Do termo à estruturação semântica: representação ontológica do domínio da Nanociência e Nanotecnologia utilizando a Estrutura Qualia

Deni Yuzo Kasama
Universidade Estadual Paulista (UNESP)
Fundação de Amparo à Pesquisa do Estado de São Paulo (FAPESP)
deni@me.com

Claudia Zavaglia
Universidade Estadual Paulista (UNESP)
zavaglia@ibilce.unesp.br

Gladis Maria de Barcellos Almeida
Universidade Federal de São Carlos (UFSCar)
gladis@ufscar.br

Resumo

O presente artigo apresenta as etapas de elaboração de uma ontologia do domínio da Nanociência e Nanotecnologia com vistas à sistematização do léxico dessa área de especialidade, por meio de formalismos descritos na Teoria do Léxico Gerativo, com ênfase na Estrutura Qualia e seus quatro papéis semânticos, a saber: Formal, Constitutivo, Agentivo e Télico. A partir de um córpus da área, e valendo-nos de métodos semiautomáticos para a extração de candidatos a termos e identificação de relações semânticas, delineamos um mapeamento semântico partindo de relações de herança conceitual, cuja representação foi feita em linguagem OWL, com o auxílio da ferramenta Protégé.

1. Introdução

No âmbito do Processamento de Línguas Naturais (doravante PLN), o léxico desempenha papel crucial para o eficiente funcionamento de sistemas que visam a tratar automaticamente a língua. Dentre algumas aplicações em PLN, podemos citar a sumarização automática, a mineração de textos, a recuperação de informação e a tradução automática, para os quais um simples elenco de palavras não é suficiente. Segundo o tipo de aplicação, outras informações linguísticas tornam-se necessárias como, por exemplo, um sistema de reconhecimento de fala que necessita de um léxico subjacente que contenha informações do tipo fonológico. Estudos dessa natureza, bem como de dados morfossintáticos, têm sido conduzidos com expressivo sucesso, no que tange a sua correta identificação e categorização por sistemas computacionais. Entretanto, a representação semântica do léxico, seja ele geral ou especializado, é ainda terreno pouco sólido para pesquisas em Linguística Computacional que fazem uso desse tipo de informação. Os formalismos representacionais hoje conhecidos não se mostram eficientes o bastante para expor e tratar a questão da significação lexical com a devida precisão que sistemas de PLN exigem.

Esse caráter pouco domesticável do léxico explica-se por sua estreita relação com a realidade extralínguistica, a qual, segundo Biderman, é demonstrada pelos signos linguísticos ou unidades lexicais "que designam os elementos desse universo segundo o recorte feito pela língua e pela cultura correlatas. Assim, o léxico é o lugar da estocagem da significac□ão e dos conteúdos significantes da linguagem humana" (Biderman, 1996, p.27). Daí o fato de o léxico de uma língua encontrar-se em constante dinamicidade, além de que, para um mesmo significante, podem-se observar múltiplos significados. A tratabilidade dessas informações por máquina depende justamente da eficácia da representação semântica adotada.

Nesse sentido, podemos apontar trabalhos como o de Reeve e Han (2007) que faz uso de relações semântico-lexicais em um sistema de sumarização automática para textos do domínio médico, ou ainda o método desenvolvido por Ercan e Cicekli (2008) que faz uso extensivo de conhecimento semântico-lexical para o funcionamento de um sumarizador. Apontamos ainda para Rino e Pardo (2003) em que são descritos alguns sistemas de sumarização que fazem uso de repositórios lexicais em língua portuguesa; em mineração de textos, Alsumait et al. (2010) apontam para a importância de conhecimento semântico agregado ao léxico para processos de inferência de assuntos tratados em um determinado texto; Fox (1980) trata da importância de relações lexicais para a recuperação da informação; e, por fim, dentro os trabalhos que apontam para uso de repositórios lexicais como estratégia para a condução de tarefas em sistemas de

tradução automática, podemos citar Dorr (1992 e 1993) e Hutchins e Somers (1992).

Os trabalhos acima mencionados fazem uso de léxicos e alguns mencionam explicitamente a melhoria dos resultados quando esse tipo de dado é levado em consideração, uma vez que muitos sistemas que visam ao tratamento de informação textual valem-se apenas de estatísticas de frequência e coocorrência.

A necessidade de um modelo de representação semântica verifica-se já em Katz e Fodor (1963), e Jackendoff (1983) com sua proposta cognitiva que se baseia em uma hipótese ontológica e epistemológica. Os modelos propostos, então, envolviam a decomposição de traços em primitivos semânticos que se mostravam eficientes apenas para uma pequena parte do léxico; mais recentemente, os modelos de representação semântica adotados apontam para um teoria composicional, geralmente utilizada quando faz-se necessário um maior formalismo. Tais fatores nos levaram à adoção de uma teoria que nos permitisse representar o léxico de um domínio por meio de relações semânticas no interior de um conjunto vocabular especializado, bem como de um modelo de representação altamente utilizado para estruturação de um conhecimento, a saber, as ontologias.

O presente artigo subdivide-se da seguinte forma: na seção 2, tratamos do conceito de ontologias, sua utilidade nesta pesquisa e de como seu conceito difere do conceito de mapa conceitual; na seção 3, apresentamos a Teoria do Léxico Gerativo, mais especificamente a Estrutura Qualia e de sua importância para a representação formal de informações semânticas; na seção 4, detalhamos o desenvolvimento da pesquisa: o córpus[1] utilizado, a extração semiautomática[2] de candidatos a termos, a definição de classes e subclasses, o método utilizado para levantamento de relações semânticas e a subsequente implementação dos dados obtidos na ferramenta Protégé; na seção 5, apresentamos alguns dos dados alcançados; na sequência (seção 6), discutimos algumas questões envolvidas em uma pesquisa deste gênero; e na seção 7, apresentamos as conclusões e possíveis desdobramentos futuros.

2. Ontologias

Filósofos, de Aristóteles a Wittgenstein, trataram da existência de categorias lógicas que levariam a uma categorização geral das coisas que existem no mundo, muito embora com visões diferentes (do realismo ao relativismo, respectivamente, passando pelo idealismo kantiano). O termo "ontologia" nasce justamente na filosofia como o estudo da natureza do ser e sua existência, sob uma ótica metafísica e hoje estende-se para áreas como as Ciências da Computação, da Informação e Linguística.

Como já dito na seção anterior, o uso de ontologias tem se mostrado um meio eficiente de representação de conceitos semanticamente relacionados, servindo não só aos propósitos de sistemas de banco de dados, como também para o PLN. Isso porque as ontologias envolvem os formalismos necessários para a descrição de um conhecimento permitindo o uso da lógica e a realização de inferências a partir das informações estruturadas.

Gruber assim define ontologias:

> *"No contexto das ciências da computação e informação, uma ontologia define um conjunto de primitivos representacionais com os quais se modela um domínio do conhecimento ou discurso. Os primitivos representacionais são tipicamente classes (ou conjuntos), atributos (ou propriedades), e relacionamentos (ou relações entre membros das classes). As definições dos primitivos representacionais incluem informações sobre seu significado e restrições sobre sua aplicação consistente de forma lógica".*[3] (Gruber, 2008)

Como forma de estruturar um conhecimento (especializado ou terminológico, neste trabalho), valemo-nos do conceito de ontologias a fim de garantir (i) uma estruturação conceitual baseada em relações de classes e subclasses (ou de hiperônimos e hipônimos) que prevê a herança de conceitos; (ii) um padrão que vem sendo extensivamente utilizado para descrição de domínios; e (iii) um formalismo capaz de garantir o tratamento computacional dos dados linguísticos levantados a partir de um córpus e com recursos à disposição para realizar inferências automáticas a partir de restrições pré-

[1] Adotamos aqui o termo "córpus", tanto para o singular quanto para o plural, grafado com o acento agudo na vogal tônica, em português, em detrimento do latinismo (ou anglicismo) *corpus* e *corpora*. É de nosso conhecimento, entretanto, que, em artigos e livros, encontram-se as duas opções de grafia em vigor, de acordo com a escolha de cada autor.

[2] Advogamos o uso de "semiautomático" uma vez que entendemos ser necessário a intervenção humana em um ou mais etapas do processo.

[3] As citações em língua estrangeira são de tradução dos autores.

determinadas que possibilitam popular classes que atendam tais restrições.

Com efeito, Guarino (1998) relata a existência de três tipos de ontologias: 1. Ontologias genéricas (*top-level ontologies*), 2. Ontologias de domínio (*domain ontologies*) e Ontologias de tarefa (*task ontologies*) e 3. Ontologias de aplicação (*application ontologies*). Este trabalho concentra-se em (2), mais especificamente sobre as ontologias de domínio, definidas pelo autor como o tipo de ontologia que "descreve o vocabulário relacionado a um domínio genérico (como medicina ou automóveis) ou uma tarefa ou atividade genérica (como diagnóstico ou venda), através de uma especialização dos termos introduzidos na ontologia genérica".

