

EXTRAÇÃO E SELEÇÃO AUTOMÁTICA DE SINTAGMAS NOMINAIS EM TEXTOS EM LÍNGUA PORTUGUESA: uma análise de citação do estado da arte

Tiago José da Silva (UFPE)

tiago.stallin@gmail.com

Renato Fernandes Correa (UFPE)

fc_renato@yahoo.com.br

EIXO TEMÁTICO: Análise de Citação

MODALIDADE: Apresentação oral

1 INTRODUÇÃO

Os sintagmas nominais (SNs) são objetos de estudo de várias áreas do conhecimento como Linguística, Ciência da Computação e Ciência da Informação. Contudo não há registro de pesquisa no Brasil que faça a análise de citação dos trabalhos da literatura acerca desse tema em língua portuguesa.

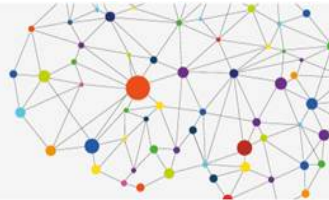
A análise de citação, segundo Primo et al (2008) permite uma autocompreensão da área, já que é possível verificar quem são os autores mais influentes acerca do tema. Leal (2005) argumenta que a análise de citação possibilita a descoberta das origens do conhecimento produzido, quais as bases e os cernes utilizados na construção do conhecimento.

Desse modo, esse trabalho faz uma análise de citação dos principais trabalhos sobre a extração e seleção automática de SNs em português no intuito de estabelecer a relação de comunicação existente entre os trabalhos e a representação cumulativa constituída por eles.

2 SINTAGMAS NOMINAIS: CONCEITO, EXTRAÇÃO E SELEÇÃO AUTOMÁTICA

SNs podem ser definidos como a menor unidade de sentido do discurso e que possui uma estrutura sintática e lógico-semântica (KURAMOTO, 1999), é o conjunto que traz o tema do enunciado (KURAMOTO, 1995). Resumidamente, SN pode ser conceituado como o conjunto de palavras que tem como núcleo o substantivo ou a palavra substantivada, tendo como exemplo de SN a seguinte expressão “*Os estudos bibliométricos*”.

A extração automática de SNs implica em soluções para diversas áreas como Recuperação de Informação, resolução de anáforas, construção de ontologias, entre outros como afirma Morellato (2010). Para Corrêa et al (2011), os SNs são extraídos do texto e



analisados a fim de facilitar o processo de indexação automática. Entende-se como extração automática de SNs o processo computacional que tem por objetivo retornar SNs contidos em um texto, tendo como entrada a frase que passa por um pré-processamento em que os léxicos são etiquetados (com categorias gramaticais) e submetidos às regras gramaticais de uma língua natural e tendo como saída o texto com SNs marcados e/ou extraídos. Algumas ferramentas podem ser citadas como indentificadores e/ou extratoras de SNs em língua portuguesa como o OGMA, PALAVRAS, LX-Parser, Curupira, Grammar Play e o SISNOP.

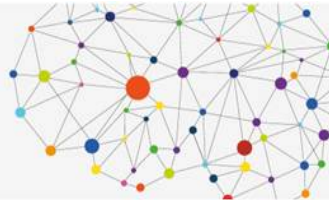
A seleção automática de SNs tem como objetivo escolher dentre os SNs extraídos automaticamente os que tem valor de descritor documental. Dentre algumas metodologias de seleção de SNs, referencia-se à formulada por Souza (2005) que consiste na verificação da relevância semântica por meio da quantificação da frequência 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 tesouro de domínio.

2 ANÁLISE DE CITAÇÃO

Segundo Leal (2005) citação é um evento ou fato social, pois ela é feita a partir do contexto do sujeito, já que se leva em consideração toda a sua experiência anterior, sua rede de conhecimento de outros autores, suas próprias indagações e as de outros. Para Cronin (1984), citações são símbolos, marcas metáforas e sinais que denotam trabalhos particulares e implicam alguma forma de relacionamento entre o citante e o citado. Enquanto que para Garfield (1979, p. 3) é “uma representação precisa, não ambígua de um assunto que não requer interpretação e é imune a mudanças terminológicas”, tendo sua precisão retida por meio do tempo e “podendo ser usada em documentos escritos em línguas diferentes”.

