram propostos através da adaptação de critérios utilizados por Lopes (2012) e Martins (2014). No total, foram aplicados seis critérios:

a) remoção de artigos no início e fim dos SNs;
b) remoção de pronomes dos SNs;
c) remoção de advérbios dos SNs;
d) remoção de numerais;
e) remoção de verbos;
f) remoção de preposições e conjunções no início e no fim dos SNs.

Os critérios de normalização foram utilizados para simplificar e generalizar os SNs. Dessa forma, os SNs finais teriam uma forma mais simples e mais próxima de um termo de indexação.

### 4.2 Análise da aplicação do método proposto

Nesta seção são analisados os resultados obtidos na aplicação das etapas do método proposto de normalização dos SNs. Primeiramente, é feita uma análise referente ao teste de revocação aplicado para a seleção dos SNs relevantes. Em seguida, são analisadas as etapas um e dois do método de normalização.

Através do teste realizado no *corpus*, foi obtida a média da revocação das palavras-chaves de 55%, com desvio padrão de 28,3%. Para 38 documentos (63,3%), foram obtidos valores de revocação superiores a 50%; para 22 documentos (36,6%), foram obtidos valores de revocação inferiores a 50%. Esses resultados são bastante positivos, pois indicam que grande parte dos documentos possuem SNs no título e resumo que possibilitam a sua recuperação em um sistema de indexação automática por extração.

A baixa revocação das palavras-chave em alguns documentos foi causada por dois fatores de indexação. O primeiro era a ausência das palavras-chave no título e no resumo do documento; isso impossibilitou o PyPLN de recuperar SNs com as palavras-chave. No total, quatro documentos foram afetados por esse problema. O segundo foi o formato da palavra-chave: em alguns casos, a palavra-chave aproximava-se de um SN, mas, como não era

idêntico, isso não foi considerado. Por exemplo, no documento 15, a palavra-chave "Ensino e pesquisa" aproximava-se do SN "Relação Ensino-Pesquisa".

Portanto, os resultados poderiam ter sido melhores caso alguns problemas nas palavras-chaves fossem tratados, como o uso de caracteres especiais e de palavras estrangeiras.

**Quadro 6 -** Número de SNs após teste de revocação

| NÚMERO DE SNs APÓS TESTE DE REVOCAÇÃO | | | | | | | |
|---|---|---|---|---|---|---|---|
| **Documento** | **SNs** | **Documento** | **SNs** | **Documento** | **SNs** | **Documento** | **SNs** |
| DOC. 1 | 2 | DOC. 16 | 1 | DOC. 31 | 5 | DOC. 46 | 21 |
| DOC. 2 | 5 | DOC. 17 | 0 | DOC. 32 | 11 | DOC. 47 | 4 |
| DOC. 3 | 11 | DOC. 18 | 3 | DOC. 33 | 7 | DOC. 48 | 9 |
| DOC. 4 | 9 | DOC. 19 | 6 | DOC. 34 | 7 | DOC. 49 | 0 |
| DOC. 5 | 11 | DOC. 20 | 1 | DOC. 35 | 28 | DOC. 50 | 3 |
| DOC. 6 | 5 | DOC. 21 | 1 | DOC. 36 | 4 | DOC. 51 | 15 |
| DOC. 7 | 9 | DOC. 22 | 15 | DOC. 37 | 4 | DOC. 52 | 7 |
| DOC. 8 | 10 | DOC. 23 | 11 | DOC. 38 | 8 | DOC. 53 | 7 |
| DOC. 9 | 12 | DOC. 24 | 17 | DOC. 39 | 4 | DOC. 54 | 6 |
| DOC. 10 | 3 | DOC. 25 | 5 | DOC. 40 | 7 | DOC. 55 | 8 |
| DOC. 11 | 3 | DOC. 26 | 3 | DOC. 41 | 17 | DOC. 56 | 2 |
| DOC. 12 | 6 | DOC. 27 | 4 | DOC. 42 | 10 | DOC. 57 | 14 |
| DOC. 13 | 1 | DOC. 28 | 9 | DOC. 43 | 10 | DOC. 58 | 11 |
| DOC. 14 | 5 | DOC. 29 | 11 | DOC. 44 | 8 | DOC. 59 | 0 |
| DOC. 15 | 0 | DOC. 30 | 15 | DOC. 45 | 7 | DOC. 60 | 14 |
| | | | | | | **MÉDIA** | **7,53** |
| | | | | | | **DESVIO** | **5,54** |

Fonte: Elaborado pelos autores.