O uso de ontologias no processo de criação de produtos terminológicos não é uma etapa necessariamente nova, mas imprescindível quanto a uma possível reutilização em aplicações como aquelas voltadas para a Web Semântica (Berners-Lee et al., 2001, p. 36), por exemplo. Ademais, como aponta Almeida (2000), o papel dos mapas conceituais interfere diretamente na própria pesquisa terminológica, visto:

> *"1) possibilitar um mapeamento mais sistemático de um campo de especialidade; 2) circunscrever a pesquisa, já que todas as ramificações da área-objeto, com seus campos, são previamente mapeadas; 3) delimitar o conjunto terminológico; 4) determinar a pertinência dos termos, pois separando cada grupo de termos pertencentes a um determinado campo, poder-se-á apontar quais termos são relevantes para o trabalho e quais não são; 5) prever os grupos de termos pertencentes à área-objeto, como também os que fazem parte de matérias conexas; 6) definir as unidades terminológicas de maneira sistemática e, finalmente, 7) controlar a rede de remissivas"* (Almeida, 2000, p. 120).

Cabré (1999, p. 144) aponta que os termos mantêm relações (não necessariamente hierárquicas) entre si, compondo dessa forma um mapa conceitual. Ainda para a mesma autora (2003), o lugar que o termo ocupa nesse mapa determina o seu significado, o que denota a importância de tais estruturas no processo de elaboração das definições em um dicionário especializado.

Algumas questões podem ser levantadas quanto ao uso dos termos "ontologia", "mapa conceitual" e "taxonomia". Entendemos haver uma diferenciação entre os conceitos, embora haja uma semelhança evidente, uma vez que, tanto terminólogos quanto ontólogos, trabalham em suas pesquisas com campos conceituais ou nocionais e com listas de unidades lexicais superordenadas em classes. Faz-se necessário, contudo, destacar conceitos como o de hereditariedade semântica e herança múltipla, presentes em ontologias. A esses conceitos agregam-se os de "atributos" e "propriedades", bem como os de "restrições" e "instâncias" ou "membros de classes", conforme citação anterior de Gruber (2008).

Nas Ciências da Computação, mapas conceituais são vistos como uma fase preliminar ao delineamento de uma ontologia, ou ainda, como se pode observar em Graudina (2008), uma reutilização de uma ontologia para fins didáticos:

> *"Levando em consideração similaridades óbvias entre ontologias e mapas conceituais, pesquisas de conversão de ontologia em mapa conceitual foram realizadas. Geração de mapas conceituais a partir de ontologias OWL existentes pode reduzir o trabalho de professores, por exemplo, para avaliação de conhecimentos. A transformação oferece aos professores um mapa conceitual inicial criado automaticamente, e ele só precisa refiná-lo, de acordo com suas necessidades, ampliando ou reduzindo-o".* (Graudina, 2008, p. 80)

Uma vez escolhido o modelo de representação semântica, foi o momento de buscar uma teoria que nos permitisse representar as relações entre os itens lexicais especializados do domínio em questão, bem como a herança conceitual lexical. A escolha recaiu sobre a Estrutura Qualia, uma das facetas do Léxico Gerativo, de James Pustejovsky (1995). O autor realiza uma distinção dicotômica para o estudo e representação da significação lexical: teorias baseadas em primitivos e teorias baseadas em relações. Pottier (1985) é um dos que trataram a semântica lexical com uma teoria de decomposição em primitivos semânticos que se opõem em positivos/negativos (possui ou não possui o sema em questão). Para Pustejovsky, contudo, uma representação semântica deve seguir uma linha composicional (que se enquadraria nas teorias baseadas em relações).

Outros modelos que estabelecem relações de significação entre itens lexicais foram observados, conforme tratado na Introdução deste artigo, contudo, acreditamos que o modelo relacional composicional adotado nos permite uma maior flexibilidade no tratamento das relações e por

estarem divididos em papéis semânticos bem definidos, conforme explicita-se na próxima seção.

3. A Teoria do Léxico Gerativo e a Estrutura Qualia

Uma visão possível para a resolução de questões inerentes ao tratamento semântico-computacional do léxico é a teoria proposta por James Pustejovsky em seu livro *The Generative Lexicon* (1995). Para o autor, os principais problemas para a semântica lexical são:

> "*(a) Explicar a natureza polimórfica da língua; (b) Caracterizar a semanticalidade de sentenças em língua natural; (c) Capturar o uso criativo de palavras em contextos novos; (d) Desenvolver uma representação semântica co-composicional mais rica*". (Pustejovsky, 1995, p. 5)

A maneira puramente morfossintática com que a maioria dos léxicos computacionais é hoje descrita pode explicar os entraves que se observam para que sistemas computacionais que necessitam do léxico funcionem adequadamente. Sem dúvida, a partir do momento que se agrega valor semântico a esses léxicos, obtêm-se resultados muito mais fiáveis e representativos concernentes àquilo que se objetiva a partir de um determinado sistema linguístico-computacional.

Para Pustejovsky, Semântica Lexical é o estudo de como e o que as palavras de uma língua denotam. Para linguistas teóricos e computacionais:

> *"o léxico é um conjunto estático de palavras-sentido, etiquetado com informações do tipo sintáticas, morfológicas e semânticas. Além disso, teorias formais do estudo da semântica de uma língua natural têm dado escassa importância a duas importantes questões: ao uso criativo de palavras em contextos novos e a uma apreciação dos modelos semântico-lexicais baseados na composicionalidade"*. (Zavaglia, 2002, p. 106 e 107)

Os componentes dessa rede de relações são classificados de acordo com o papel que desempenham, divididos da seguinte forma, conforme Pustejovsky (1995, p. 85 e 86):

- **Formal**, papel que faz a distinção do objeto em um domínio maior: i. Orientação, ii. Magnitude, iii. Forma, iv. Dimensionalidade, v. Cor, vi. Posição;
- **Constitutivo** ou **Partes Constituintes**, evidencia a relação entre objeto e suas partes constituintes que lhe são próprias: i. Material, ii. Peso, iii. Partes e elementos componentes";
- **Télico**, mostra o propósito e função do objeto: i. Propósito que um agente tem ao realizar uma ação, ii. Função integrada ou objetivo que especifica certas atividades;
- **Agentivo**, fatores que tratam da origem ou "causas" de um objeto: i. Criador, ii. Artefato, iii. Classe natural, iv. Cadeia causal

Uma abordagem do gênero, i.e. de caráter relacional, elimina entraves de natureza extensiva, pois não se limita, por exemplo, a uma lista exaustiva de traços semânticos e admite uma maior caracterização do léxico pelo próprio léxico. Sobre isso, a Teoria do Léxico Gerativo e, mais especificamente, a Estrutura Qualia, permite que se descreva um léxico valendo-se dos papéis semânticos que atribuem significado a um vocabulário finito e capturam a constituição, função, caracterização e origem dos referentes extralinguísticos que esse léxico representa no interior do sistema linguístico.

4. Metodologia da pesquisa

Antes de detalharmos o delineamento da ontologia em si, acreditamos fazer-se necessário explicitar a composição do córpus de pesquisa, a extração semiautomática dos candidatos a termos que compuseram o mapa ontológico do domínio, o levantamento de classes e subclasses, bem como do método semiautomático utilizado para o levantamento de relações semânticas segundo a Estrutura Qualia e a implementação dos dados na ferramenta Protégé.

4.1 O córpus da pesquisa

O córpus da Nanociência e Nanotecnologia (doravante N&N) foi compilado pelo Grupo de Estudos e Pesquisas em Terminologia, GETerm,[4] e apresenta 2.565.790 palavras (1057 textos, extraídos de 57 fontes diferentes), divididas tipologicamente da seguinte forma:

- Científico: composto por textos extraídos de revistas científicas, do Banco de Teses da

[4] Mais detalhes sobre a compilação do córpus podem ser obtidos em Coleti et al., 2008.

Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES), doadas por CD-ROM;

- Informativo: constituído por jornais, revistas, portais, textos publicados em sites de órgãos de fomento à pesquisa;
- Científico de Divulgação: constituído por documentos extraídos de sites especializados, revistas, da Fundação de Desenvolvimento da Pesquisa (FUNDEP);
- Técnico-Administrativo: textos retirados do portal do Ministério da Ciência e da Tecnologia brasileiro;
- Outros: formado por textos presentes em slides de apresentações, prospectos de empresas e institutos de pesquisas e demais documentos avulsos obtidos em feiras e congressos da área.

