A semântica dos compostos nominais em língua inglesa: um estudo de *corpus*[1]

Rove Luiza de Oliveira Chishman (UNISINOS)
Lílian Figueiró Teixeira (UNISINOS)

RESUMO: Neste trabalho, investiga-se a semântica dos compostos nominais em inglês formados por dois substantivos (compostos NN). O objetivo é identificar as relações mais frequentes em um *corpus* formado por dez edições da revista *National Geographic*, a fim de se propor uma tipologia que expresse a composicionalidade semântica dessas construções. Com o auxílio de recursos da Linguística de *Corpus*, chegou-se a um total de 4.693 possíveis compostos. As relações semânticas mais frequentes entre os dois elementos dos compostos presentes no *corpus* são: telicidade, agentividade, meronímia, localização, posse e hiponímia.

Palavras-chave: Linguística de *Corpus*; Semântica lexical; Compostos nominais; *Frames* semânticos.

Introdução

A semântica dos compostos nominais em inglês formados por dois substantivos tem sido um tema de interesse para as investigações sobre o Processamento Automático de Línguas Naturais (PLN), basicamente porque fazer com que os sistemas reconheçam tais construções, formadas por duas palavras, como uma unidade constitui um grande desafio. Diferentemente de outras línguas, como o português, em que a relação entre dois substantivos é expressa através de alguma outra palavra, como uma preposição (*pano de prato*), em inglês, um substantivo pode ser anteposto a outro para formar um composto (*bus stop*), sem outra palavra que possa sinalizar a relação entre esses dois elementos. Além disso, as construções em inglês formadas por dois substantivos são extremamente numerosas e compostos novos desse tipo podem ser facilmente criados.
Como exemplos de trabalhos que se ocupam desta temática, pode-se citar o de Ó Séaghdha (2007), que tem como objetivo utilizar um *corpus* de treinamento e a partir de métodos probabilísticos etiquetar automaticamente as relações semânticas entre os elementos de compostos em um *corpus*. Copestake (2003), por sua vez, utiliza uma abordagem

[1] Este trabalho é baseado na dissertação de mestrado "A semântica dos compostos nominais – um estudo de *corpus* paralelo inglês/português" de Lílian Figueiró Teixeira, orientada por Rove Luiza de Oliveira Chishman, UNISINOS, 2009.

VEREDAS ON-LINE – LINGUÍSTICA DE CORPUS E COMPUTACIONAL – 2/2009, P. 84-99 – PPG LINGUÍSTICA/UFJF – JUIZ DE FOR A – ISSN 1982-2243

composicional para os compostos nominais, com o propósito de integrar esses dados em uma gramática (um sistema de dados computacionais) da língua inglesa. Johnston e Busa (1999) comparam os compostos em inglês e italiano, buscando identificar padrões semânticos entre os elementos que constituem essas estruturas através de uma interpretação a partir da estrutura *qualia* de Pustejovsky (1995). Em português, destaca-se Smarsaro (2004), que se ocupa do estabelecimento de critérios formais para a identificação de uma expressão N de N como um composto, de forma que esses critérios possam auxiliar o processamento automático dessas estruturas.

O propósito deste estudo é analisar a semântica dos compostos nominais (doravante compostos NN[2]); em outras palavras, interessa aqui compreender como os elementos dos compostos em língua inglesa se relacionam semanticamente. Os compostos nominais analisados são as sequências de dois substantivos separados por espaço. Para tal, seguem-se os princípios da Linguística de *Corpus* (LC), priorizando dados reais da língua, e parte-se para a construção de um *corpus* composto por dez edições da revista *National Geographic*, em sua versão eletrônica.

Para alcançar tal meta, discutem-se inicialmente alguns conceitos básicos acerca da temática. Esses conceitos são apresentados na próxima seção, que é dividida em duas subseções conforme a base teórica: a Teoria do Léxico Gerativo (PUSTEJOVSKY, 1995) e a teoria de *templates* e *frames* (RYDER, 1994, FILLMORE, 2006). Utilizando-se os recursos da LC, foram extraídos os candidatos a compostos do *corpus*, do qual 200 compostos foram selecionados para a análise. Por fim, tem-se as relações semânticas identificadas nos compostos nominais do *corpus* coletado, assim como sua frequência.

Espera-se com este estudo aprofundar a reflexão sobre a semântica dos compostos NN, contribuindo para a Linguística Computacional, em especial para o desenvolvimento de programas que se valham de anotação semântica.

1. Compostos nominais

Nesta seção, são apresentados alguns conceitos pertinentes ao estudo dos compostos, sendo estes: produtividade, transparência semântica, nuclearidade, composição exocêntrica e endocêntrica, lexicalização e nominalização. Também são expostas diferentes perspectivas quanto à definição de compostos nominais e algumas de suas principais características.

Há diversos critérios que podem ser utilizados para a identificação de dois substantivos como compostos. Para este trabalho, consideramos dois, o sintático e o semântico.

Levando-se em conta as considerações sintáticas, pode-se dizer que os compostos comportam-se como palavras, e que, dessa forma, não é possível inserir outra palavra entre os dois elementos. Além disso, não é possível modificar parte da palavra sem modificar o composto todo. Critérios como esses são adotados por Adams (1973 *apud* RYDER, 1994) e Bloomfield (1933 *apud* RYDER, 1994).

Optamos por priorizar o critério semântico, que se baseia na noção de unidade referencial, ou seja, tem-se duas palavras, mas apenas um referente. Em caso de dúvida, podem ainda ser aplicados critérios sintáticos, como a impossibilidade de inserção de outra palavra entre os dois elementos e o fato de que não é possível modificar apenas parte da palavra sem modificar o todo. Esses critérios parecem interessantes, pois reforçam a unidade referencial.

Os compostos são produtivos porque novas combinações de palavras em contextos variados podem ser criadas. Ainda assim, apresentam características idiossincráticas, já que não é

[2] Essa denominação vem do inglês, em que os compostos formados por dois substantivos são chamados de *Noun-Noun Compounds*; por isso, utiliza-se de forma abreviada a expressão compostos NN.

possível estabelecer padrões quanto à sua produção. Esse é o principal problema relacionado aos compostos nominais, o que justifica não apenas o interesse de muitos pesquisadores por essa temática, como também a falta de consenso no que tange ao seu tratamento teórico. É difícil criar regras, estabelecer padrões, pois o fenômeno é idiossincrático, mas também não é possível criar um léxico com todos os compostos, haja vista o número de novos casos que surgem. No entanto, a partir do reconhecimento de graus de produtividade ou de semi-produtividade em algumas construções, é possível se pensar em algum tratamento para os compostos.

