Um estudo do córpus COMPARA: a semântica dos compostos nominais

Lílian Figueiró Teixeira
Universidade do Vale do Rio dos Sinos - Brasil
lilianjoy@gmail.com

Rove Luiza de Oliveira Chishman
Universidade do Vale do Rio dos Sinos – Brasil
rove@unisinos.br

1 O córpus COMPARA e o estudo semântico

Com os dados disponíveis no córpus paralelo COMPARA [3, 5], é possível estudar uma série de fenômenos lingüísticos a partir de equivalências de tradução nas línguas portuguesa e inglesa. Neste trabalho, dedicamo-nos ao estudo da semântica dos compostos nominais, tendo como ponto de partida a sua tradução do inglês para o português. As linhas de concordância obtidas no sítio do COMPARA serviram como ponto de partida para o estudo da semântica dos componentes destas construções e a identificação de padrões de tradução destes compostos. Salientamos que, por padrão, compreendemos apenas características semânticas que sejam recorrentes nos resultados de tradução na língua alvo, neste caso, no português.

Os compostos nominais são extremamente produtivos na língua inglesa, o que representa um desafio para os sistemas de análise e produção da linguagem natural, em especial para a tradução automática. A grande dificuldade encontrada por um sistema de tradução automática é o reconhecimento de mais de duas palavras como uma unidade. Uma frase como *I went to the night school* seria traduzida como *Eu fui à escola de noite*, pois *night* não seria identificado como um modificador de *school*. Este sistema não chegaria à tradução esperada *escola noturna*.

Nosso interesse, neste trabalho, é fazer um estudo da semântica dos compostos formados por dois substantivos (NN), identificando, dentre as abordagens que se ocupam deste fenômeno, as que se prestam à sua representação. Considerando como motivação as tarefas de processamento computacional, vale salientar que algumas das dificuldades em processar os compostos estão relacionadas à complexidade deste fenômeno lingüístico, o que se evidencia na própria diversidade de tratamento teórico que o fenômeno vem recebendo. Alguns estudos elegem a teoria do Léxico Gerativo [7] como um modelo representativo para os compostos. É o caso de [2], que analisaram, com base nos papéis da estrutura qualia, ocorrências nas línguas inglesa e italiana a fim de identificar os padrões semânticos dos compostos. [4] segue na mesma linha, adaptando a estrutura qualia para uma classificação dos compostos NN, utilizando-a como base, mas incluindo outras classes que dêem conta dos dados do estudo.

Neste estudo, seguimos [1], que consideram que um composto formado por dois substantivos (NN) em inglês apresenta um pré-modificador seguido por um substantivo núcleo. Adotamos a estrutura qualia para a interpretação dos dados, mas outras etiquetas semânticas também são consideradas, como as que foram propostas por [6], tais como tempo, posse e local. Optamos por esta abordagem, já que os papéis qualia não cobrem os diferentes tipos de compostos. Em língua inglesa, geralmente o modificador é o substantivo da esquerda e o núcleo é o da direita. Em *samba school*, *samba* é o

modificador e *school* o seu núcleo. A ordem muda em português, pois o modificador aparece após o seu núcleo, conforme visto em *escola de samba*.

Um outro conceito importante, quando trabalhamos com compostos, são as *core words*, traduzidos aqui como *nódulos*. Segundo [8], tanto o modificador quanto o núcleo podem ser o nódulo, já que é esta palavra que vai ser encontrada em outros compostos, formando o que chamamos de *família de compostos*. Assim, uma palavra como *school* serve de nódulo participando de diferentes compostos, tais como *grammar school, summer school, law school, sister school, pottery school* e *state school*.

2. Extração dos dados do COMPARA

A ferramenta de busca desenvolvida pela equipe do COMPARA se mostrou extremamente útil e capaz de fornecer os dados necessários para a realização deste estudo. Precisávamos extrair seqüências de dois substantivos em inglês seguidos pela sua tradução em português. O fato de o córpus estar etiquetado foi o que possibilitou este tipo de busca. Para obter estas informações, adotamos os seguintes passos:

i) Foi feita uma busca por linhas de concordância em que dois substantivos aparecem juntos. Consideramos tanto os substantivos no singular quanto no plural e, com a fórmula **[pos="N.*"][pos="N.*"]** digitada na busca avançada, obtivemos, como resultado, 32.216 ocorrências. A partir deste primeiro resultado, alguns nódulos recorrentes foram selecionados: *hall, room, house, door, floor, table, window* e *school*.

ii) As linhas de concordância para cada combinação das palavras de busca seguidas ou antecedidas por outro substantivo foram analisadas, e os equivalentes de tradução foram identificados. Utilizamos a fórmula **[pos="N.*" & word="school"] @[pos="N.*"]** para cada busca, sendo que os diferentes nódulos eram digitados onde se encontrava a palavra *school*. Também invertemos a ordem do nódulo e selecionamos as expressões com um número maior de resultados. Um outro recurso interessante no sítio do COMPARA é a possibilidade de visualização do número de ocorrências e da lista de palavras que ocupam o lugar de N em cada fórmula. Isto é possível através dos itens "especifique os resultados" e “distribuição dos lemas” encontrados no formulário de busca avançada. O quadro abaixo sistematiza estes primeiros resultados.