A Tabela 1 apresenta a distribuição do número de palavras por tipologia dos textos.

tipos de textos	extensão do córpus
Científico	1.846.763
Informativo	361.607
Científico de divulgação	310.018
Técnico-administrativo	26.877
Outros	20.525

Tabela 1: Número de ocorrências no córpus por tipos textuais.

4.2 Extração semiautomática de candidatos a termos

A partir desse córpus, procedemos à extração semiautomática dos candidatos a termos utilizando-se do pacote *NSP – N-gram Statistics Package* (Banerjee e Pedersen, 2003).

Por meio do pacote NSP, foi possível gerar listas de unigramas, bigramas, trigramas e tetragramas, que correspondem a termos compostos por uma, duas, três ou quatro *tokens*, respectivamente. As listas geradas pelo pacote NSP necessitaram passar por uma limpeza manual, uma vez que, muito do que foi obtido não era necessariamente um termo, como ilustrado no Quadro 1 (os candidatos a termo que foram submetidos ao especialista, neste exemplo, encontram-se em negrito).

DIFRAÇÃO<>DE<>RAIOS<>214 528 31477 436 528 214 436
DAS<>AMOSTRAS<>DE<>209 1684 1438 51641 490 923 672
A<>QUANTIDADE<>DE<>209 20683 470 51641 209 9609 470
O<>NÚMERO<>DE<>209 10266 635 51641 231 5757 613
DENSIDADE<>DE<>CORRENTE<>208 460 31477 580 436 208 373
NA<>FIGURA<>A<>207 4264 2130 9308 318 340 405
DE<>ÓXIDO<>DE<>202 21743 444 51641 249 5655 350
FILME<>DE<>ÓXIDO<>199 590 31477 384 514 199 276
DO<>CAMPO<>ELÉTRICO<>199 7131 1107 507 424 199 485
PARA<>A<>AMOSTRA<>192 5491 9247 861 1928 192 381
DA<>CONCENTRAÇÃO<>DE<>191 7434 724 51641 222 3561 601
AS<>AMOSTRAS<>DE<>190 2149 1438 51641 593 1210 672
DO<>NÚMERO<>DE<>189 7131 635 51641 189 3913 613
A<>FIGURA<>ILUSTRA<>189 20683 2130 189 1077 189 189
CEO<>-AL<>O<>189 189 294 1963 189 189 294
A<>TÉCNICA<>DE<>187 20683 405 51641 187 9609 405
A<>ADIÇÃO<>DE<>187 20683 346 51641 206 9609 327
TAXA<>DE<>CORROSÃO<>187 493 31477 705 493 187 408
TAXA<>DE<>CRESCIMENTO<>93 493 31477 362 493 93 333
CARGA<>E<>DESCARGA<>124 188 2873 168 124 124 124

Quadro 1 – Exemplo de lista de trigrama gerada pelo pacote NSP

No Quadro 1, os *tokens* encontram-se separados pelo sinal "<>", os número que se observam logo após o último sinal "<>" referem-se a frequência no córpus daquele trigrama (neste exemplo, "taxa de corrosão" ocorreu 187 vezes), os demais valores não foram utilizados nesta pesquisa.

Uma vez feita a extração e limpeza das listas geradas, essas foram submetidas à análise do especialista da área, o Prof. Osvaldo Novais de Oliveira Jr. do Instituto de Física da Universidade de São Paulo, que validou os termos e sua pertinência ao domínio em questão.

Os números de candidatos a termos obtidos imediatamente após a utilização do NSP foram muito díspares em relação ao número de termos validados pelo especialista e os que, de fato, figuram na lista final de termos, conforme a Tabela 2. Essa diferença resulta da exclusão de falsos candidatos a termos (do Quadro 1: "das amostras de", "a quantidade de", "o número de" e assim por diante) e de possíveis candidatos a termos enviados ao especialista, mas que não foram confirmados, por ele, como termos da área (é o caso de "carga e descarga" e "taxa de crescimento", do Quadro 1).

É possível afirmar que, geralmente, quanto maior o número de unidades que compõe o termo, maior o número de candidatos que são, efetivamente, termos. Isso porque o pacote NSP não utiliza nenhuma medida de associação para unigramas, apenas a medida de frequência. Nos demais casos, o pacote disponibiliza medidas de Informação Mútua, *log-likelihood* e Coeficiente *Dice* (Banerjee e Pedersen, 2003 e Almeida et al., 2003) entre outras que otimizam os resultados.

	Número de candidatos do NSP	Número final de termos
unigramas	1.081.552	1.795 (0,16%)
bigramas	314.194	587 (0,18%)
trigramas	579.491	591 (1,01%)
tetragramas	123.760	152 (1,22%)
Total	2.098.997	3.125 (0,14%)

Tabela 2: Número de candidatos a termos e número final de termos.

A Tabela 3 apresenta uma parte da lista final de trigramas já validada pelo especialista e da qual partimos para o delineamento da ontologia.

TERMOS	FREQÜÊNCIA	TIPO DE TEXTO
ABSORÇÃO DE RAIOS X	1	TA
ACETATO DE CELULOSE	4	OU
AÇO INOXIDÁVEL DUPLEX	22	CI
AEROSOL EM CHAMA	34	CI
ALARGAMENTO DO PICO	21	CI
ALGINATO DE SÓDIO	6	OU
ALTA RESOLUÇÃO ESPACIAL	22	CI
ALTURA DO PICO	20	CI
ANALISADOR DE ESPECTRO	29	CI
ANALISADOR DE REDE	21	CI
ANÁLISE TÉRMICA DIFERENCIAL	28	CI
ÁREA SUPERFICIAL ESPECÍFICA	111	CI

Tabela 3: Lista de trigramas final.

4.3 Definição de classes e subclasses

Em uma ontologia, a principal relação que se observa é a formal, mais especificamente a relação *é_um*, *é_uma* (do inglês, *is_a)* a qual representa, de maneira objetiva, a herança conceitual de uma classe por sua subclasse. Sendo assim, essa foi a primeira relação que procuramos observar para que a ontologia tivesse uma estrutura hierárquica primária. O exemplo da Figura 1 apresenta uma estrutura indicando relações *é_uma* entre a classe "microscopia eletrônica" e suas subclasses: "microscopia de varredura por sonda" herda os conceitos de "microscopia eletrônica de varredura" que, por sua vez, herda os conceitos de "microscopia eletrônica". Para "microscopia eletrônica de transmissão", esta herda também conceitos de "microscopia eletrônica", mas possui traços diferenciais em relação à "microscopia eletrônica de varredura".

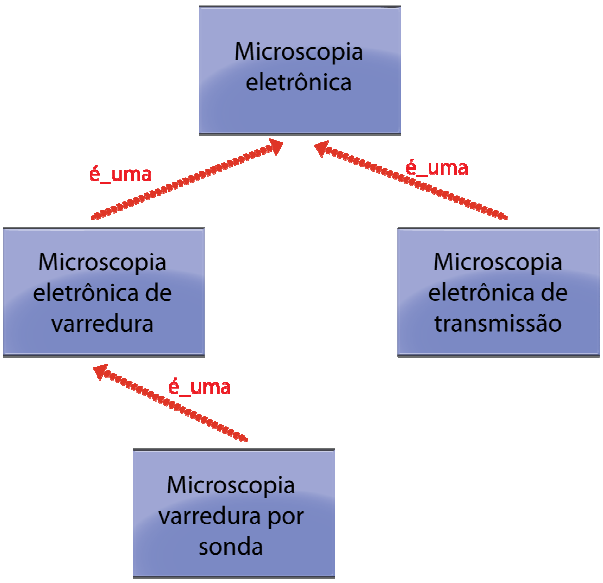

Figura 1: Classe "microscopia eletrônica" e suas subclasses.

A fim de agrupar os termos semanticamente relacionados, baseamo-nos na divisão de subdomínios feita no projeto *Desenvolvimento de uma Estrutura Conceitual (Ontologia) para a Área de Nanociência e Nanotecnologia* (Aluísio et al., 2006), para o qual havia também seis subdomínios principais:

1. "**Synthesis, Processing and Fabrication**";
2. "**Materials**";
3. "**Properties and Characterization techniques**";
4. "**Machines and Devices**";
5. "**Theories and Computational methods**";
6. "**Applications**".

Nesta pesquisa, a divisão foi realizada da seguinte forma:

1. **Aplicações**: Termos relacionados a campos científicos e usos específicos da N&N;
2. **Equipamentos**: Dispositivos utilizados na síntese, processamento e construção de nanomateriais;
3. **Materiais**: Matéria utilizada para a confecção de nanomateriais, os nanomateriais propriamente ditos ou foco de atuação de materiais nanoestruturados;
4. **Métodos e técnicas**: Processos envolvidos na manipulação de nanomatéria;

5. **Propriedades**: Características diversas intrínsecas aos materiais;
6. **Teorias**: Teorias que confluem na manipulação de materiais em nanoescala.