Segundo Vale (2001), se uma expressão apresenta transparência semântica, o seu significado é obtido a partir da soma do significado dos seus componentes, o que indica uma forte ligação entre transparência semântica e produtividade. Smarsaro (2004) complementa o raciocínio de Vale, quando destaca que uma palavra é composicional quando é produtiva e apresenta transparência semântica. A relação entre produtividade e composicionalidade se justifica, pois os elementos que formam uma expressão composicional são livres e podem dar origem a outras construções, sendo desta forma também produtivos. Quando isso não é possível, dizemos que a palavra não é composicional. Jespersen (1909 *apud* RYDER, 1994) considera compostos somente as expressões que não são transparentes semanticamente, como *blackmail*[3] e *honey moon*. Uma construção transparente como *glass door*, uma porta que é feita de vidro, para o autor, não seria um composto. Como o tema de estudo são os compostos NN em geral, busca-se alguma forma de interpretar tanto as construções transparentes quanto as opacas, ambas consideradas compostas, seguindo Copestake (2003) e Ryder (1994), para quem as combinações de dois substantivos separados por um espaço podem constituir compostos.

Os compostos em língua inglesa são geralmente formados por modificador seguido de núcleo, sendo este geralmente o elemento da direita. Para Jespersen (1924 *apud* MCDONALD, 1995), o núcleo semântico é a palavra mais importante, que é definida ou modificada por outra. Assim, em *apple cake*, o núcleo semântico é *cake*, pois a expressão refere-se a um bolo que é feito de maçã, sendo *apple* o seu modificador. Há também casos em que o composto NN não possui núcleo, como em *poet-painter*, em que o referente é as duas coisas ao mesmo tempo ou em compostos com sentido pejorativo, tal qual *bonehead.* Quando não é possível identificar um núcleo, considera-se a referência; ou seja, se os dois elementos constituem uma unidade, um único referente, tem-se um composto.

Quando um composto é considerado o hipônimo de seu núcleo, tem-se um endocêntrico. Veja-se o composto *desktop computer*. Como é possível dizer que *desktop computer* é um tipo de computador, esse composto pode ser considerado endocêntrico. Em alguns casos de compostos nominais não é possível estabelecer essa relação de classe e subclasse. Se a relação "é um" ou "é um tipo de" não podem ser identificadas, a expressão é exocêntrica, com em *bird brain*. A relação "é um" não se aplica nesse caso, pois não podemos afirmar que um *bird brain* é um tipo de *brain*; se chamamos alguém dessa forma, referimo-nos figurativamente a um tipo de pessoa, cujo tamanho do cérebro está sendo comparado ao de um pássaro.

O conceito de lexicalização está relacionado à transparência semântica dos constituintes de um composto. Para Sandmann (1997), são lexicalizados os compostos que não são transparentes semanticamente, também chamados de opacos. Neste trabalho, adota-se esse conceito de lexicalização, seguindo a mesma linha de alguns trabalhos em PLN, como Copestake (2003), que sugere a construção de uma lista no léxico para armazenar os compostos lexicalizados.

[3] Embora o conceito de composto NN adotado seja o de dois substantivos separados, nas obras de referência, são encontrados exemplos de diversos tipos, palavras escritas juntas ou não. Como as possibilidades de relações semânticas são possíveis tanto com palavras escritas juntas como separadas, apresentaremos exemplos dos dois tipos ao nos referirmos a trabalhos lingüísticos de outros autores.

Para Lieber (2004), os compostos são divididos em dois grandes grupos conforme a classe gramatical do segundo elemento. Os compostos nominais são aqueles cujo N2 não é derivado de verbos, enquanto os compostos sintéticos (ou deverbais) apresentam um N2 deverbal, como em *truck driver*, *load tolerance*, *city employee*, etc. Apesar de o escopo deste trabalho se limitar aos compostos nominais, quando se faz uma busca por compostos formados por dois substantivos em um *corpus* de língua inglesa, depara-se (conforme dados de TEIXEIRA e CHISHMAN, 2008) com expressões em que o segundo elemento é deverbal. Casos como *flood losses*, *foundation investigation*, *horseback gathering* e *drainage improvement*, em que o N2 é deverbal, estão presentes nos dados deste estudo, pois, sendo deverbais ou não, esses itens sofreram uma nominalização e são usados como substantivos no contexto específico de estudo.

1.1. Teoria do Léxico Gerativo

A Teoria do Léxico Gerativo, de Pustejovsky (1995), propõe níveis de representação para as entradas lexicais. Segundo Chishman (2002), essa abordagem vê o léxico como um componente gerativo, em vez de tentar descrevê-lo de forma enumerativa, como ocorre em um método mais tradicional. O léxico, para Pustejovsky (1995), não é considerado um conjunto estático de palavras, já que a língua é usada de forma criativa, gerando novos significados para as palavras de acordo com os seus contextos.

Neste trabalho, interessa-nos, do maquinário da proposta de Pustejovsky (1995), o nível denominado de estrutura *qualia*, que é uma representação dos aspectos essenciais do significado de um nome. A estrutura *qualia* é composta por quatro papéis, que constituem os aspectos do significado de uma palavra, quais sejam: a) papel constitutivo, que expressa a relação entre um objeto e suas partes; b) papel formal, que distingue o objeto dentro de um domínio maior, trazendo os seus atributos físicos; c) papel télico, expressando o propósito ou função do objeto; d) papel agentivo, trazendo os fatores envolvidos no surgimento ou na criação de um objeto.

Decompondo-se as palavras dessa forma, a descrição lógica torna-se mais detalhada, possibilitando a relação entre os itens lexicais de tal forma que possam ser criados novos significados de acordo com as combinações lexicais. Isso se dá basicamente por meio do mecanismo gerativo chamado de co-composição. Para Pustejovsky (1991), o significado de uma sentença é determinado não só a partir aplicação do verbo sobre o argumento, mas também pela aplicação da função do argumento sobre o verbo.

Nos compostos também é possível aplicar esse tipo de interpretação bidirecional. Busa e Johnston (1999) propõem uma interpretação para os compostos nominais nas línguas inglesa e italiana a partir da estrutura *qualia*. Descreve-se aqui como os autores utilizaram cada um dos papéis para representar a semântica dos compostos. O aspecto formal remete à relação “é um”, em que o hiperônimo do composto será o seu núcleo. Vejamos a expressão *bread knife*. Como o hiperônimo de *knife* é *artifact_tool*, esse será o mesmo para *bread knife*.

Por modificação télica, entende-se o propósito de algo. Em *bread knife*, *bread* modifica telicamente *knife*, pois essa expressão pode ser definida como “uma faca usada para cortar pão”.

A modificação agentiva pode ser identificada em *bullet wound*. O elemento *bullet* especifica como o ferimento foi feito, exercendo o papel agentivo, já que o nome relaciona a origem do objeto, como ele surgiu. Um *bullet wound* é um ferimento que surgiu através do ato de disparar uma arma.

A relação “parte de” é expressa no papel constitutivo, em que o modificador especifica uma parte ou subparte do núcleo, como em *glass door*. Uma *porta de vidro* é uma porta feita de

vidro, em que *vidro* expressa o material do qual esse objeto é feito. Para os autores (BUSA; JOHNSTON, 1999), o composto é interpretado como um hipônimo do seu núcleo; assim, pode-se dizer que uma porta de vidro é uma porta. Para interpretarmos o composto *porta de vidro*, basta representar a estrutura *qualia* de porta, preenchendo o papel constitutivo com vidro.

Considerando-se as relações semânticas entre os elementos de um composto NN sugerida por outros autores, é possível criar subdivisões para cada papel da estrutura qualia.