Composto	**Ocorrências**	**Exemplo**
N hall	67	*concert hall*
N room	78	*hotel room*
N house	226	*country house*
N door	135	*kitchen door*
N floor	54	*ground floor*
N table	94	*dinner table*
N window	55	*train window*
school N	47	*school gate*
N school	59	*summer school*

Compostos nominais do córpus

iii) Como o objetivo do estudo é analisar os compostos nominais formados por dois substantivos, os compostos formados por mais de dois substantivos e os com algum elemento deverbal (como –*ing*) foram excluídos.

iv) Para uma melhor visualização das opções de tradução para cada composto, os equivalentes de tradução de uma mesma expressão foram agrupados em um arquivo separado.

3. Análise das relações semânticas

Feita a extração dos compostos, passamos para a análise. Valemo-nos dos papéis télico e constitutivo, tal como propostos por [7] na formulação da estrutura qualia. Por papel télico, compreende-se que um dos elementos expressa a função ou propósito do composto, geralmente o modificador se presta a isto. Já o constitutivo estabelece a relação entre o todo e as suas partes. Também utilizamos as categorias de posse, local e tempo, tal como propostas por [6]. Nosso propósito, a partir deste estudo semântico dos compostos, foi verificar como estes sentidos vêm a se expressar nos equivalentes de tradução. A seguir, apresentamos um quadro sistematizando esta análise comparativa a partir dos papéis semânticos.

Composto	**Exemplos**	**Tradução**	**Relação semântica**
N hall	concert hall	sala de concertos	papel télico
	entrance hall	átrio de entrada/entrada	papel télico
	church hall	salão de igreja	papel constitutivo
	parish hall	salão paroquial	papel constitutivo
	school hall	salão da escola/refeitório	papel constitutivo
N room	hotel room	quarto de hotel	papel constitutivo
	laundry room	quarto de engomados/lavanderia	papel télico
	emergency room	pronto-socorro	papel télico
N house	station house	delegacia	papel télico
	summer house	casa de verão	tempo
	family house	casa da família	posse
	brick house	casa de tijolo	papel constitutivo
	beach house	casa da praia	local
	hen house	galinheiro	papel constitutivo
N door	kitchen door	porta da cozinha	papel constitutivo
	trap door	alçapão	papel constitutivo
	glass door	porta de vidro/porta envidraçada	papel constitutivo
	street door	porta da rua	local
	garden door	porta que dava para o jardim	local
N floor	ground floor	andar térreo	local
	kitchen floor	chão da cozinha	papel constitutivo
	metal floor	chão metálico	papel constitutivo
N table	kitchen table	mesa da cozinha	papel constitutivo
	bedside table	mesa-de-cabeceira	local
	dinner table	mesa de jantar	papel télico
	coffee table	mesinha	papel télico
	tin table	mesa metálica	papel constitutivo
N window	kitchen window	janela da cozinha	papel constitutivo
	picture window	janela panorâmica	papel télico
	ticket window	guichê	papel télico
school N	school holiday	férias	tempo
	school gate	portão do colégio	papel constitutivo
	school report	boletim escolar	local
N school	summer school	curso de verão	tempo

	night school Sunday school	escola noturna escola dominical/catequese	tempo tempo

Análise dos compostos

A relação entre os dois substantivos de uma expressão composta, na maioria dos casos, pode ser explicada através de dois papéis da estrutura qualia, o constitutivo e o télico. Em *dinner table*, o substantivo modificador (N1) indica o propósito desta mesa, que é o de ser utilizada durante a janta. Já o papel constitutivo estabelece a relação entre o todo e as suas partes, como em *school gate*, em que *portão* é parte de *escola*.

Analisando os equivalentes em português, identificamos diversos significados para a preposição "de" como parte de uma expressão composta. Além dos papéis constitutivo e télico, identificamos a relação de posse e outras relações como tempo e local. Sentimos a necessidade de incluir estas relações na análise, por não conseguir incluir os exemplos nos papéis e por percebermos uma relação diferente entre os substantivos. Se em *church hall* interpretamos que o salão faz parte da igreja, em *street door* não temos a mesma relação. Não se pode dizer que a porta faça parte da rua, no entanto, o que importa é o fato de alguém poder chegar até a rua ao passar por está porta. Desta forma, a localização da porta é o que motiva a criação deste composto. Os compostos que trazem alguma informação relacionada ao tempo, como em *summer school* e *sunday school*, também não se ajustaram aos papéis estudados e mereceram uma classificação diferenciada. Entre os casos estudados, houve apenas uma única ocorrência em que a relação de posse pudesse ser percebida. Uma *family house* pode ser interpretada como uma casa que pertence à família.