Assim, a classe "microscopia eletrônica", ilustrada acima, faz parte do subdomínio *Métodos e técnicas*.

A nova nomenclatura na divisão foi feita visando a facilitar o agrupamento de conceitos, além de deixar mais claro sobre o que cada subdomínio trata. Nesse sentido, indagou-se como abarcar em um mesmo subdomínio propriedades e técnicas de caracterização. Pareceu-nos que técnicas possuem mais afinidade semântica com métodos de processamento e fabricação, uma vez que, em ambos os casos, tratam-se de processos envolvidos na composição/manipulação dos nanomateriais. E ainda, "Equipamentos" engloba tanto o conceito de "máquinas" quanto o de "dispositivos" utilizados em N&N.

Ademais, a taxonomia em inglês da N&N (desenvolvida no âmbito do projeto acima citado) não corresponde propriamente a uma ontologia formalizada: alguns conceitos encontram-se agrupados em uma mesma classe, como é o caso de "Óxidos e sais", mas não se pode afirmar que as suas subclasses serão, todas elas, um óxido e ao mesmo tempo um sal.

4.4 Levantamento de relações semânticas

Esta etapa consiste na definição de relações semânticas, segundo a Estrutura *Qualia* de James Pustejovsky. Muitas das relações semânticas foram sendo delineadas concomitantemente ao processo de definição de classes e subclasses, uma vez que a observação dos contextos trazidos pelo processador de córpus já evidenciavam tais relações. Contudo, uma forma semiautomática que pudesse destacar tais relações foi útil e proveitosa, na medida em que essas são formadas, em geral, por expressões regulares. Para relações do tipo *Constitutivo*, observamos expressões como *é feito(a) de*, *é constituído(a) de/por*, *tem/têm como parte*, *é composto(a) de/por*, entre outras. Para as relações *Formal*, levantamos diversos termos a partir do subdomínio ao qual pertencem por meio de expressões de busca do tipo *é um equipamento*, *é um material*, *é uma aplicação* e assim sucessivamente para cada subdomínio eleito e elencado na seção anterior.

Visando a facilitar tal trabalho, utilizamos o recurso de *grafos* da ferramenta Unitex,[5] por meio do qual foi possível descrever um conjunto de regras recursivas de busca, permitindo assim um levantamento semiautomático de expressões que pudessem indicar relações semânticas nos quatros tipos descritos pela Estrutura Qualia (seção 3). A avaliação da eficácia do método, comparada ao número de resultados obtidos, pode, a princípio, parecer insatisfatória uma vez que muito do que obtivemos como *output* da ferramenta não foi utilizado; contudo, resultados que efetivamente foram aplicados à ontologia, após nossa análise, não teriam sido facilmente detectados, em uma busca manual, em um córpus de mais de dois milhões de palavras.

Apresentamos a seguir os *grafos* utilizados para cada um dos papéis semânticos da Estrutura Qualia, descritos na seção 3.

A Figura 2 apresenta o *grafo* utilizado para as buscas por relações do tipo *Formal*.

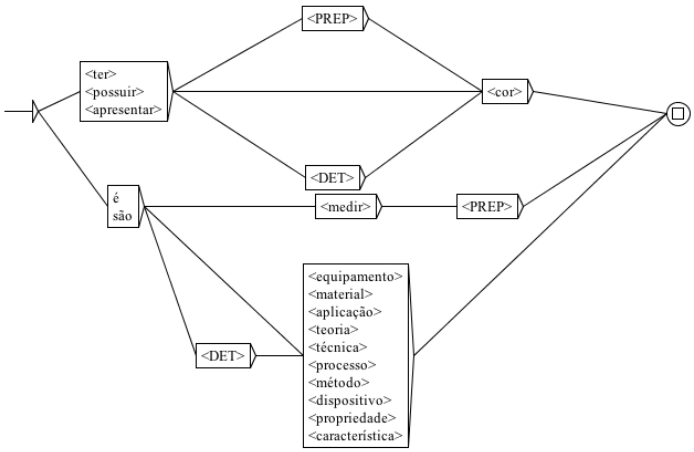

Figura 2: Grafo para busca de relações *Formal*.

O *grafo* representado na Figura 2 permite a realização de buscas que atendam aos seguintes critérios:

[5] Unitex é um sistema de processamento de corpus, baseado na tecnologia autômato-orientada. É um software criado no LADL (Laboratoire d'Automatique Documentaire et Linguistique), sob a direção de Maurice Gross. Com esta ferramenta, tem-se acesso a recursos eletrônicos, tais como dicionários e gramáticas, os quais podem ser aplicados em determinado corpus. O Unitex permite análises nos níveis da morfologia, do léxico e da sintaxe. O programa pode ser obtido gratuitamente em: www-igm.univ-mlv.fr/~unitex/.

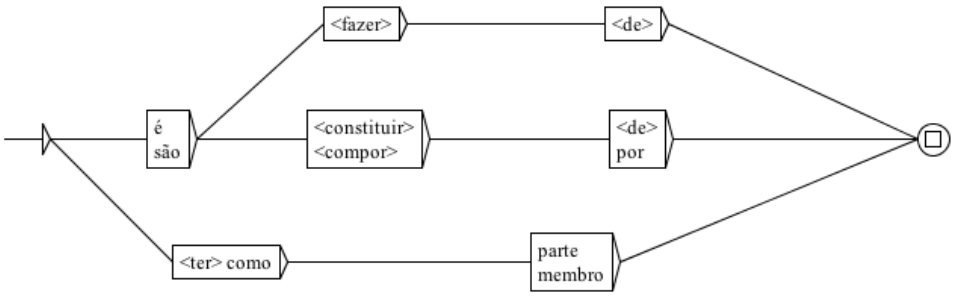

Figura 3: *Grafo* para busca de relações do tipo *Constitutivo*.

1. verbos "ter", "possuir" ou "apresentar" flexionados em qualquer tempo, modo ou pessoa seguidos de uma preposição e esta seguida pela palavra "cor" com flexão;
2. verbos "ter", "possuir" ou "apresentar" flexionados em qualquer tempo, modo ou pessoa seguidos por um determinante qualquer, seguido pela palavra "cor" com flexão;
3. "é" ou "são" seguido pelo verbo "medir" flexionado em qualquer tempo, modo ou pessoa, seguido por uma preposição;
4. "é" ou "são" seguido ou não por um determinante, seguido pelas palavras "equipamento", "material", "aplicação", "teoria", "técnica", "processo", "método", "dispositivo", "propriedade" ou "técnica" incluindo flexões dessas.

Obtivemos com essa busca 293 resultados. A título de exemplo, reproduzimos no Quadro 2 (no Anexo) algumas concordâncias para os critérios descritos no item (4), com os quais foi possível chegar a termos, ausentes até então na ontologia, como: "constante dielétrica", "perfilômetro" e "redução carbotérmica".

A Figura 3 apresenta o *grafo* utilizado para buscar relações do tipo *Constitutivo*.

Em um primeiro momento, o verbo "fazer" estava na mesma caixa dos verbos "constituir" e "compor", contudo constatamos que a combinação "fazer" seguida da preposição "por" não apontava para relações constitutivas (como *feito de*), mas para relações do tipo *Agentivo* (i.e., aquelas envolvidas na origem do objeto), como podemos observar nas concordâncias do Quadro 3 (Anexo).

O *grafo* da Figura 3 permitiu uma busca que retornou 243 resultados e atendeu aos seguintes critérios:

1. "é" ou "são" seguido do verbo "fazer" flexionado em qualquer tempo, modo ou pessoa, seguido da preposição "de", contraída com artigo ou não, e com flexão de número;
2. "é" ou "são" seguido dos verbos "constituir" ou "compor" flexionados em qualquer tempo, modo ou pessoa, seguidos da preposição "de", contraída com artigo ou não, e com flexão de número ou da preposição "por";
3. verbo "ter" flexionado em qualquer tempo, modo ou pessoa, seguido da preposição "como", seguido das palavras "parte" ou "membro"

Para as relações do tipo *Télico*, estabelecemos os seguintes critérios:

1. verbo "é" ou "são", seguido do verbo "utilizar" ou "usar" flexionado em qualquer tempo, modo ou pessoa, seguido da preposição "em" ou "para";
2. verbo "ter" flexionado em qualquer tempo, modo ou pessoa, seguido ou não da preposição "como" ou "a", seguido do substantivo "finalidade", "objetivo" ou "escopo" flexionado em número, seguido ou não da preposição "de";
3. verbo "fazer" flexionado em qualquer tempo, modo ou pessoa, seguido da palavra "uso", seguida da preposição "de";
4. verbo "utilizar" ou "usar" flexionado em qualquer tempo, modo ou pessoa, com próclise ou ênclise do pronome "se", seguido da preposição "de";
5. locução prepositiva "a fim de" ou preposição "para", seguida do verbo "obter" flexionado em qualquer tempo, modo ou pessoa ou seguida do substantivo

"obtenção", seguido ou não da preposição "de";

6. verbo "é" ou "são", seguido do verbo "fazer" flexionado em qualquer tempo, modo ou pessoa, seguido da preposição "para".