A relação de função é apresentada nos trabalhos de Jespersen (1909), Marchand (1969), Warren (1978), Adams (1973) (*apud* RYDER, 1994) e Copestake (2003). Pode-se dizer que o primeiro substantivo (N1) indica para que serve o segundo (N2). Como exemplo, temos os seguintes compostos: *flagstaff, beehive, keyhole, birdcage, wineglass, cigar-case, fuel oil, gear wheel*. Para Warren (1978 *apud* RYDER, 1994), tanto o tempo como o lugar podem expressar a finalidade de algo. Em *nightdress*, que é traduzido como camisola, pode-se entender a expressão a partir do significado das duas palavras que a constituem: tem-se uma espécie de vestido que é geralmente usado à noite, esta é a sua função. Já *weekend guests* são os convidados que se hospedam um uma determinada residência durante o fim de semana, sendo que *weekend* não é a sua função, apenas a sua localização temporal.

Em Jespersen (1909 *apud* RYDER, 1994), a relação de agentividade é explicada da seguinte maneira, o N1 é um instrumento ou uma ferramenta que dá origem ao N2, como em *gunshot* e *sabre-cut*. Um dos predicados de Levi (1978 *apud* RYDER, 1994) também expressa que um dos substantivos deu origem ao outro, o predicado CAUSE, em que o N1 pode ser o sujeito (*drug death, birth pains*) ou o objeto direto (*tear gas, disease germ*). O N1 também pode ser a origem do N2, conforme outro predicado de Levi, FROM. Este predicado indica uma relação de direção, de onde algo veio (*olive oil, candlelight, battle fatigue*). O último desdobramento do papel agentivo refere-se ao N1 como fonte de energia do N2, também relacionado com um dos predicados de Levi, USE de Levi, em que o N1 é o objeto direto (*voice vote, steam iron*). Desta forma, o papel agentivo pode ser dividido em quatro subgrupos: um instrumento que dá origem ao N2, um substantivo causa o outro, o N1 é a origem de N2, N1 é a fonte de energia de N2.

A relação de parte e todo está presente nos cinco trabalhos estudados por Ryder (1994), sendo que duas direções são possíveis: N1 é uma parte de N2, como em *stone fruit*, e N2 é uma parte de N1, como em *broomstick*. Jespersen (1909), Marchand (1969) e Adams (1973) (*apud* RYDER, 1994) também criaram mais uma categoria, material ou "N2 é feito de N1", em que também se poderia dizer que o material está de alguma forma contido em um dos elementos. No entanto, há uma diferença entre a relação de parte e todo e a relação de material: em *feather-bed*, as penas fazem parte da cama; se tirarmos as penas, ainda temos uma cama; mas em *gold ring*, o ouro é o material do qual o anel todo é feito; sem o ouro não temos um anel. Quanto à relação de parte e todo, Warren inclui uma subcategoria em que a parte constitui uma característica abstrata, como em *room temperature*. Ryder (1994) sugere a relação de recipiente e conteúdo, como em *wine glass*, em que o copo é o recipiente para o vinho. Desta forma, podemos desmembrar o papel constitutivo em quatro: todo e parte constituinte, todo e característica abstrata, material, recipiente e conteúdo.

Outros tipos de compostos que, caso fossem analisados através da teoria da estrutura *Qualia*, seriam incluídos no papel constitutivo, são os que se referem a local e tempo. Em *garden-party*, o N1 indica onde o N2 ocorre. Para Adams (1973 *apud* RYDER, 1994), o local pode ser tanto o N1, como em *field mouse* e *pocket handkerchief*, como o N2, como em *biscuit factory* e *law court*. Já Warren (1978 *apud* RYDER, 1994) inclui tanto as relações de tempo como as de espaço na categoria localização. Nos compostos *eveningsong* e *nightclub*, o N1 indica o tempo. Conforme já mencionado anteriormente, Warren (1978) (*apud* RYDER, 1994) inclui tanto o tempo como o local em um conceito mais geral, a localização. Propomos,

neste trabalho, que mesmo tempo e lugar aparecendo em outras relações como parte ou função, o seu uso como localização precisa de alguma forma ser diferenciado e a sua principal característica é ocupar o N1. Da mesma forma, o material que é indicado no N1 pode ser considerado parte do todo, mas, como geralmente o objeto é completamente feito de um determinado material e não há como separar a parte desse todo, há uma relação diferente de alguma forma.

Outra relação que também é incluída no papel constitutivo, segundo Warren (1978 *apud* RYDER, 1994), é a relação entre posse e possuído. No entanto, a relação é diferente. Um *family car* é um carro possuído pela família, mas soa estranho dizer que o carro faz parte da família. Desta forma, acredita-se que a relação de posse também deve ser tratada separadamente.

Os compostos que expressam o papel formal são os compostos endocêntricos, em que o composto como um todo é um tipo, uma subclasse do seu núcleo, como em *tax law*, que é um tipo de lei ou em *desktop computer*, que é um tipo de computador.

1.2. Gramática Cognitiva e a Semântica de *Frames*

Nesta seção reúnem-se duas abordagens que têm em comum o compromisso com os princípios da Linguística Cognitiva. Ryder (1994) parte dos princípios da gramática cognitiva (Langacker, 1987) para tratar da semântica dos compostos NN, enquanto Fillmore (2006), ainda que seu foco não seja necessariamente a semântica interna dos compostos, propõe a teoria dos *frames* semânticos, abordagem adotada para a criação da base de dados lexical, *FrameNet*[4] (BAKER *et al*, 1998).

Ryder (1994) acredita que explicar os compostos por meio de regras e incluir as exceções em um léxico não é a melhor forma de análise. O que ela defende é a existência de padrões com diferentes graus de produtividade. Para expressar as relações semânticas possíveis em compostos NN, Ryder (1994) utiliza *templates*, que são estruturas abstratas do conhecimento que resumem o que se sabe sobre a variedade de casos e representa as relações entre as variáveis. Por exemplo, o *template* de *jogo* evoca os elementos que fazem parte deste evento, como local, jogadores, tempo, etc. Esses *templates* também podem ser considerados *frames*. Segundo Fillmore (2006), um *frame* é um sistema de conceitos relacionados de tal modo que, para entendê-lo, é necessário entender a estrutura toda na qual ele se encaixa.

Os *templates* linguísticos formados a partir de grupos de compostos que possuem um elemento em comum, ou seja, que constituem famílias de compostos, servem de bases de analogia. A palavra que se repete em várias construções compostas é chamada de nódulo ou *core word* por Ryder (1994) e pode ocupar tanto o lugar do núcleo quanto o do modificador. Exemplos: *sea lion, seaman, sea cow, seaweed* ou *boathouse, warehouse, tree house, firehouse*.

Ryder (1994) analisou 1.600 compostos extraídos dos livros *American Heritage* e *American Heritage Word Frequency Book* (1971) e identificou cinco *templates* linguísticos frequentes, em que um dos elementos refere-se a: localização, recipiente, ser humano, parte do corpo e animal. Os *templates* linguísticos, também chamados de padrões pela autora, apenas diferem pelo fato de que alguns são mais frequentes do que outros.