Antes do teste de revocação, 2786 SNs foram extraídos dos documentos, resultando em uma média de aproximadamente 46 SNs por documento (Quadro 5). Após o teste de revocação, o total de SNs remanescente foi de 452, restando então uma média de aproximadamente oito SNs por documento (vide Quadro 6).

Após a análise do teste de revocação, apresentam-se os resultados da aplicação das duas etapas do método proposto de normalização de sintagmas nominais: antes da aplicação da etapa um, existiam no total 452 SNs; após a sua aplicação, o número de SNs passou a ser de 569, ou seja, houve um aumento de 117 SNs, aproximadamente 25,8%. As regras foram aplicadas em 138 SNs e ocasionaram o surgimento de 255 novos SNs.

A aplicação das regras foi bastante satisfatória, pois possibilitou minimizar SNs que eram muito extensos, dividindo-os em SNs de menor nível que se adequavam melhor como um termo descritor.

As regras definidas na etapa um foram aplicadas manualmente e não apresentaram ambiguidade ou subjetividade na aplicação. Portanto, é possível que essas regras sejam utilizadas para a aplicação automática por um sistema.

Após a análise da etapa um, apresenta-se a análise dos resultados da etapa dois. Para facilitar a compreensão, os resultados foram analisados separadamente, seguindo a ordem de aplicação de cada critério.

A aplicação do critério de remoção de artigos do início e fim dos SNs foi a mais expressiva, pois dos 569 SNs, 231 SNs (40,5%) sofreram alterações devido à existência de artigos irrelevantes em sua estrutura. Os artigos definidos foram mais excluídos em comparação com os indefinidos. No total, 231 artigos foram removidos dos SNs: 211 artigos definidos e 20 artigos indefinidos. Todos os artigos removidos estavam posicionados no início dos SNs.

Logo, tendo por base tais resultados, pode-se afirmar que esse critério mostrou-se bastante eficiente, pois diversos artigos que não eram necessários estavam presentes na estrutura dos SNs e foram excluídos, deixando os SNs mais genéricos, simples e de fácil compreensão.

A aplicação do critério de remoção de pronomes resultou na modificação de seis dos 569 SNs, cerca de 1,05% do total de SNs. Os pronomes foram dos tipos indefinido, demonstrativo e possessivo.

Apesar de apresentar um número inferior ao critério de remoção de artigos, o critério de remoção de pronomes também demonstrou ser eficaz, pois permitiu remover pronomes desnecessários dos SNs.

O critério de remoção de advérbios não resultou em nenhuma alteração nos SNs. A possível causa para esse resultado é que houve a remoção de advérbios durante a aplicação das regras da etapa um.

Em seguida, foi aplicado o critério de remoção de numerais, visando eliminar numerais escritos por extenso ou em algarismos (romanos e arábicos). Como resultado, seis SNs foram alterados (aproximadamente 1%) de um total de 569. Assim, como o critério de remoção de pronomes, o critério de remoção de numerais possibilitou a normalização dos SNs, eliminando termos

desnecessários. Durante a aplicação desse critério, os SNs "1995 e 2000" e "545% titulados" foram excluídos, pois não se qualificavam mais como SNs.

O critério de remoção de verbos visa remover verbos desnecessários da estrutura dos SNs. A sua aplicação foi condicionada apenas aos verbos que não estivessem no particípio passado. Através desse critério, não houve alterações nos SNs, fator esse atribuído a algumas regras da etapa um que removiam verbos existentes nos SNs; portanto, não sobraram verbos para serem removidos durante a etapa dois.

Os resultados dos critérios de remoção de verbos e advérbios foram idênticos. Logo, é possível afirmar que eles não se mostraram necessários para o corpus desta pesquisa. Todavia, existe a possibilidade desses critérios serem úteis para a normalização de SNs de outro *corpus*.

O critério de remoção de preposições e conjunções do início e fim dos SNs foi aplicado de forma semelhante ao critério de remoção de artigos. No total, 110 SNs (19%) foram alterados através desse critério. Foram duas conjunções removidas no fim dos SNs e 108 preposições removidas do início dos SNs.

Após o processo de normalização, foi aplicada a remoção de sufixo nos SNs finais, possibilitando a comparação com os termos presentes no Tesauro Brasileiro de Ciência da Informação.