Utilizando o método aqui descrito, é possível ter um foco maior nas relações que se busca e que, numa busca manual, poderiam passar despercebidas. Os critérios descritos em (1) nos levaram às concordâncias reproduzidas no Quadro 4 (Anexo), a partir do córpus. Destacamos a última delas, que nos apontou para uma relação *Télica* importante entre os termos "óxido misto" e "coprecipitação", ilustrada na Figura 4.

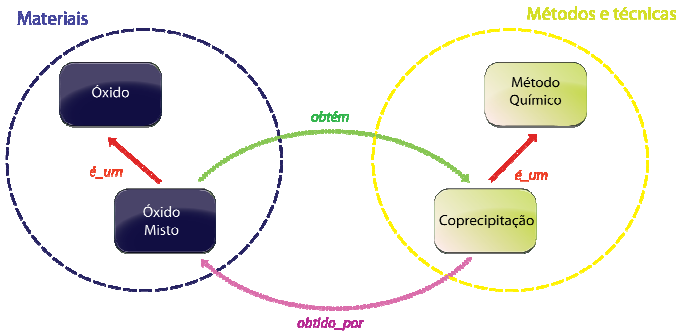

Figura 4: *Relação Télica, obtém.*

Cumpre aqui dizer que, conforme ilustrado na Figura 4, *obtém* e *obtido_por* são relações inversas, sendo *Télica* (função do objeto) e *Agentiva* (origem do objeto), respectivamente.

Da mesma forma, *utilizado_em* e *utiliza* (respectivamente, relações *Télica* e *Agentiva*) são inversas, segundo a Figura 5, em que são representados os termos "nitrogênio" (do subdomínio de "Materiais") e "secagem" (do subdomínio de "Métodos e técnicas").

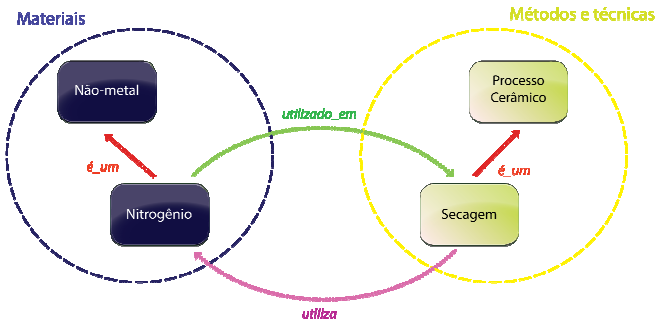

Figura 5: Relações inversas *utilizado_em* e *utiliza*.

4.5 A linguagem OWL

Neste trabalho, a linguagem adotada para a representação do domínio da N&N foi a OWL (*Web Ontology Language*), considerada atualmente, o padrão mais corrente para a representação de informações ontológicas na Web. A OWL (Smith et al., 2004) foi antecedida pelas linguagens RDF (*Resource Description Framework*) e RDFS (*RDF-Schema*), mostrando-se mais potente em termos de descrição e instanciação. Essas duas últimas correspondem a linguagens em que os recursos são descritos como trios de objetos-atributos-valores, semelhantes ao sujeito-verbo-objeto das redes semânticas.

4.6 Implementação dos dados na ferramenta Protégé.

A implementação dos resultados alcançados em uma ferramenta computacional específica para ontologias garante que os formalismos adotados para a representação do domínio escolhido sejam respeitados. Além disso, as possibilidades existentes de reuso de uma ontologia, quando expressa em uma linguagem computacional corrente e atual, são variadas. Nesse sentido, buscamos utilizar um software que possuísse facilidade de uso aliada a potencialidades de funções. A escolha incidiu sobre a ferramenta Protégé[6] (Noy *et al.*, 2000), uma vez que atende a esses quesitos.

Em consonância com os princípios de construção de ontologias, a ferramenta permite que ontologias sejam constantemente alimentadas e representadas em diferentes formatos e linguagens. Segundo Noy *et al.* (2001, p. 62), a ferramenta possui: um "modelo de conhecimento extensível", sendo possível redefinir seus primitivos representacionais; um "formato de arquivo de saída customizável", o que permite gerar arquivos em qualquer linguagem formal; "uma interface com o usuário customizável", possibilitando adaptar os componentes da interface com o usuário para a nova linguagem escolhida; "uma arquitetura extensível que permite integração com outras aplicações", isso torna a ferramenta conectável a módulos semânticos externos.

[6] Desenvolvida pela Divisão de Informática Médica do Departamento de Medicina da Universidade de Stanford, o Protégé foi inicialmente idealizado para modelar o domínio da medicina e traçar relações entre os muitos conceitos que englobam tal campo de especialidade. A ferramenta encontra-se disponível para download gratuitamente em http://protege.stanford.edu/

A representação do conhecimento no Protégé se dá por meio de três entidades básicas:

- *Classes* – define conceitos no domínio;
- *Propriedades* (*Properties*) – define atributos das classes;
- *Facetas* (*Facets*) – define restrições nos valores de classes (por exemplo: tipos, cardinalidade,[7] padrões).

Essa ferramenta permite a definição de propriedades inversas (Figura 6), o que facilita a representação de questões como aquelas ilustradas pelas Figuras 4 e 5.

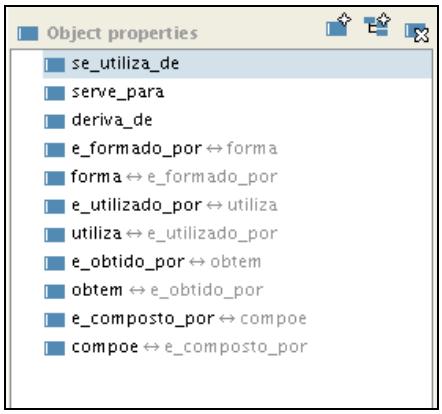

Figura 6: Relações semânticas representadas como Propriedades.

6. Resultados

O subdomínio que apresentou mais classes e subclasses foi o de *Materiais* (230), seguido de *Métodos e Técnicas* (68), *Propriedades* (42), *Equipamentos* (18), *Aplicações* (17) e *Teorias* (3), totalizando 378 classes e subclasses.

Estruturalmente, uma classe como "Material nanoestruturado" (do subdomínio de "Materiais") e suas subclasses podem ser assim representadas, de acordo com a figura 7.

Com o auxílio do *plugin* OWLViz, essa mesma classe e suas subclasses ficam ilustradas, conforme a Figura 8.

Algumas das relações semânticas identificadas ao longo do processo foram implementadas na ferramenta. Elas encontram-se elencadas, quantificadas e exemplificadas na Tabela 4.

7 A cardinalidade diz respeito a um dado expresso em valor numérico ou por um conjunto deles.

Formal		
é_medido_em	43	<Nanoporo> é_medido_em <Nanômetro>
é_um, é_uma	376	<Densidade> é_uma <Grandeza_física>
Constitutivo		
compõe	1	<Carbono> compõe <Nanotubo_de_carbono>
é_composto_por	8	<Vitrocerâmica> é_composto_por <Cálcio>
é_feito_de	4	<Nanotubo de carbono> é_feito_de <Carbono>
é_formado_por	3	<Nanocompósito polimérico> é_formado_por <Borracha>
forma	6	<Quinona> forma <Nanocápsula>
Télico		
obtém	2	<Precursor_polimérico> obtém <Óxido_de_estanho>
produz	2	<Bactéria> produz <Antígeno>
utilizado_em	5	<Vidro> utilizado_em <Vitrocerâmica>
Agentivo		
deriva_de	4	<Plástico> deriva_de <Petróleo>
é_produzido_por	2	<Antígeno> é_produzido_por <Vírus>
obtido_por	7	<Óxido_de_estanho> obtido_por <Precursor_polimérico>
originado_de	1	<Vitrocerâmica> originado_de <Vidro>
utiliza	2	<Fotoalinhadora> utiliza <Luz_ultravioleta>

Tabela 4: Relações individuadas a partir das buscas na ferramenta Unitex.