O ato citar, na concepção de Alvarenga (1998), remete a aspectos psicológicos, sociológicos e históricos, tendo também outras dimensões como influências, a saber: o fenômeno da autocitação, influências entre autores e instituições.

Análise de citação se baseia na hipótese de que ao citar o autor de um trabalho prévio, há sempre algum significado que permite estabelecer uma relação de comunicação que vai se constituir nas representações de certos campos do saber (SENGUPTA, 1992). Segundo esse mesmo autor, a responsabilidade pelo fluxo de informação e de comunicação entre aquele que cita e o que lê é da pessoa citante.



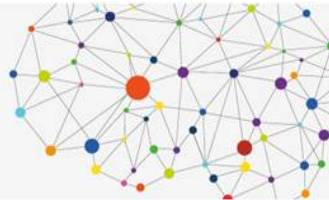
Oliveira, Gracio e Silva (2010) apontam que análise de citação identifica grupos de cientistas e suas publicações, evidencia os pesquisadores de maior impacto na área, apontando seus paradigmas, aportes metodológicos pertinentes, bem como pesquisadores “vanguardistas”.

3 METODOLOGIA

Os trabalhos escolhidos foram os que falavam sobre a identificação e/ou seleção de SNs em textos escritos em língua portuguesa. Assim, têm-se: Kuramoto (1999), Souza (2005), Maia (2008), e Corrêa et al (2011) da área de Ciência da Informação; Bick (2000), Othéro (2004), Costa (2007) e David (2007) da Linguística; Vieira et al (2000), Miorelli (2001), Santos (2005), Arcoverde (2007), Lopes (2009), Pinheiro (2009), Morellato (2010) da Ciência da Computação. Daí foram analisadas as referências bibliográficas e o contexto de cada citação que mantivesse qualquer relação com algum desses trabalhos supracitados, logo em seguida foi feito um diagrama estabelecendo os elos entre esses pesquisadores. Vale ressaltar que foram considerados os artigos que deram origem ou derivados dos trabalhos supracitados, mas considerando-os como um só trabalho. Não foram levadas em consideração as autocitações, focando essa pesquisa apenas na relação existentes entre os pesquisadores e suas áreas de domínios.