A avaliação do método de normalização proposto foi realizada com base na verificação da existência de termos canonizados do TBCI na estrutura de cada SN, sendo considerado SN normalizado em caso afirmativo.

No Quadro 7, é possível observar os quantitativos de SNs finais (alterados ou não) com e sem as palavras-chave dos autores. É apresentado também o quantitativo de SNs finais que são idênticos a termos do TBCI, o quantitativo de SNs que contemplam termos do TBCI e o quantitativo de SNs que não contêm termos do TBCI.

**Quadro 7 -** Relação dos SNs e TBCI

| | Idênticos a termo do TBCI | Contemplam termo do TBCI | Não contém termo do TBCI | TOTAL |
|---|---|---|---|---|
| **SNs finais com palavras-chave alterados** | 95 | 117 | 66 | 251 |
| **SNs finais sem palavras-chave alterados** | 12 | 7 | 50 | 69 |
| **SNs finais com palavras-chave não alterados** | 68 | 70 | 50 | 188 |
| **SNs finais sem palavras-chave não alterados** | 1 | 5 | 26 | 32 |
| **TOTAL** | 176 | 199 | 192 | 567 |

Fonte: Elaborado pelos autores.

Dos 567 SNs finais, 375 apresentaram termos do TBCI em sua estrutura, sendo, então, considerados normalizados. Em termos percentuais, 66,1% dos SNs finais foram considerados SNs normalizados. No total, 192 SNs finais (33,9%) não foram considerados normalizados, pois não estavam presentes no tesauro. Esses SNs não representam falhas do método de normalização, mas consistem em SNs que precisam ser traduzidos por indexadores para termos canônicos da linguagem documentária (116 SNs) ou simplesmente descartados por terem surgido da quebra de um SN extenso e não contemplarem as palavras-chave (76 SNs).

No Quadro 7, o total de SNs sem palavras-chave que não contêm termos do TBCI é de 50, porque, entre esses, não foram contabilizados os SNs "1995 a 2000" e "545% titulados", que foram excluídos na aplicação do critério de remoção de numerais por não se qualificarem mais como SNs.

Ao analisar os SNs considerados normalizados, ficou evidente que todos os SNs que passaram pelos critérios de normalização expressavam valor descritivo para o documento e, portanto, podem ser utilizados como descritores em bases de dados.

## 5 Conclusão

Com base nos resultados alcançados na avaliação do método de normalização de SNs, é possível afirmar que o objetivo proposto por este artigo foi alcançado.

O teste de revocação possibilitou uma amostragem dos SNs mais relevantes (maior valor descritivo) e mais adequados para a normalização.

Com relação às etapas do método proposto de normalização de SNs, as restrições de preservação dos SNs e das palavras-chave mostraram-se pertinentes durante o desenvolvimento do método e foram obedecidas integralmente na aplicação do mesmo.

A etapa um do método de normalização obteve um resultado bastante positivo, pois conseguiu diminuir o número de SNs extensos, criando novos SNs mais adequados como termos descritores dos documentos.

Para etapa um, ficou claro que sua utilização em um sistema automático de normalização de SNs é viável, uma vez que suas regras conseguiram, sem falha, gerar SNs descritores de nível menor para os documentos.

Os seguintes critérios da etapa dois apresentaram resultados positivos para a normalização de SNs: remoção de artigos, remoção de pronomes, remoção de numerais, remoção de preposição e conjunção. Através dos resultados, é possível afirmar que eles podem ser aplicados a SNs de outros *corpora* e também podem ser utilizados em um sistema automático de normalização de SNs.

Os critérios de remoção de verbos e advérbios para o *corpus* do presente artigo não se mostraram aplicáveis na segunda etapa. Entretanto, não é possível afirmar que são descartáveis para o método de normalização de SNs, pois existe a possibilidade de que, para outro *corpus,* eles apresentem resultados diferentes.

Portanto, os resultados indicam que o método proposto de normalização de SNs permite o alcance de resultados positivos, e sua aplicação em sistemas de indexação automática é viável.

Com relação aos resultados obtidos nesse artigo, em comparação com os resultados obtidos pelas pesquisas de Lopes (2012) e Martins (2014), é possível afirmar que esta pesquisa traz avanços em relação ao tratamento e normalização de SNs. Primeiramente ao formalizá-la como uma etapa distinta na indexação automática por sintagmas nominais; em seguida, ao propor método visando especificamente à normalização dos SNs, e, por último, através da avaliação empírica do método proposto.