5. Discussões

A Estrutura Qualia permite-nos ter um maior controle sobre as relações semânticas do domínio da N&N, uma vez que os delimita em quatro papéis funcionais. O método que aqui se motiva é um primeiro passo para trabalhos terminológicos que fazem uso de grandes córpus. As relações pré-estabelecidas podem não cobrir todas as relações que podem figurar no domínio mas já apontam para aquelas fundamentais. Além disso, os *grafos* podem ser ampliados e adaptados de acordo com as necessidades de cada pesquisa. Uma vez individuadas as relações básicas do tipo *Formal* *é_um*, *é_uma* (presentes em qualquer ontologia), as demais podem ser estendidas partindo-se das listas geradas pelo Unitex.

Delineamos, neste trabalho, a área técnico-científica da N&N, uma ciência interdisciplinar e inovadora, cujas técnicas de manipulação de materiais têm obtido investimentos enormes e cujas possibilidades de aplicação são inúmeras. A definição de sua estrutura conceitual permitirá que o produto terminográfico seja coeso e uniforme. Por outro lado, essa mesma estrutura, quando dotada de formalismos, pode também servir como um léxico computacional que sirva para alimentar sistemas de PLN.

A observação dos fenômenos linguísticos por meio de um processador de córpus ressalta a importância desse tipo de ferramenta e a necessidade de automatização das pesquisas em Linguística, de um modo geral. O alto nível de conhecimento do uso dessas ferramentas, por parte do pesquisador, aprimora os resultados da pesquisa e permite uma adaptação dessas ferramentas às necessidades particulares de cada investigação científica.

A importância da utilização de métodos computacionais é grande, na medida em que o volume de informações, com que muitos trabalhos científicos se deparam tem sido cada vez maior. Nesta pesquisa, a extração semiautomática de termos e o levantamento de candidatos a relações semânticas mostraram-se um fim cujos meios para alcançá-los foram enormemente facilitados pelo auxílio de recursos informatizados. Entretanto, a observação cautelosa e criteriosa desses dados por parte do pesquisador foi o elemento-chave para que chegássemos aos resultados esperados.

Os recursos aqui descritos encontram-se disponíveis no Portal de Ontologias OntoLP.[8]

6. Conclusões

A Engenharia Ontológica é um vasto campo a ser explorado por pesquisadores de disciplinas diversas que têm estudado e aplicado, cada vez mais, seus conhecimentos na criação de uma metodologia que permita a criação e reuso de ontologias. Há, nessas disciplinas distintas, conceitos que se interpolam e se confundem, permitindo que se trate de conceitos relativos às ontologias de maneiras diversas e complementares. Aquilo que a Computação entende por ontologias, os formalismos que ela adota para sua criação e manipulação beneficiam o poder de descrição semântica de um dado vocabulário por parte de um lexicólogo/terminólogo, conferindo-lhe também a possibilidade de realizar uma aplicação computacional para seu trabalho, se assim desejar. Logo, podemos afirmar que tais formalismos garantem um processo definitório mais consciente, uma vez que, para o tratamento informático do léxico, as ambiguidades, inconsistências e imprecisões devem ser minimizadas. Para tanto, deve-se ter à disposição um modelo semântico eficiente que estenda a exposição lexical a um nível superior ao da morfologia e da sintaxe fornecendo à máquina condições de inferir e interpretar dados linguísticos. A esse propósito, a Estrutura Qualia representa um método eficaz para uma representação semântica inter-relacional, e garantiu a esta pesquisa meios de estabelecer relações de tipologias diversas ao léxico em questão, permitindo que a sua semântica fosse exposta e computacionalmente tratável.

Embora a N&N seja uma área de especialidade multidisciplinar que se utiliza de conceitos e técnicas da Física, Química, Biologia, Medicina, Engenharia de Materiais e áreas afins, o que percebemos é que pesquisadores em N&N têm criado novos materiais (em sua maioria, aqueles em escala nanométrica) e esses devem ser nomeados. Procuramos, dessa forma, estudar também esses novos termos e os processos aí envolvidos. Destacamos, assim, a partícula *nano-* como formadora desses itens neológicos especializados e os métodos envolvidos nesse levantamento.[9]

Salienta-se ainda que os resultados alcançados podem ser estendidos a partir do modelo proposto. As 361 classes e subclasses apresentadas representam o domínio da N&N, mas não integralmente. Essa delimitação deve-se, em primeiro lugar, à extensão do domínio e, posteriormente, ao grande tempo requerido por tarefas como:

- resgate de conceitos;
- observação das diversas ocorrências de um mesmo termo;
- correlações com termos semelhantes;
- real estatuto de termo de determinadas lexias;
- identificação do equivalente em português, em casos nos quais preferiu-se pelo uso de um termo estrangeiro;
- pertinência de um termo a duas superclasses distintas – em qual delas o termo estaria melhor representado e de qual superclasse há uma herança conceitual mais clara?

[8] Acessível em http://www.inf.pucrs.br/~ontolp/index.php.

[9] Uma análise dos processos de formação neológica no domínio da N&N pode ser encontrada em Kasama et al. (2008).

Esses são alguns exemplos de dificuldades encontradas no desenvolvimento da pesquisa aqui relatada, mas que lhe são inerentes.

As contribuições deste trabalho fazem-se sentir em áreas como:

- a Linguística: por meio do estabelecimento de uma metodologia, fundamentalmente embasada em ferramentas computacionais, que permite a observação de termos em uso e sua estruturação a partir de critérios semânticos;
- as Ciências da Computação: que se beneficia de conceitos linguísticos no seu fazer e pode reaproveitar os resultados obtidos para avaliação e uso real de uma ferramenta computacional que se sirva de informações semânticas;
- a área de N&N: cuja sistematização vocabular permite que pesquisadores da área possuam uma fonte de referência no que tange suas práticas. Ademais, a recolha de termos em língua portuguesa, variante brasileira, contribui para o desenvolvimento da área no país;
- o ensino em geral: seja para alunos de graduação ou pós-graduação em cursos afins à N&N, mas também para alunos de Ensino Médio, tendo em vista que a multidisciplinaridade da N&N promove o conhecimento da Física, da Química e da Biologia.

As possibilidades iniciadas neste trabalho vão além daquilo que obtivemos. Esperamos que a ontologia ora proposta auxilie, de fato, no processo de elaboração do dicionário de N&N em língua portuguesa do Brasil, mas também que possa ter utilidade e aplicação real em sistemas de PLN.

Além da elaboração da ontologia em si, esperamos ter proposto uma metodologia para a elaboração de novas representações do conhecimento valendo-nos de preceitos observados na Linguística de Córpus, na Terminologia e nos formalismos computacionais que buscamos seguir.

Agradecimentos

Agradecemos à Fundação de Amparo à Pesquisa do Estado de São Paulo (FAPESP), pelo financiamento da pesquisa (Processo nº 06/59144-8), aos Professores Oto Araújo Vale, Sandra Maria Aluísio e Maria Cristina Parreira pela leitura e valiosas contribuições ao trabalho; ao Professor Osvaldo Novais de Oliveira Jr., do Instituto de Física da Universidade de São Paulo, pela consultoria técnica na área, sem a qual um trabalho desta natureza não poderia ser realizado; e, por fim, aos Professores António Teixeira e Patrícia Cunha França pelas leituras finais e sugestões que enriqueceram este artigo.

Referências

Almeida, Gladis Maria de Barcellos. 2000. Teoria Comunicativa da Terminologia: uma aplicação. Araraquara (Tese de doutorado).

Almeida, Gladis Maria de Barcellos; Aluísio, Sandra Maria; Teline, Maria Fernanda. 2003. Extração manual e automática de terminologia: comparando abordagens e critérios. In: 1o. Workshop em Tecnologia da Informação e da Linguagem Humana, 2003, São Carlos. Anais do TIL'2003.

Alsumait, Loulwah; Wang, Pu; Domeniconi, Carlota; Barbará, Daniel. Embedding semantics in LDA topic models. 2010. In: Berry, Michael W.; Kogan, Jacob. Text Mining: Application and Theory. John Wiley & Sons, Ltd., p. 183-203

Aluísio, Sandra Maria; Oliveira Jr., Osvaldo Novais; Almeida, Gladis Maria de Barcellos; Nunes, Maria das Graças Volpe; Oliveira, Leandro Henrique Mendonça de; Felippo, Ariani Di; Antiqueira, Lucas; Genoves Jr, Luiz Carlos; Caseli, Luciano; Zucolotto, Valtencir ; Santos Jr., David Sotero dos. 2006. Desenvolvimento de uma estrutura conceitual (ontologia) para a área de Nanociência e Nanotecnologia. (Relatório técnico)

Banerjee, Satanjeev; Pedersen, Ted. 2003. The Design, Implementation, and Use of the Ngram Statistics Package In: Conference On Intelligent Text Processing And Computational Linguistics, 4., 2003, Cidade do México. Proceedings…, Cidade do México, p. 370-381.