Algumas vezes, os substantivos modificadores são traduzidos como um adjetivo em português. Se é possível traduzir o composto de duas formas, N de N ou N Adjetivo, os dois casos são encontrados no córpus. Geralmente o uso do adjetivo está relacionado a algum material do qual o objeto é feito. Exemplos deste caso são *metal floor* e *tin table*, cujos equivalentes de tradução são *chão metálico* e *mesa metálica*. Quando não há um adjetivo correspondente em português para o material, mantém-se a construção N de N (*brick house*). Como uma casa de tijolo possui tijolos, consideramos que o modificador representa o papel constitutivo.

Quando existe uma única palavra em português correspondente ao composto em inglês, o seu uso é preferido. Enquanto há três ocorrências para *lavanderia*, *quarto de engomados* só aparece uma única vez. Outros equivalentes são escolhidos, pois se percebe certo grau de lexicalização no seu uso. *Coffee table* foi considerado um composto télico, pois é uma mesa utilizada para servir café. No entanto, se observarmos o seu equivalente (*mesinha*), a informação mais importante aqui não é o seu uso, mas o seu tamanho.

4. Considerações Finais

O estudo aqui empreendido e a definição de uma tipologia semântica para descrever os compostos nominais do tipo NN em inglês e seus correspondentes em português pode servir de base para pesquisas voltadas para o aprimoramento de sistemas de tradução automática. Quando padrões da língua são conhecidos, é possível identificar automaticamente os compostos e criar léxicos que possam ser usados em tarefas relacionadas ao processamento da língua natural.

O acesso a um córpus paralelo se mostrou útil para um estudo bilíngüe, podendo contribuir para outros estudos sobre diferentes fenômenos lingüísticos e inclusive multilíngües. Cumprimentamos a iniciativa dos organizadores do córpus COMPARA

em compilar este material e disponibilizá-lo gratuitamente. A comunidade acadêmica carece de recursos desta qualidade e de livre acesso. Sugerimos a disponibilização de alguma ferramenta ou documento que apresente uma lista de palavras do córpus. Para este estudo em especial, uma lista com dois substantivos que ocorrem juntos seguidos pela sua freqüência no córpus teria ajudado.

O foco deste trabalho foi verificar as equivalências de tradução considerando o inglês como língua fonte e o português como língua alvo. No entanto, acreditamos que seja interessante, para um futuro estudo, analisar como os compostos são traduzidos do português para o inglês. Observar quais os equivalentes de tradução em inglês dos compostos formados por NdeN na língua portuguesa poderia ser um propósito de estudo. Também não procuramos separar os resultados de acordo com as variantes da língua portuguesa, português europeu e brasileiro, pois com isso acabaríamos diminuindo os dados de estudo. No entanto, a ferramenta de busca do COMPARA poderia trazer apenas os resultados de uma variante específica.

Referências

[1] Ken Barker e Stan Szbapakowicz. Semi-Automatic Recognition of Noun Modifier Relationships. Em *Proceedings of COLING-ACL '98*, Montreal, 16 de Agosto de 1998, p. 96-102.

[2] Federica Busa e Michael Johnston. Qualia Structure and the Compositional Interpretation of Compounds. Em Evelyne Viegas, organizadora, *Breath and Depth of Semantic Lexicons.* Kluwer, Londres, Inglaterra, 1999, p. 167-187.

[3] COMPARA 10.1.2. http://www.linguateca.pt/COMPARA/

[4] Ann Copestake. Compounds revisited. Em *2nd International Workshop on Generative Approaches to the Lexicon, GL'2003*, Genebra, 15-17 de Maio de 2003. CD-ROM.

[5] Ana Frankenberg-Garcia e Diana Santos. COMPARA, um corpus paralelo de português e de inglês na Web. *Cadernos de Tradução*, IX(1): 61-79, 2002.

[6] Roxana Girju, Dan Moldovan, Marta Tatu e Daniel Antohe. On the semantics of noun compounds. *Computer Speech and Language*, 19:479-496. Março, 2005.

[7] James Pustejovsky. *The Generative Lexicon.* MIT Press, Londres, Inglaterra, 1995.

[8] Mary Ellen Ryder. *Ordered Chaos: The Interpretation of English Noun-Noun Compounds*. University of California Press, Berkeley, Estados Unidos, 1994.