O padrão mais frequente encontrado nos dados de Ryder (1994) é: Localização Y + X = X localizado em Y[5]. Exemplos: *camp stool, altarpiece, contrywoman.*O reverso desse padrão

[4] Disponível em: <http://framenet.icsi.berkeley.edu/>

[5] Quando refere-se aos *templates*, Ryder (1994) utiliza Y para referir-se ao primeiro substantivo do composto NN e X, para o segundo. O sinal de igual relaciona o padrão com uma paráfrase equivalente.

também é possível: Y + Localização X = X no qual Y é tipicamente encontrado. Exemplos: *apple orchard, cranberry bog, hen house.*

Nos dados de Ryder (1994), o segundo padrão comum é: Y + Recipiente X = X que tipicamente contém Y. Exemplos: *suitcase, teapot, ice bag.* O padrão recipiente apresenta as mesmas características de localização, porém ele é móvel, ou seja, é geralmente menor do que localização e na maioria das vezes é artificial. O reverso desse padrão – Y Recipiente + X – não apresenta uma interpretação homogênea. Isso ocorre porque os recipientes são artificiais e foram criados com um determinado propósito, podendo conter diversos itens. Entretanto informar que algo pode ser contido em um recipiente não é uma informação relevante. Assim, duas interpretações são possíveis: "X habitualmente contido em Y" (*bag lunch, box wine*) e "X semelhante a Y em formato" (*boxcar, box stall, box office, box bed*).

Substantivos que se referem a seres humanos também apresentam *templates* altamente frequentes: animal doméstico + humano = um humano que cria, cuida ou treina animal doméstico. Exemplos: *horseman, cattleman, poultryman*; veículo/maquinaria + humano = um humano que opera veículo/maquinaria. Exemplos: *boatman, cabman, trainman*; instrumento/ferramenta + humano = um humano que trabalha usando instrumento/ferramenta. Exemplos: *ploughboy, brakeman, cameraman*; arma + humano = um humano que usa a arma, geralmente como um assassino ou caçador profissional. Exemplos: *gunman, spearman, bowman*.

Há um *template* mais geral que normalmente se sobrepõe aos outros ou os substitui: produto + humano = um humano que faz, vende, entrega, transporta ou processa um produto como uma profissão. Exemplos: *mailman, milkmaid*. Por produto, compreende-se "qualquer coisa cuja produção ou desenvolvimento é influenciado por pessoas" (RYDER, 1994, p. 101). Dessa forma, como uma mesma pessoa pode executar mais de uma atividade em relação a um mesmo item, como produzir, vender ou processar, esses *templates* se relacionam.

Os dois últimos padrões são menos frequentes, mas apresentam certa regularidade. Um deles se refere à parte do corpo: parte do corpo + roupa/jóias = roupa/jóias vestidas na parte do corpo. Exemplos: *headband, necktie, earmuffs;* parte do corpo + roupa = roupa que se estende até a parte do corpo. Exemplos*: waistcoat, ankle socks, knee socks*; parte do corpo + algo que não é roupa = algo que não é roupa é operado por/usado na parte do corpo. Exemplos: *foot pedal, foot brake, handcart.*

O quinto padrão é com os compostos formados por algum animal: animal + animal carnívoro = animal carnívoro que come/caça animal. Exemplos: *bee fly, bee moth, bee louse.* Ryder (1994) identificou também outro *template,* que relaciona uma planta com o que ela produz: produto + planta = planta que produz o produto. Exemplos: *apple tree, cranberry bush, tomato plant*.

Ao unir dois substantivos formando um composto, é possível estabelecer também a relação que Ryder (1994) chama de equivalência, em que duas coisas são comparadas de alguma forma. Assim decidiu-se desmembrar essa relação de Ryder da seguinte forma: a comparação propriamente dita, em que os dois elementos do composto apresentam alguma característica em comum, tal como em *bell-flower* e *goldfish*; os copulativos, em que um dos substantivos é hipônimo do outro, como em *bossman*, pois trata-se de um homem que é o chefe. Dentro da classe de homens, há subgrupos em que alguns são chefes, outros são empregados; os aditivos, como em *secretary-treasurer e poet-painter*. Nestes casos, o referente constitui os dois elementos ao mesmo tempo em um mesmo nível.

Tendo em vista todas as relações apresentadas até aqui, a estrutura *Qualia* de Pustejovsky e os *templates* de *Ryder*, seguidos de seus desmembramentos e adaptações para este trabalho, outro recurso se mostrou útil para a análise da semântica dos compostos, o *FrameNet*. Ao se fazer uma busca pela palavra *field* na base de dados do *FrameNet*, chega-se ao *frame Locale_by_Event* (localização por evento), que é descrito como "um local definido em termos

de um evento que ocorreu ou ocorrerá lá". Entre os seus elementos *frames*, tem-se: evento definidor, local, partes constituintes, descritor, nome e localização relativa. A partir deste tipo de busca, é possível saber quais os seus argumentos, ou seja, quais as características dos itens lexicais que podem se combinar com essa palavra formando um composto.
Um estudo mais recente na área que também analisa os compostos NN sob uma perspectiva cognitivista é o de Baroni, Guevara e Pirrelli (2007), cujo objetivo é oferecer uma tipologia que expresse as possíveis relações entre o modificador e o núcleo de um composto NN. Assim como em Ryder (1994), que organiza as relações presentes nos compostos em *templates*, há uma proposta de interpretação a partir de representações conceituais denominadas cenários, em que padrões de combinações são identificados. A principal diferença entre o estudo de Baroni, Guevara e Pirrelli (2007) e nosso estudo é que eles sugerem uma classificação mais rasa para os compostos, dividindo-os em: primários, copulativos, atributivos, exocêntricos e sintéticos. Já em nosso estudo, há a proposição de uma tipologia mais específica em que se busca detalhar os significados evocados por determinados *templates* ou esquemas.

2. Compostos NN: seleção de *corpus* e extração de dados

Nesta seção, os procedimentos metodológicos que foram seguidos para a realização da análise proposta no presente estudo são apresentados. Como nem sempre é possível encontrar um *corpus* disponível e apropriado para os objetivos de pesquisa, optou-se pela compilação[6] de um *corpus*.

2.1. Construção do *corpus*

Como o objetivo é compreender as relações semânticas entre os elementos de compostos NN do inglês, optou-se pelas edições da revista *National Geographic*. Um fator importante que contribuiu para esta escolha foi o fato de que a disponibilização do material é simples, pois os textos já estão em formato eletrônico, na página on-line da revista. Além disso, segundo Biber (1993, p. 233), os artigos de revistas "incluem uma grande variedade de propósitos e mostram vastas diferenças linguísticas entre os textos do registro".
O *corpus* compilado é formado por 10 edições da revista *National Geographic*, publicadas entre agosto de 2007 e maio de 2008[7]. No *corpus*, com 208.201 *tokens*, 24.327 *types* e 11.052 frases, chegou-se a quase 4.700 ocorrências de possíveis compostos NN.
Coletado o *corpus*, passou-se para a tarefa de itemização, já que esse formato é pré-requisito para o etiquetador morfossintático. O itemizador que formata o texto em uma palavra por linha utiliza a arquitetura Java J2SE. A principal vantagem dessa arquitetura é o fato de ser multiplataforma, ou seja, independente de sistema operacional, podendo funcionar em *linux*, *windows*, entre outros.
Para separar cada palavra, o programa identifica os espaços em branco e os substitui por um símbolo de nova linha. Dessa forma, os sinais de pontuação não são separados das palavras. Expressões compostas separadas por hífen e siglas também são mantidas na mesma linha. A seguir, será descrito o processo de como os compostos NN foram selecionados para a fase da análise semântica.