Berners-Lee, Tim; Hendler, James; Lassila, Ora. 2001. The Semantic Web. Scientific American. p. 35-43.

Biderman, Maria Tereza Camargo. 1996. Léxico e vocabulário fundamental. Alfa. São Paulo, v.40, p. 27-46.

Cabré, Maria Tereza. 1999. La terminología. Representación y comunicación. Barcelona: IULATERM.

Cabré, Maria Tereza. 2003. Theories of terminology: their description, prescription and explanation. Terminology, v.9, n.2, p.163-200.

Coleti, Joel S.; Mattos, Daniela F.; Genoves Jr., Luiz Carlos; Candido Jr., Arnaldo; Di Felippo, Ariani; Almeida, Gladis Maria de Barcellos;

Aluísio, Sandra M.; Oliveira Jr., Osvaldo Novais. 2008. A compilação de corpus em língua portuguesa na área de nanociência/nanotecnologia: problemas e soluções. In: Tagnin, Stella E. O.; Vale, Oto Araújo (Org.). Avanços da Lingüística de Corpus no Brasil. 1 ed. São Paulo: Humanitas, p. 167-191.

Dorr, Bonnie J. 1992. The use of lexical semantics in interlingual machine translation. v.7, n.3, Springer Netherlands, p. 135-193.

Dorr, Bonnie J. 1993. Machine Translation: a view from the lexicon. Cambridge: MIT Press.

Ercan, Gonenc; Cicekli, Ilyas. 2008. Lexical Cohesion Based Topic Modeling for Summarization. Lecture Notes in Computer Science. v. 4919, p. 582-592.

Fox, Edward A. 1980. Lexical relations: Enhancing effectiveness of information retrieval systems. SIGIR Forum, v.15, n.3, p. 5-36.

Graudina, Vita. 2008. OWL Ontology Transformation into Concept Map. Scientific Proceedings of Riga Technical University. 5th Series, Computer Science, Applied Computer Science, Vol. 34, 79-90.

Gruber, Tom. 2008. Ontology. In: Liu, Ling; Özsu, M. Tamer (Eds.) Encyclopedia of Database Systems, v. 1, Springer-Verlag.

Guarino, Nicola. 1998. Formal Ontology in Information Systems. Proceedings of FOIS'98, Trento, Itália, 6-8 Junho 1998. Amsterdam, IOS Press, p. 3-15.

Hutchins, W. John; Somers, Harold L. An introduction to machine translation. London: Academic Press, 1992.

Jackendoff, Ray. 1983. Semantics and cognition. Cambridge: The MIT Press.

Kasama, Deni Y.; Almeida, Gladis Maria de Barcellos; Zavaglia, Claudia. 2008. A influência das novas tecnologias no léxico: processos de formação neológica no domínio da nanociência e nanotecnologia. Debate Terminológico, v. 4, p. 3.

Katz, Jerrold J.; Fodor, Jerry A. 1963. The Structure of a Semantic Theory. Language, v. 39, n. 2, p. 170-210.

Noy, Natalya F.; Sintek, Michael; Decker, Stefan; Crubézy, Monica; Fergerson, Ray W.; Musen, Mark A. 2001. Creating Semantic Web Contents with Protégé-2000. IEEE Intelligent Systems, v. 16, n. 2, p. 60-71.

Pottier, Bernard. 1985. Linguistique Générale: théorie et description. 2. ed. Paris: Éditions Klincksieck.

Pustejovsky, James. 1995. The Generative Lexicon. Cambridge: The MIT Press.

Reeve, Lawrence H.; Han, Hyoil. 2007. The Use of Domain- Specific Concepts in Biomedical Text Summarization. Information Processing and Management, v.43, n.6, p. 1765–1776.

Rino, Lúcia Helena Machado; Pardo, Thiago Alexandre Salgueiro (2003). A Sumarização Automática de Textos: Principais Características e Metodologias. In: Anais do XXIII Congresso da Sociedade Brasileira de Computação, Vol. VIII: III Jornada de Minicursos de Inteligência Artificial, p. 203-245.

Zavaglia, Claudia. 2002. Análise da homonímia no português: tratamento semântico com vistas a procedimentos computacionais. Araraquara (Tese de doutorado).

3.14. Material nanoestruturado / Nanomaterial
3.14.1. Material nanoestruturado bidimensional / Nanomaterial bidimensional
3.14.1.1. Filme fino / Poço quântico
3.14.2. Material nanoestruturado unidimensional / Nanomaterial unidimensional
3.14.2.1. Fio quântico
3.14.2.1.1. Nanotubo
3.14.2.1.1.1. Nanotubo de carbono
3.14.2.1.1.1.1. Nanotubo de carbono de parede múltipla/ Nanotubo de carbono de múltiplas paredes
3.14.2.1.1.1.2. Nanotubo de carbono de parede simples / Nanotubo de carbono de parede única
3.14.2.1.2. Nanofio
3.14.2.1.3. Nanofita
3.14.2.2. Nanobastonete
3.14.3. Material nanoestruturado zero dimensional / Nanomaterial zero-dimensional
3.14.3.1. Nanofibra
3.14.3.1.1. Nanofibra de carbono
3.14.3.2. Nanopartícula
3.14.3.2.1. Nanopartícula de hidrogel
3.14.3.2.2. Nanopartícula de metal
3.14.3.2.2.1. Nanopartícula de ferrita
3.14.3.2.2.2. Nanopartícula de ferro
3.14.3.2.2.3. Nanopartícula de ouro
3.14.3.2.2.4. Nanopartícula de prata
3.14.3.2.3. Nanopartícula de ni
3.14.3.2.4. Nanopartícula de óxido
3.14.3.2.5. Nanopartícula de semicondutor
3.14.3.2.6. Nanopartícula de sílica
3.14.3.2.7. Nanopartícula polimérica
3.14.3.2.7.1. Nanocápsula / Lipossoma
3.14.3.2.7.2. Nanoesfera
3.14.3.3. Ponto quântico / Quantum dot
3.14.3.3.1. Nanocristal
3.14.4. Material nanoporoso
3.14.5. Nanocompósito
3.14.5.1. Nanocompósito cerâmico / Nanocompósito de matriz cerâmica
3.14.5.2. Nanocompósito polimérico / Nanocompósito de matriz polimérica
3.14.5.3. Nanoporo
3.14.6. Nanohélice
3.14.7. Nanoimã
3.14.8. Nanomola
3.14.9. Nanomotor
3.14.10. Nanorobô
3.14.11. Nanorotor
3.14.12. Nanossensor

Figura 7: Classe “Material nanoestruturado” e suas subclasses.