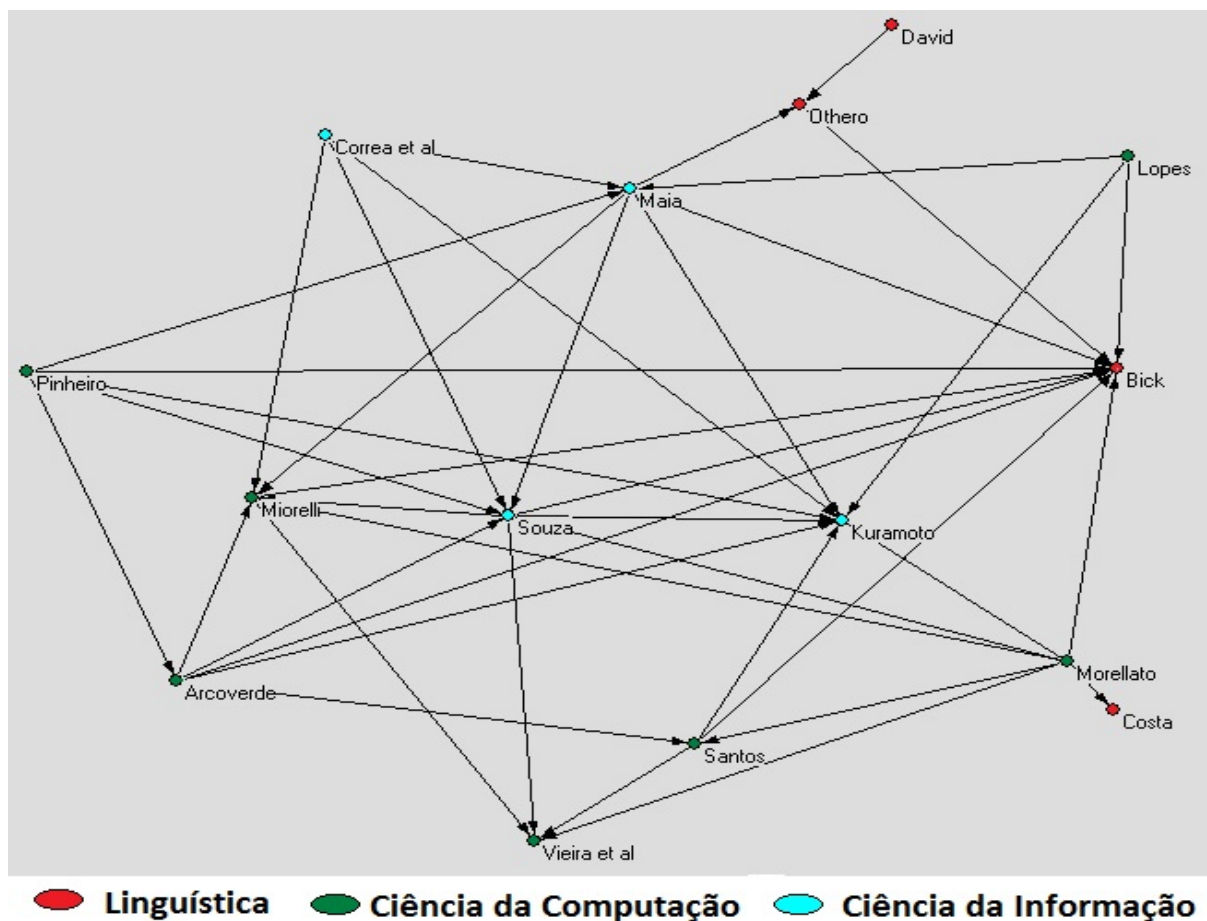
4 RESULTADOS E ANÁLISE

Os resultados apresentados pela análise de citação demonstram que Bick foi referenciado por nove autores: Pinheiro, Morellato, Lopes, Maia, Arcoverde, Santos, Souza, Othéro, Miorelli. Oito trabalhos referenciaram a Kuramoto: Morellato, Arcoverde, Santos, Lopes, Souza, Pinheiro, Corrêa et al e Maia. Souza foi citado por cinco trabalhos: Arcoverde, Maia, Morellato, Corrêa et al e Pinheiro. Para Miorelli houve quatro menções: Arcoverde, Maia, Souza e Corrêa. Já Vieira et al foram mencionados por três: Miorelli, Santos e Souza. Maia também teve três menções: Pinheiro, Lopes e Corrêa. Othéro foi mencionado duas vezes: Maia e David. Santos foi citado uma vez por Morellato. Costa que foi mencionado por Morellato, como também Arcoverde que foi citado por Pinheiro. Os trabalhos de Corrêa et al, Lopes, Pinheiro e David não foram referenciados por trabalhos posteriores dentro desse levantamento com os quinze trabalhos.



A Figura 1 demonstra a relação de trabalhos citantes e citados. Na qual a seta vai do citante ao citado.