[6] Compilar, para a área de Linguística de *Corpus*, refere-se à tarefa de reunir textos para a confecção de um *corpus*. Berber Sardinha (2002) refere-se à compilação como a criação de *corpus*.
[7] *National Geographic Magazine*, disponível em <http://ngm.nationalgeographic.com/>.

VEREDAS ON-LINE – LINGUÍSTICA DE CORPUS E COMPUTACIONAL – 2/2009, P. 84-99 – PPG LINGUÍSTICA/UFJF – JUIZ DE FOR A – ISSN 1982-2243

2.2. Seleção dos dados para o estudo

Como este estudo trata dos compostos nominais formados por dois substantivos na língua inglesa, era preciso extrair do *corpus* uma sequência de dois substantivos sem que houvesse outro substantivo antes ou depois. Também se tornou necessário obter uma lista de todas as expressões seguidas pela quantidade de vezes em que elas ocorrem no *corpus*. Para o levantamento desses dados, é necessário que o *corpus* esteja anotado morfologicamente, pois só assim é possível fazer uma busca por expressões formadas por substantivos. Como o *corpus* deste estudo não estava etiquetado, ele precisou passar por esse processamento, pois só assim outro programa poderia identificar a informação necessária.

Optou-se pelo etiquetador *TreeTagger* (SANTORINI, 1990) para a língua inglesa, por ser uma ferramenta gratuita e com bons resultados, com uma média de 96% de precisão. Com as anotações morfológicas, o *corpus* estava pronto para a extração das sequências de dois substantivos. Como não foi encontrado nenhum extrator apropriado e que fosse gratuito, foi necessário criar essa ferramenta, que tem como base as etiquetas do *TreeTagger* e utiliza a mesma arquitetura do itemizador Java J2SE[8].

O *TreeTagger*[9] é um etiquetador de *part-of-speech (POS)*, ou seja, é um sistema que faz automaticamente o reconhecimento das categorias morfossintáticas. Ele foi desenvolvido na Universidade de Stuttgart, na Alemanha, e é utilizado em mais de 10 idiomas diferentes, dentre eles o inglês, o francês, o alemão e o italiano.

Foram obtidos bons resultados com o etiquetador, ainda que algumas palavras não tenham sido classificadas corretamente, como verbos e adjetivos que foram etiquetados como substantivos. Algumas palavras apareceram equivocadamente anotadas como substantivos, mas isso não dificultou a análise, pois quando se examinou cada composto, o seu contexto de uso também foi considerado.

Conforme mencionado anteriormente, após o *corpus* estar etiquetado, era preciso extrair as sequências formadas por dois substantivos com o objetivo de se chegar aos compostos nominais. Para esse fim, foi desenvolvido um extrator com a mesma arquitetura utilizada pelo itemizador Java J2SE.

Esse extrator busca pelas sequências de dois substantivos a partir das etiquetas do *TreeTagger*. Assim, ele busca por: NN NN, NN NNS, NNS NN e NNS NNS. Durante o desenvolvimento do extrator, houve também o cuidado de que não ocorresse um substantivo antes ou depois dessa sequência, pois o foco deste trabalho são apenas os compostos formados por dois substantivos. Assim, quando três substantivos consecutivos ocorrem, o programa verifica isso e descarta. Como saída, o programa oferece uma lista com possíveis compostos nominais e o seu número de ocorrências no *corpus*, ou seja, a sua frequência.

Cada resultado do extrator precisou ser conferido, pois não havia garantias de que as palavras listadas pelo programa fossem realmente compostos nominais, já que erros de etiquetação ou de formatação poderiam ter ocorrido.

A partir dos resultados do extrator de sequências NN, chegou-se a 4.693 candidatos a compostos, sendo que, desse total, 690 ocorreram mais de uma vez no *corpus*. A grande maioria dos possíveis compostos ocorreu apenas uma vez no *corpus* todo, fenômeno identificado com sendo um *hapax legomenon*. Isso já era previsto, pois há pouca probabilidade de a mesma combinação de duas palavras específicas ocorrer repetidamente.

---

[8] Para a confecção do itemizador e do extrator de sequências de expressões formadas por dois substantivos, foi fundamental a colaboração do aluno Lucas Lermen, bolsista de Apoio Técnico do projeto FrameCorp (coordenado por Rove Luiza de Oliveira Chishman) em 2008.

[9] Disponível em: <http://www.ims.uni-stuttgart.de/projekte/corplex/TreeTagger/>.

Tendo-se como resultado uma grande quantidade de dados, resolveu-se selecionar algumas sequências. Como critério, foram adotadas as *core words*, ou seja, aquelas palavras que aparecem como elementos de mais de um composto. Um exemplo de *core word* no presente *corpus* é *day*, pois foram encontrados compostos em que essa palavra é utilizada tanto como núcleo quanto como modificador em uma construção composta, tais como *day care*, *day jobs*, *day laborer*, *day pack*, *election day*, *feast days*, *field day* e *harvest days*. Esse é o mesmo critério utilizado por Ryder (1994) no seu estudo. Partindo-se de *core words*, é mais fácil chegar a possíveis *templates* que servem como um indicativo para as relações entre os elementos de um composto no presente *corpus*. Dessa forma, também os casos de *hapax legomena* não representam um problema para a análise, já que um mesmo substantivo pode fazer parte de diversos compostos NN.

Foram selecionadas as *core words* que ocorrem 10 ou mais vezes, da lista de 4.693 candidatos a compostos, com o objetivo de se poder caracterizar os *templates* mais frequentes no *corpus*. Não se levou em consideração, nesse primeiro momento, se a palavra ocorre no primeiro ou segundo substantivo. A partir dos substantivos, ou palavras que receberam a etiqueta de substantivo pelo etiquetador, foi realizada a conferência manual de cada candidato a composto. Para saber se cada sequência de dois substantivos constitui de fato um composto NN, foram analisadas as 1.641 ocorrências por meio do concordanceador do *WordSmith Tools* (SCOTT, 2008).

Através da conferência manual de cada composto, o número de compostos diminuiu bastante, chegando a 842 compostos NN. Para a realização dessa tarefa de conferência, as concordâncias fornecidas pelo *WordSmith Tools* (SCOTT, 2008) se mostraram extremamente importantes.

Durante a aplicação das etapas da metodologia, percebeu-se que as ferramentas apresentaram alguns problemas, como erro na etiquetação morfológica do *corpus* e erro de configuração. Mesmo com parte da análise tendo de ser feita manualmente, as ferramentas facilitaram o trabalho. Sem o auxílio dessas ferramentas não haveria como se realizar a busca por *core words* e se obter a sua frequência de forma precisa e rápida. O concordanceador do *WordSmith* (SCOTT, 2008) também foi extremamente útil para a conferência dos candidatos a compostos.