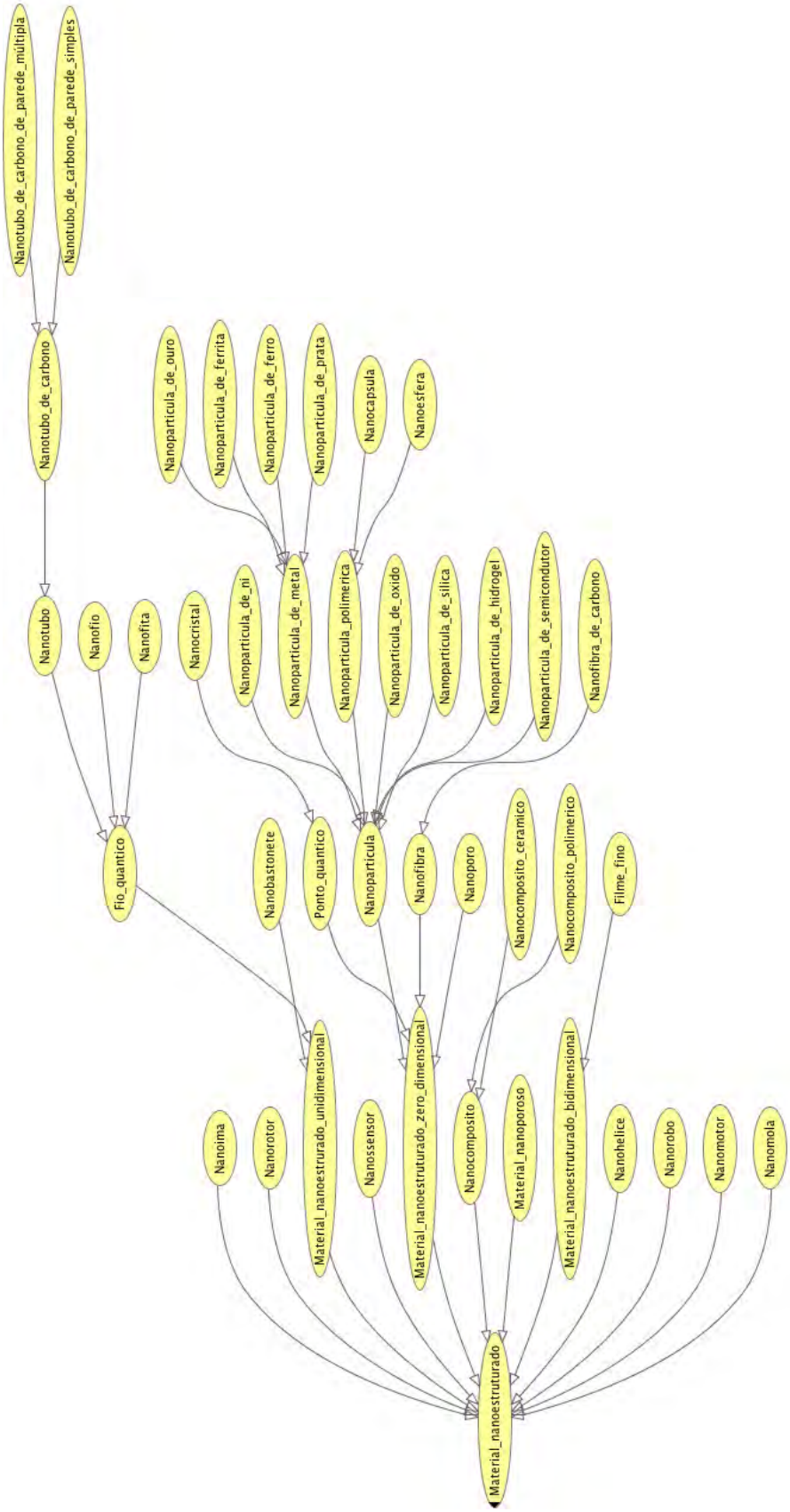

Figura 8: Classe “Material nanoestruturado” e suas subclasses.

Anexo: Concordâncias

Reproduzimos aqui, a título de exemplo, algumas das concordâncias geradas pelo Unitex segundo os critérios de busca descritos na seção 4.4.

```
eja realmente na superfície do material é a aplicação de um "coating" (recobrimento) do aditivo recobri
] {S}O objetivo desta etapa do processo é a aplicação de um filme uniforme de fotorresiste sobre o subs
létrica. {S}A constante dielétrica (k') é a propriedade determinante da capacitância do circuito, sendo
James (2005) a resistência ao glifosato é a propriedade mais frequente nestes cultivos, presente em 72%
izadas no projeto.{S} A fonte principal é a Teoria do Controle Ótimo, que aborda entre outros fatores,
izadas no projeto.{S} A fonte principal é a Teoria do Controle Ótimo, que aborda entre outros fatores,
droga, que teve sua fórmula patenteada, é material de estudo do doutorando Raul Ribeiro, orientado pela
ndutores.{S} A hidroxiapatita sintética é material inorgânico composto por fosfato de cálcio que tem si
a dimensão atómica.{S} O sistema de MBE é um equipamento sofisticado.{S} Equipamentos mais versáteis po
rredura por tunelamento (STM).{S} O STM é um equipamento sofisticado e de uso dedicado, permitindo um o
rredura por tunelamento (STM).{S} O STM é um equipamento sofisticado e de uso dedicado, permitindo um o
pa 2: {S}? O precipitador eletrostático é um equipamento apto para a remoção de nanopartículas, obtendo
}B.3.{S} Perfilômetro {S}O perfilômetro é um equipamento de medida mecânica de perfis ou topologia de f
ubstrato de silício. {S}Metalização:{S} É um método de deposição de um filme de metal que pode ser feit
am que o método de redução carbotérmica é um método viável para o crescimento de nanoestruturas unidime
lhos demonstrando que a moagem mecânica é um método eficiente de obtenção de espinélios LiMn2O (KOSOVA
rabalhos demonstram que moagem mecânica é um método eficiente para controle das característica morfológ
 de CVD [17]. {S}- A Implantação Iônica é um método de modificação superficial no qual um feixe de íons
ipais componentes.{S} Em síntese, a PCA é um método que tem por finalidade básica a redução de dados a
m a carne morta em tecido vivo.{S} Esse é um método para vencer a morte e promover a ressurreição dos s
. "O que mais nos entusiasma é que este é um método modular de montagem que irá nos permitir juntar pra
e o método de redução carbotérmica, que é um método na qual os óxidos são misturados com carbono para p
conhecido por redução carbotérmica, que é um método de simples utilização, mas que não tem sido muito e
Image @2005 AIST. {S}A tecnologia LIBWE é um método de uma etapa para a microfabricação de placa de vid
o por Lift-off [01, 14, 22] {S}Lift-off é um método simples que é muito utilizado na definição de linha
```

Quadro 2: Expressões que apontam para subdomínios.

```
esquisa, assim como a de pesquisadores, é feita por uma quantificação mais abrangente, o número de arti
an a decomposição da radiação espalhada é feita por meio de grades d difração, enquanto que no espalham
 substrato. {S}A desidratação da lâmina é feita por evaporação, pelo aquecimento do substrato em uma es
dores, enquanto que a avaliação de água é feita por análise química em laboratório e são bastante demor
dores, enquanto que a avaliação de água é feita por análise química em laboratório e são bastante demor
 potável. {S}A regeneração do nanofilme é feita por aquecimento do material. {S}Também é possível se ob
ente, a avaliação do sabor dos produtos é feita por pessoas especialmente treinadas, que analisam senso
gravação, leitura e desgravação do bits é feita por agulhas do tipo usado em microscópios de varredura
ssim como de todos os programas do PPA) é feito por meio do sistema de informações gerenciais do Minist
500.{S} O controle da pressão na câmara é feito por um sensor Pirani Balzers modelo TPR250, os fluxos d
ntrada/saída especificados.{S} O ajuste é feito por ciclos em que a cada entrada apresentada à rede os
ão, menos de 15% do gasto privado total é feito por empresas com menos de 250/300 empregados.{S} O mesm
forme. {S}Por isso, as medidas de cores são feitas por meio de métodos espectrais.{S} Neste caso, o equ
m tempo e custo elevados, e as análises são feitas por amostragem ao invés de medidas em tempo real. {S
rios com alguns nanômetros de diâmetro, são feitos por feixes ("jatos") de elétrons, obtidos de um micr
```

Quadro 3: Relação “é/são” <fazer> “por” denota relação *Agentiva*.

```
m redox/eletrodo/vidro/camada espelhada é usada em espelhos eletrocrômicos automotivos (industr
icas). {S}Hoje em dia, essa mesma idéia é usada em computadores de alto desempenho, com micropr
. {S}Nanotecnologia {S}A nanotecnologia é usada em cosméticos para trazer vantagens sensoriais
m redox/eletrodo/vidro/camada espelhada é usada em espelhos eletrocrômicos automotivos (industr
icas). {S}Hoje em dia, essa mesma idéia é usada em computadores de alto desempenho, com micropr
. {S}Nanotecnologia {S}A nanotecnologia é usada em cosméticos para trazer vantagens sensoriais
e a baixa pressão, hidrogênio molecular é usado em abundância na alimentação do gás para gerar
itrofenil-ß-D galactopiranosídeo (ONPG) é usado para detectar a enzima ß-D-galactosidase, a qua
metil-umbeliferil- ß glicuronídeo (MUG) é usado para detectar a enzima ß-glicuronidase, a qual
m chamado de BOE (Buffered Oxide Etch), é usado para corroer {S}SiO2 (óxido de silício) e SiNx
, Anritsu MS2601B. {S}A terminação (9d) é utilizada para a observação da presença da linha Bril
 fibra, (FBG - Fiber Bragg {S}Grating), é utilizada para refletir os campos ópticos chegando ao
4.3.2, extraída da Lei de Lambert-Beer, é utilizada para correlacionar a intensidade (I), a abs
ologia será baseada no esquema que hoje é utilizado em computação quântica com ressonância magn
of microparticles - LAM", Juang (1994), é utilizado para fabricação de nanopartículas em pequen
lador acusto-óptico, Intra-Action ME40, é utilizado para induzir um desvio conhecido na frequên
stão interconectados, o modelo de M. S. é utilizado para predizer a taxa de 1 densificação (ma
onente é denominada coprecipitação, que é utilizada para a obtenção de óxidos mistos, pois, per
```

Quadro 4: Busca por relações *Télicas* utilizando os verbos “utilizar” e “usar”.