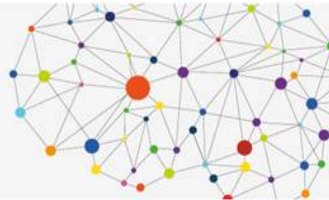
Figura 1 – Relação entre citante e citado nos trabalhos acerca da extração e/ou seleção de SNs



Fonte: os autores

De acordo com os resultados representados na Figura 1, percebe-se que os trabalhos desenvolvidos na área da Linguística não fazem referência a nenhum trabalho da Ciência da Informação nem da Ciência da Computação. Infere-se que isso acontece devido ao foco diferenciado dado ao objeto de estudo SN, enquanto na Linguística a ênfase está nas regras de formação de SNs, na Ciência da Informação os estudos objetivam encontrar dentre os SNs extraídos de um texto os que mais se enquadram como descritores de um documento, já os trabalhos desenvolvidos na área da Ciência da Computação estão focados na maneira em que o computador possa fazer a extração e seleção de SNs automaticamente.

Bick (2000) desenvolveu o seu trabalho à mesma época de Kuramoto (1995; 1999). Com o parser PALAVRAS permitiu que análises morfossintáticas fossem feitas com maior



precisão. Esse software foi utilizado por vários autores como Miorelli (2001), Santos (2005), Souza (2005), Morellato (2010), entre outros, ou para fazer parte de umas das etapas das metodologias ou para fazer as validações das metodologias propostas por esses autores.

Dos trabalhos desenvolvidos por diferentes autores sobre a extração e/ou seleção de SNs, os que seguem uma continuidade são Kuramoto (1995) com a proposição de oferecer regras que possibilitem a identificação de sintagmas nominais, Souza (2005) que cria taxas de relevância para as os sintagmas nominais, Maia (2008) que cria uma ferramenta que implementa essas regras e metodologias de pontuação de relevância em um *software* e Corrêa et al (2011) que avalia esse *software* na indexação automática de resumos de teses e dissertações.

Morellato (2010) cita os trabalhos de Kuramoto (1995; 1999), Miorelli (2001), Souza (2005), Santos (2005) e Costa (2007) porque percebe que esses estudiosos almejavam a obtenção de altos valores de precisão e de abrangência na extração de SNs para que pudessem ser aplicados nos desenvolvimentos de sistemas em diversas áreas.

4 CONSIDERAÇÕES FINAIS

Percebe-se com a análise de citação que boa parte dos trabalhos posteriores não citavam os trabalhos anteriores, independentemente da área. Isso demonstra que a Linguística, por exemplo, mesmo focando suas pesquisas na Linguística Computacional, não creditavam alguma menção aos trabalhos das outras áreas como Ciência da Informação e Ciência da Computação citados nessa pesquisa. Mas vale ressaltar que havia citação a outros autores da área de Ciência da Computação mas que trabalhava a extração e/ou seleção de SNs em outras línguas como a francesa e a inglesa.

Constata-se que a análise de citação permite perceber as relações multidisciplinares entre as ciências. Também se percebe as relações existentes entre as pesquisas dos autores e aprofundando na análise dos documentos podemos chegar as razões da citação a determinado trabalho.

REFERÊNCIAS

ALVARENGA, L. Bibliometria e Arqueologia do saber de Michel Foucault: traços de identidade teórico-metodológicas. **Ciência da Informação**, Brasília, v. 27, n. 3, set./dez. 1998.

ARCOVERDE, J. M. A. **Indução de filtros linguisticamente motivados na recuperação de informação**. 2007. 107 f. Dissertação (Mestrado em Ciências de Computação e Matemática Computacional), Instituto de Ciências Matemáticas e de Computação, USP, São Carlos, 2007.

BICK, E. **The Parsing System Palavras**: Automatic Grammatical Analysis of Portuguese in a Constraint Grammar Framework. 2000. 505 f. Tese (Doutorado em Linguística), Department Of Linguistics, University Of Aarhus, Aarhus, 2000.

CORREIA, R. F. et al. Indexação e recuperação de teses e dissertações por meio de sintagmas nominais. **AtoZ**: Novas Práticas em Informação e Conhecimento, Curitiba, v. 1, n. 1, p. 11-22, 2011.

COSTA, F. N. Q. M. C. **Deep Linguistic Processing of Portuguese Noun Phrases**. 2007. 214 f. Dissertação (Mestrado em Informática), Faculdade de Ciências da Universidade De Lisboa, Lisboa, 2007.