3. Análise semântica

Partindo-se dos resultados de extração, foram selecionados 200 compostos aleatoriamente. Esta análise tem como objetivo principal compreender as relações semânticas entre os elementos de compostos NN em língua inglesa. Um dos propósitos é reconhecer quais são as propriedades semânticas presentes nos compostos do *corpus* de estudo. Para tanto, foram considerados os seguintes conceitos: produtividade, transparência semântica, nuclearidade, composição exocêntrica e endocêntrica, lexicalização e nominalização. Os dados numéricos desses resultados gerais são apresentados na tabela 1, a seguir:

200 compostos			
Nuclearidade	*Composição exocêntrica e endocêntrica*	*Transparência semântica e lexicalização*	*Nominalização*
195 compostos permitem a identificação do núcleo semântico	199 compostos endocêntricos	198 compostos com transparência total	44 compostos constituídos por algum tipo de nominalizaçao

5 compostos apresentam dificuldades quanto à identificação do núcleo semântico	1 composto exocêntrico	1 composto com transparência parcial	156 compostos totalmente nominais
		1 composto lexicalizado	
Total: 200	Total: 200	Total: 200	Total: 200

TABELA 1: Resultados gerais da análise inicial dos 200 compostos selecionados

Conforme os dados da tabela 1, entre os 200 compostos analisados, em 195 foi possível identificar o núcleo semântico do composto, sendo que ele é representado no N2. Em *banana trees*, por exemplo, o seu núcleo é *trees*, pois se refere a uma árvore; o núcleo de *car accidents* é também o seu N2, já que o seu referente é *accidents*. Entre os 200 compostos analisados, apenas cinco foram problemáticos quanto à identificação do seu núcleo. O primeiro composto que causa dúvida é *car bombs*, em que o referente é as duas coisas ao mesmo tempo, trata-se de um carro e de uma bomba. Sendo assim, tem-se um composto aditivo e não há núcleo. O mesmo pode-se dizer dos compostos *island home* e *home village*. Já em *water ice*, o núcleo é o N1, pois se tem como referente a *água* e o modificador (*ice*) informa o seu estado. Outro composto em que não foi possível identificar um núcleo, principalmente porque não se trata de um composto composicional, é *water hole*. Na oração *Fongoli chimps have been exhibiting some other novel behaviors: soaking in a water hole, passing the afternoon in caves*, o composto poderia ser traduzido como *nascentes d'água*. Trata-se de uma expressão lexicalizada, aspecto que será retomado a seguir.

Vale retomar que os compostos endocêntricos são os que constituem um hipônimo do seu núcleo. Como foi possível identificar um núcleo na maioria dos 200 compostos analisados, também é possível afirmar que eles são endocêntricos. Um *fruit bat* é um tipo de morcego, assim como uma *rain forest* é um tipo de floresta. Parece relevante informar que o composto é endocêntrico quando este é um exemplar de uma categoria maior, como em *fruit bat* e *rain forest*. No entanto, também se pode dizer que uma *gorilla family* é um tipo de família e que uma *business licence* é um tipo de licença. Mas é possível trazer mais detalhes sobre a relação entre *gorilla* e *family* que a relação de hiponímia não abrange. O gorila é um membro da família, faz parte deste grupo, enquanto que uma licença para abrir negócios tem um propósito específico, uma função, serve para abrir negócios. Essa questão será retomada na próxima seção. Apenas um composto não foi considerado endocêntrico. O composto *water hole* é denominado exocêntrico, pois é lexicalizado.

Dos 200 compostos analisados, 199 são transparentes. Eles também são produtivos, pois são expressões livres, já que há outras construções formadas a partir do núcleo. Como exemplo, consideremos o composto *baseball field*: além de *baseball field*, é possível encontrar na mídia *hockey field*, *soccer field*, entre outros. Vê-se que há uma relação entre a transparência semântica e a produtividade. Os compostos lexicalizados, como *water hole*, apresentam um grau menor de produtividade, já que não há outros compostos do mesmo tipo.

O composto *rock salt* é parcialmente composicional, pois se trata de um sal, mas não há pedras no sal. O modificador *rock* foi utilizado para informar que o sal encontra-se em forma de pedra. Em língua portuguesa não se utiliza esse mesmo modificador, pois, ao nos referirmos a esse tipo de sal, utilizamos a expressão *sal grosso*.

Os compostos parcialmente transparentes podem ser interpretados de alguma forma a partir do ponto de vista escolhido pelo analisador, pois há alguma relação entre os dois substantivos. Já nos compostos lexicalizados, como *water hole* não há uma transparência evidente.

Compreende-se por compostos nominalizados aqueles em que um dos substantivos é formado por um verbo e um sufixo nominal. Entre os 200 compostos escolhidos para esta análise, 44 apresentam um elemento nominalizado, tais como *cane cutter, killing field, research coordinator, storage facility, space exploration, city government* e *construction efforts*. O verbo de origem do elemento nominalizado do composto geralmente traz alguma informação referente ao significado do composto. Neste estudo, uma nominalização como *cooking oil* é interpretada pela relação *serve para*, em que o N1 é a função e o N2 é o produto. Se a interpretação fosse mais específica, sugerindo uma paráfrase diferente para cada composto, teríamos algo como "óleo que serve para cozinhar". O verbo *cook* (cozinhar) é a função do núcleo do composto.

Além das conclusões gerais apresentadas até aqui, foi feita uma análise da semântica dos 200 compostos nominais do *corpus National Geographic* de forma que se pudesse chegar a um grupo de padrões semânticos, ou seja, relações semânticas recorrentes entre os elementos dos compostos NN em língua inglesa.

Partindo das conclusões referentes ao estudo teórico e dos *templates* sugeridos por Ryder (1994), chegou-se a um conjunto de 26 relações, como as de função, instrumento, material, local, tempo, animal, humano, substância, artefato, produto, planta, parte do corpo, arma, etc. Iniciou-se com uma classificação para cada substantivo a partir dos dados encontrados nos *templates* de Ryder (1994) e em outros estudos, como Pustejovsky (1995), Warren (1978 *apud* RYDER, 1994), Marchand (1969 *apud* RYDER, 1994) e Levi (1978 *apud* DOWNING, 1977).

Além dessas relações, considerou-se importante incluir informações mais específicas, como doença, vírus, fonte de energia, possuidor, possuído e profissão. A partir do predicado CAUSE de Levi (1978 *apud* DOWNING, 1977), percebeu-se que os exemplos sempre traziam algo de negativo, como uma doença. Encontramos compostos em que a relação semântica é de causa, como em *skin cancer*. Ao realizar-se uma busca no *FrameNet* pelo item lexical *cancer*, chegou-se ao *frame Medical_condictions* (condições médicas), que possui os seguintes elementos *frame*: doença, paciente, parte do corpo, causa e grau. Os dados do *FrameNet* contribuíram para uma maior compreensão sobre a relação semântica dos compostos. Assim, utilizando-se a etiqueta CAUSE, pode-se parafrasear o composto *skin cancer* como *um câncer causado na pele*.