Como limitações deste artigo, podemos apontar que o método proposto produz alguns SNs normalizados que contêm mais de um termo do TBCI, necessitando de uma análise posterior da estrutura desses, a fim de levantar possíveis regras para quebrá-los em outros menores contendo os descritores encontrados; o método de avaliação, baseado na presença de termos do TBCI nos SNs para categorizá-los como normalizados, pode apresentar falhas na precisão em alguns casos, e talvez a presença das palavras-chave seja uma alternativa para tal categorização.

Como trabalhos futuros, seria interessante analisar a estrutura dos SNs normalizados que contêm mais de um termo do TBCI, a fim de levantar possíveis regras para a sua quebra; investigar outros métodos de categorização de SNs como normalizados; rever a aplicação dos critérios de remoção de advérbios e verbos em outro *corpus*, visando confirmar a sua eliminação da etapa dois de normalização; mensurar os índices de revocação e precisão das palavras-chave para os SNs normalizados; analisar a normalização de SNs de *corpus* de outras áreas do conhecimento; analisar a aplicação dos critérios de normalização a todos os SNs, necessitando para isso de ferramenta automatizada de normalização e de categorização dos SNs como normalizados.

## Referências

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

COELHO, F. C. et al. PyPLN: a distributed platform for natural language processing. **ArXiv e-prints**, [s.l.], v. 2, arXiv:1301.7738, p. 1-8, Feb. 2013.

CORRÊA, R. et al. Indexação e recuperação de teses e dissertações por meio de sintagmas nominais. **AtoZ**, Curitiba, v. 1, n. 1, p. 11-22, jan./jun. 2011.

CORRÊA, R. F.; LAPA, R. C. Panorama de estudos sobre indexação automática no âmbito da ciência da informação no Brasil (1973-2012). **Ciência da Informação**, Brasília, v. 42, n. 2, p.255-273, maio/ago. 2013.

CORRÊA, R. F.; BAZÍLIO, L. H. T. Análise da extração de descritores como sintagmas nominais através do software OGMA. **Encontros Bibli**: revista

eletrônica de biblioteconomia e ciência da informação, Florianópolis, v. 22, n. 50, p. 44-58, set. 2017.

EDMUNDSON, H. P. New methods in automatic extracting. **Journal of the Association for Computing Machinery**, New York, v. 16, n. 2, p. 264- 285, Apr. 1969.

GARVIN, P. L. et al. **Some opinions concerning linguistics and reformation processing**. Washington: Center for Applied Linguistics, 1969.

GIL LEIVA, I. **La automatización de la indización, propuesta teórico-metodológica**: aplicación al área de biblioteconomía y documentación. 1997. Tese (Doutorado em Informação e Documentação) - Universidad de Murcia, Murcia, 1997.

KURAMOTO, H. Sintagmas Nominais: uma nova proposta para a recuperação de informação. **DataGramaZero**: Revista de Ciência da Informação, Brasília v. 3, n. 1, 9 p., fev. 2002.

KURAMOTO, H. Uma abordagem alternativa para o tratamento e a recuperação de informação textual: os sintagmas nominais. **Ciência da Informação**, Brasília, v. 25, n. 2, p. 1-18, 1995.

LE GUERN, M. Un analyseur morpho-syntaxique pour l'indexation automatique. **Le Français Moderne**, Paris, v. 59, n. 1, p. 22-35, juin 1991.

LOPES, L. **Extração automática de conceitos a partir de textos em língua portuguesa**. 2012. Tese (Doutorado em Ciência da Computação) - Faculdade de Informática, Pontifícia Universidade Católica do Rio Grande do Sul, Porto Alegre, 2012.

LUKE. [S.l.]: GitHub, 2018. Disponível em: https://github.com/DmitryKey/luke. Acesso em: 23 nov. 2018.

MAIA, L. C. G. **Uso de Sintagmas Nominais na classificação automática de documentos eletrônicos**. 2008. Tese (Doutorado em Ciência da Informação) - Escola de Ciência da Informação, Universidade Federal de Minas Gerais Minas Gerais, Belo Horizonte, 2008.