CRONIN, B.. **The citation process**. London: Taylor Graham, 1984.

DAVID, K. A. **Sintaxe das expressões nominais no português do Brasil**: uma abordagem computacional. 2007. 118 f. Dissertação (Mestrado em Linguística), Faculdade de Letras, UFC, Fortaleza, 2007.

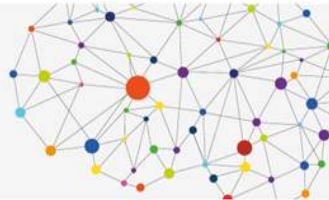
GARFIELD, E. **Citation Indexing**: Its Theory and Application in Science, Technology and Humanities. New York: John Willey & Sons, 1979.

KURAMOTO, H. **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.

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. 182-192, maio/ago. 1995.

LEAL, I. C. Análise de citações da produção científica de uma comunidade [manuscrito]: a construção de uma ferramenta e sua aplicação em um acervo de teses e dissertações do PPGCI-UFMG. 2005. 94f. Dissertação (Mestrado em Ciência da Informação), ECI, UFMG, Belo Horizonte, 2005.

LOPES, L. **Extração automática de conceitos a partir de textos em língua portuguesa**. 2011. 156 f. Tese (Doutorado em Ciência da Computação), Faculdade de Informática, PUCRS, Porto Alegre, 2011.



MAIA, L. C. G. **Uso de sintagmas nominais na classificação automática de documentos eletrônicos**. 2008. 158 f. Tese (Doutorado em Ciência da Informação), ECI, UFMG, Belo Horizonte, 2008.

MIORELLI, S. T. **Extração do sintagma nominal em sentenças em português**. 2001. 98 f. Dissertação (Mestrado em Ciência da Computação) – Faculdade de Informática, PUCRS, Porto Alegre, 2001.

MORELLATO, L. V. **Metodologia Computacional para Identificação de Sintagmas Nominais na Língua Portuguesa**. 2010. 112 f. Dissertação (Mestrado em Ciência da Computação) – Departamento de Informática, UFES, Vitória, 2010.

OLIVEIRA, E. F. T; GRACIO, M. C. C.; SILVA, A.C.C. Investigadores de mayor visibilidad en Organización y Representación del Conocimiento: un estudio desde el análisis de cocitaciones. **Scire**, v. 16, p. 39-46, 2010.

OTHERO, G. de Á. **Grammar Play: um parser sintático em Prolog para a língua portuguesa**. 2004. 265 f. Dissertação (Mestrado em Letras) – Faculdade de Letras, PUCRS, Porto Alegre, 2004.

PINHEIRO, M. S.. **Uma abordagem usando sintagmas nominais como descritores no processo de mineração de opiniões**. 2009. 110 f. Tese (Doutorado em Engenharia Civil) – COPPE, UFRJ, Rio de Janeiro, 2009.

PRIMO, Alex et al. Análise de citações dos trabalhos da Compós 2008. **E-compós**, Brasília, v.11, n.3, set./dez. 2008.

SANTOS, C. N. dos. **Aprendizado de máquina na identificação de sintagmas nominais: o caso do português brasileiro**. 2005. 104 f. Dissertação (Mestrado em Sistemas e Computação), IME, Rio de Janeiro, 2005.

SENGUPTA, I. N. Bibliometrics, informetrics, scientometrics and librametrics: na overview. **Libri**, v. 42, n. 2, p. 75-98, 1992.

SOUZA, R. R. **Uma proposta de metodologia para escolha automática de descritores utilizando sintagmas nominais**. 2005. 197 f. Tese (Doutorado em Ciência da Informação) - ECI, UFMG, Belo Horizonte, 2005.

VIEIRA, Renata et al. Extração de sintagmas nominais para o processamento de co-referência. In: Encontro para o Processamento Computacional do Português Escrito e Falado (PROPOR), 5. **Anais...** São Carlos: ICMC/USP, 2000. p. 165-173.