Os dados sobre cada substantivo encontrado no *FrameNet* também trazem informações desse tipo. Além disso, cada palavra, ou elemento lexical, faz parte de um *frame* que está relacionado com alguns verbos. É o verbo que vai instanciar o *frame*, relacionando os substantivos. Ryder (1994) também utiliza verbos para expressar a relação entre os substantivos de um composto. Ao optar por um verbo, chegou-se a uma única palavra que expressa a semântica do composto e que pode servir como uma etiqueta semântica a ser utilizada em tarefas de processamento da língua.

Os padrões semânticos recorrentes nos compostos do *corpus* de estudo referem-se às seguintes relações: telicidade, agentividade, meronímia, localização, posse e hiponímia. Na tabela 2, é apresentada a frequência de cada relação encontrada no *corpus*. O número de ocorrências refere-se aos compostos diferentes, sem contar os casos de repetições. Por exemplo, mesmo que *memory drugs* tenha ocorrido três vezes no *corpus*, para esta tabela, ele foi contabilizado uma única vez.

Relação / Verbo	**Types**	**Exemplos**
hiponímia / o composto é um tipo de núcleo	49	*palm trees*
localização / é localizado em (local)	40	*school play*
meronímia / possui (parte integrante ou característica abstrata)	26	*church floor, island culture*

VEREDAS ON-LINE – LINGUÍSTICA DE CORPUS E COMPUTACIONAL – 2/2009, P. 84-99 – PPG LINGUÍSTICA/UFJF – JUIZ DE FOR A – ISSN 1982-2243

telicidade / serve para	25	*memory drugs*
agentividade / dá origem a, é causada em, vem de, é feito a partir de, funciona a partir de	14	*car accidents, brain infection, cane juice, life force*
posse / tem	11	*family mosque*
localização / ocorre em (tempo)	9	*night school*
meronímia / é feito de (material)	7	*metal armor*
meronímia / contém (recipiente e conteúdo)	7	*rice bag*

TABELA 2: Relações encontradas no *corpus*

Quando não foi possível estabelecer uma relação específica entre os elementos dos compostos, estes foram considerados endocêntricos, e incluídos na categoria de hipônimos. Os compostos que constituem uma instância do seu núcleo trazem no modificador características bem específicas, tais como: o assunto, sobre o que se trata determinado evento (*biomimetics research, carbon emissions)*; uma subclasse ou exemplar que faz parte da classe (*snack food, palm trees, law school, fire brigade)*; órgão ou grupo social responsável por uma determinada instituição (*state banks*, *city government).*

Também foi possível identificar tipos de pessoas, animais, plantas, substâncias, artefatos ou eventos. O composto pode expressar a profissão da pessoa (*construction workers*), a sua área de atuação (*rocket scientist*) ou algum vício que possui (*drug users*). Em relação a tipos de animais, o modificador traz alguma característica específica do bicho, como *fruit bat*. Por tipos de plantas, deparou-se basicamente com tipos de árvores no *corpus*. Entre os tipos de substâncias, líquidos ou produtos, encontraram-se no *corpus* os seguintes: *water ice* e *warming seas*. Como exemplo de tipo de artefato, temos o *cell phone*, que se diferencia dos outros por apresentar a característica de ser móvel. *Opium trade, ethanol production* e *computer classes* são tipos de eventos, pois os substantivos *trade*, *production* e *classes* denotam uma série de ações específicas em contextos variados.

Entre os 200 compostos, 46 foram classificados apenas como endocêntricos, fato que indica que a maioria dos compostos analisados, mais de 75%, apresenta alguma relação semântica mais específica entre os seus elementos, tais como função, localização e parte/todo.

Os compostos relacionados ao local trazem informações quanto à origem de alguém (*country boy)* ou onde trabalha *(lab people).* Fica mais evidente que a palavra *country* refere-se à origem quando analisamos a ocorrência: *This son of Hazarajat is the proverbial country boy who came to the big city and made good.* A palavra *country* está em oposição a *big city.* Quanto a animais, plantas, artefatos e eventos, foi possível identificar os seguintes *templates*: local + animal: *desert beetle* = besouro que vive no deserto; local + planta: *water plants* = plantas que nascem na água; planta + local de cultivo: *orchid farm* = fazenda onde orquídeas são cultivadas; artefato + localização: *electronics shops* = lojas onde aparelhos eletrônicos são vendidos; evento + local: *baseball field* = um jogo que ocorre em um lugar específico, o campo.

As buscas realizadas no *FrameNet* também contribuíram para a classificação dos compostos cuja relação é de local. Ao realizar-se a consulta pelo item lexical *farm*, obteve-se a informação de que ele faz parte do *frame Locale_by_use* (localização pelo uso). Esse *frame* possui como elementos principais um local e o seu uso, ou seja, o local é descrito a partir da forma como ele é usado. Voltando ao exemplo *orchid farm*, não se pode dizer que a fazenda é a localização natural das orquídeas, mas que a fazenda é um local utilizado para o cultivo, a plantação de orquídeas. Com o auxílio do *FrameNet*, foi possível identificar tipos diferentes de local, o local destinado a um uso específico e o local como localização, que informa onde algo se encontra.

Entre os *templates* relacionados à relação de meronímia, como parte e todo, foram encontrados os seguintes: grupo + pessoa: *family member* = pessoa que faz parte deste grupo; animal + parte do corpo: *leopard skin* = pele de leopardo; animal + grupo: *gorilla families* = família da qual o gorila faz parte; planta + grupo de plantas: *eucalyptus forest* = floresta constituída por eucaliptos; planta + parte: *tree branches* = galhos que fazem parte da árvore; artefato + parte: *computer keyboard* = o teclado faz parte do computador;
Além disso, foram encontrados os seguintes casos que expressam a relação de função: função + pessoas: *monitoring groups* = um grupo de pessoas cuja função é monitorar algo; função + planta: *feed corn* = milho destinado à alimentação de animais; função + artefato: *car keys* = chaves usadas para abrir o carro; função + substância: *cooking oil* = óleo destinado o uso culinário;
Os compostos que expressam a agentividade geralmente estão relacionados com a função de uma substância, como em *corn ethanol*. Uma ocorrência no *corpus* chamou a atenção por não estar dentro de nenhum padrão recorrente: *chimp feces*. Não se pode dizer que o chimpanzé é o dono ou possuidor das fezes que produziu; entende-se que o animal é a fonte, a origem das fezes. Só se encontrou um caso desse tipo.
A relação de posse foi identificada em pessoas (*family mosque)* ou em animais *(gorilla health)*, pois temos como referente em cada caso algo que pertence a pessoas (*mesquita da família*) ou a animais (*saúde dos gorilas*).
Enquanto que a localização de tempo refere-se a eventos (*weekend class),* o material é relacionado com algum artefato *(stone axes).*
Entre os casos de conteúdo e recipiente, foram encontrados os seguintes casos: *rice bag, cattle trailers, ethanol tanks, gas tanker, water bottles, water bowl* e *water tanks*.