MAIA, L. C. G.; SOUZA, R. R. Uso de sintagmas nominais na classificação automática de documentos eletrônicos. **Perspectivas em Ciência da informação**, v. 15, n. 1, p. 154-172, jan./abr. 2010.

MARTINS, A. L. **O uso do sintagma nominal na recuperação de documentos [manuscrito]**: proposta de um mecanismo automático para classificação temática de textos digitais. 2014. Tese (Doutorado em Ciência da Informação) - Escola de Ciência da Informação, Universidade Federal de Minas Gerais Minas Gerais, Belo Horizonte, 2014.

MIORELLI, S. T. **Extração do sintagma nominal em sentenças em português**. 2001. Dissertação (Mestrado em Ciência da Computação) - Faculdade de Informática, Pontifícia Universidade Católica do Rio Grande do Sul, Porto Alegre, 2001.

NASCIMENTO, G. D. do. **Dos sintagmas nominais aos descritores documentais:** estudo de caso na indexação de teses e dissertações da área de direito. 2015. Dissertação (Mestrado em Ciência da Informação) - Departamento de Ciência da Informação, Universidade Federal de Pernambuco, Recife, 2015.

NASCIMENTO, G. D. do; CORREA, R. F. Avaliação de critérios para seleção de sintagmas nominais com valor para a recuperação da informação. **Transinformação,** Campinas , v. 30, n. 2, p. 179-192, ago. 2018.

PERINI, M. A. **Gramática descritiva do português**. 4. ed. São Paulo: Ática, 2005.

SALTON, G.; MCGILL, M. J. **Introduction to modern information retrieval**. New York: McGraw-Hill, 1983.

SANTOS, C. N. dos. **Aprendizado de máquina na identificação de sintagmas nominais:** o caso do português brasileiro. 2005. Dissertação (Mestrado em Sistemas e Computação) - Instituto Militar de Engenharia, Rio de Janeiro, 2005.

SILVA, T. J. da. **Indexação automática por meio da extração e seleção de sintagmas nominais em textos em língua portuguesa**. 2014. Dissertação (Mestrado em Ciência da Informação) - Departamento de Ciência da Informação, Universidade Federal de Pernambuco, Recife, 2014.

SILVA, T. J.; CORRÊA, R. F. Ferramentas para indexação automática: uma análise comparativa entre o ogma, parser palavras, lx-parser e a extração manual de sintagmas nominais. In: ENCONTRO NACIONAL DE PESQUISA EM CIÊNCIA DA INFORMAÇÃO, 16., 2015, João Pessoa. **Anais...** João Pessoa: UFPB, 2015.

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

SOUZA, R. R.; RAGHAVAN, K. S. A methodology for noun phrase-based automatic indexing. **Knowledge Organization**, Würzburg, v. 33, n. 1, p. 45-56, 2006

SOUZA, R. R.; RAGHAVAN, K. S. A extração de palavras-chave a partir de textos: um estudo exploratório utilizando sintagmas. **Informação & Tecnologia**, Marília, v. 1, n. 1, p. 5-16, 2014.

VIEIRA, S. B. Indexação automática e manual: revisão de literatura. **Ciência da Informação**, Brasília, v. 17, n. 1, p. 43-57, jan./jun. 1988.

## Standardisation method of noun phrases for automatic indexing

**Abstract:** This work proposes and evaluates a method of standardisation of noun phrases in canonical terms. This procedure aims to contribute to the qualitative improvement of automatic indexing avoiding the terminological dispersion and preserving the keywords present within the noun phrases. The research is exploratory and empirical, based on bibliographic research and an experiment in a corpus composed of scientific articles in Information Science. The proposed standardisation method contains rules and criteria that follow the constraints of preserving the valid structure of the noun phrase and the keywords. The method evaluation consists of the analysis of the presence of terms of the Brazilian Thesaurus in Information Science (TBCI) in the noun phrases resulting from the application of the proposed rules and criteria. The method consists of two stages: the first consists of 85 rules to reduce the size of the noun phrases, and the second stage contains seven criteria responsible for eliminating unnecessary grammatical elements from the noun phrases. The results of the evaluation indicate that the proposed method allows the achievement of positive results, even with two criteria of the second stage not presenting results for the corpus. It concludes that the application of the method in automatic indexing system is feasible and brings good results.

**Keywords:** Automatic indexing. Noun phrases. Standardisation of noun phrases. Keywords. Thesaurus.

Recebido: 10/04/2018
Aceito: 11/07/2018