Conclusão

O estudo da semântica dos compostos nominais é tema interdisciplinar, pois é recorrente em trabalhos da área da linguística e da computação. A grande diferença está na forma como cada área aborda este fenômeno linguístico. Enquanto os estudos semânticos sugerem que os tipos de relação entre os elementos de um composto NN são infinitos e não buscam esgotar as possibilidades de interpretação, os trabalhos na área de PLN procuram identificar um grupo limitado de relações semânticas procurando dar conta senão de todos, da maioria dos compostos NN. A partir do aprofundamento de diferentes perspectivas teóricas, chegou-se a uma proposta de análise dos compostos nominais. Esta proposta partiu das diferentes relações semânticas responsáveis pela combinação dos elementos, mas sem ser estanque, pois parte do princípio de que os dois substantivos exercem um papel específico a partir do seu uso.
Uma das grandes inovações deste trabalho foi relacionar uma teoria baseada em *templates*, como a de Ryder (1994), que propõe uma interpretação semântica para os compostos nominais, com a teoria de *frames* semânticos de Fillmore (2006). A consulta à base de dados *FrameNet* trouxe informações baseadas no uso para confirmar a tipologia proposta neste trabalho. Mesmo que os *frames* não considerem as relações entre os elementos de expressões compostas, foi possível utilizar os conceitos semânticos que compõem os frames para confirmar as relações semânticas dos compostos deste estudo.
Este trabalho trouxe sugestões de análise semântica para os compostos nominais em inglês. No entanto, este estudo ainda pode ser ampliado no que diz respeito ao potencial dos *frames* para expressar as relações semânticas dos compostos. Espera-se que este artigo sirva de

inspiração para outros estudos que contribuam para a solução de problemas computacionais relativos às expressões como os compostos.

ABSTRACT. The purpose of this paper is to study the semantics of noun-noun compounds in English. A specific objective is to identify the most frequent relations in a corpus formed by ten editions of the National Geographic Magazine. It is also proposed a typology which expresses the semantic compositionality of these constructions. Through the Corpus Linguistics resources, it was possible to achieve a list of 4.693 compounds candidates. The most recurrent semantic relations in the corpus are: telicity, agentivity, meronimy, localization, possession and hyponymy.

Keywords: Corpus Linguistics; Lexical semantics; Nominal compounds; Semantic Frames.

Referências

BAKER, Collin F.; FILLMORE, Charles J.; LOWE, John B. The Berkeley FrameNet project. In.: *Proceedings of the COLING-ACL*. Montreal, Canada, 1998.

BARONI, M.; GUEVARA, E.; PIRRELLI, V. NN compounds in Italian: Modelling category induction and analogical extension. In: PIRRELLI, Vito (ed.). *Psycho-Computational Issues in Morphology Learning and Processing* (Special issue of Lingue e Linguaggio, 6.2). Bologna: il Mulino, 2007. p. 263-290.

BERBER SARDINHA, Tony. Tamanho de Corpus. *The ESPecialist*, São Paulo, v. 23, n. 2, p. 103-122, jul./dez. 2002.

BIBER, Douglas. Using Register-Diversified Corpora for General Language Studies. *Computational Linguistics*, Cambridge, v. 19, n. 2, p. 219-241, jun. 1993.

BUSA, Federica; JOHNSTON, Michael. Qualia Structure and the Compositional Interpretation of Compounds. In.: VIEGAS, Evelyne (org.). *Breath and Depth of Semantic Lexicons.* London: Kluwer, 1999. p. 167-187.

CHISHMAN, Rove Luiza de Oliveira. A teoria do léxico gerativo: uma abordagem crítica. In.: IBAÑOS, Ana Maria T.; SILVEIRA, Jane Rita Caetano (org.). *Na Interface Semântica/Pragmática.* Porto Alegre: EDIPUCRS, 2002. p. 51-82.

COPESTAKE, Ann. Compounds revisited. In.: *2nd International Workshop on Generative Approaches to the Lexicon, GL'2003.* Geneva, maio, 2003. CD-ROM.

DOWNING, Pamela. On the creation and use of English compound nouns. *Language, Journal of the Linguistic Society of America*, Baltimore, v. 53, n. 4, p. 810-842, dez. 1977.

FILLMORE, Charles J. Frame Semantics. In.: GEERAERTS, Dirk. *Cognitive Linguistics:* Basic Readings. Berlim, Nova Iorque: Mouton de Gruyter, 2006. p. 373-400.

GRANGER, Sylviane (org.). *Learner English on computer.* New York: Longman, 1998. 228p.

LANGACKER, Ronald W. *Foundations of cognitive grammar.* Volume I: Theoretical prerequisites. Standford: Standford University, 1987. 540p.

LIEBER, Rochelle. *Morphology and Lexical Semantics.* Cambridge: Cambridge University, 2004. 196p.

MCDONALD, Scott. *Learning Compound Order:* Towards a Functional Explanation. 1995. 48f. Dissertação de Mestrado – Centre for Cognitive Science, University of Edinburgh. Edinburgh, Scotland, 1995.

Ó SÉAGHDHA, Diarmuid. Annotating and Learning Compound Noun Semantics. In: THE ACL 2007 STUDENT RESEARCH WORKSHOP, Prague, 2007. *Proceedings...* Prague, p. 73-78, 2007.

PUSTEJOVSKY, James. *The Generative Lexicon.* Cambridge: MIT, 1995. 298p.

PUSTEJOVSKY, James. The Generative Lexicon. *Computational Linguistics*, v. 17, n. 4, p. 409-440, dez. 1991.
RYDER, Mary Ellen. *Ordered Chaos:* The Interpretation of English Noun-Noun Compounds. Berkeley: University of California, 1994. 449p.
SANDMANN, Antônio José. *Morfologia Geral.* 3. ed. São Paulo: Contexto, 1997. 79p.
SANTORINI, Beatrice. *Part-of-Speech Tagging Guidelines for the Penn Treebank Project.* Pennsylvania: Departament of Computer & Information Science, 1990. 32p.
SCOTT, Mike. *WordSmith Tools 5.0.* Disponível em: <http://www.lexically.net/downloads/version5/HTML/index.html>. Acesso em: 22 dez. 2007.
SMARSARO, Aucione. Um estudo de palavras compostas com estrutura N de N para processamento automático. *Revista Palavra*, Rio de Janeiro, n. 12, p.164-171, 2004.
TEIXEIRA, Lílian Figueiró; CHISHMAN, Rove Luiza de Oliveira. Um estudo do córpus COMPARA: a semântica dos compostos nominais. In.: COSTA, Luis; SANTOS, Diana; CARDOSO, Nuno. (Ed.). *Perspectivas sobre a Linguateca:* Actas do encontro Linguateca: 10 anos. Porto: Linguateca, 2008, p. 35-41.
VALE, Oto Araújo. *Expressões cristalizadas do português do Brasil:* uma proposta de tipologia. 2001. 213f. Tese (Doutorado em Linguística e Língua Portuguesa) - Faculdade de Ciências e Letras, Universidade Estadual Paulista Julio Mesquita Filho. Araraquara, 2001.