

AN ENGLISH-PORTUGUESE PARALLEL CORPUS FOR THE STUDY ON THE SEMANTICS OF NOUN-NOUN COMPOUNDS

Lilian Figueiró Teixeira • Unisinos
lilianjoy@gmail.com

Rove Luiza de Oliveira Chishman • Unisinos - Pq CNPq
rove@unisinos.br

Universidade do Vale do Rio dos Sinos - UNISINOS

INTRODUCTION

This study suggests a semantic analysis of the noun-noun compounds in English in a magazine context, presenting an initial reflection on the translation correspondents.



fruit bat



Ice age



rice wine

THEORIES AND CONCEPTS

- **Noun-noun compounds:** two nouns (NN) in which there is a pre-modifier followed by a head noun (BARKER; SZPAKOWICZ, 1998).
- **Qualia structure:** Johnston and Busa (1999); Generative Lexicon theory (PUSTEJOVSKY, 1995).
- **Parallel Corpus:** each line of an original text is aligned with its translation in the second language (BERBER SARDINHA, 2004).
- **Translation equivalence:** as part of a translational unit, and the textual context has to be taken into account (HERNÁNDEZ, 1996).
- **Semantic patterns:** Ryder (1994) suggests a classification based on linguistic templates, which are schemas that include semantic characteristics of the components and of the compound structures.
- **Frames:** provide a description of a specify context, through the identification of related actors and lexical units (FILLMORE, 1982).

OBJECTIVE

To identify how the semantics of the compounds could suggest some predictability on the structure of the noun compounds in the target language.

METHODOLOGY

- Ten editions of the **National Geographic magazine**¹ - English and Portuguese, from August 2007 to May 2008.
- Itemization and morphological annotation of the articles in English (**TreeTagger**²).
- An extractor was used, a tool that provides a list of possible compounds, by Lucas Lermen.
- Then some semantic patterns were identified (RYDER, 1994) - **WordSmith Tools**³ for the concordance lines.
- The two texts were machine aligned (**Vanilla Alligner**⁴) and the translation equivalents were identified.

Results	Methodology
4,693 possible compounds	Extractor - NN sequences
1,641 possible compounds	Core words (RYDER, 1994) - 10 or more times in the corpus
842 NN compounds	Human conference
200 NN compounds	Randomly selected

MOST FREQUENT SEMANTIC RELATIONS

• N. OF OCCURRENCES	• RELATION	• ENGLISH	• PORTUGUESE
40	IS LOCATED IN	school play	peça escolar
37	HAS CONSTITUENT PART	church floor	solo da igreja
25	SERVES TO	car keys	chave do carro
9	COMES FROM	cane juice	caldo da cana
9	HAPPENS IN - TIME	night school	escola noturna

¹NATIONAL GEOGRAPHIC MAGAZINE: <http://ngm.nationalgeographic.com/ngm/2007-11/tableofcontents.html>. REVISTA NATIONAL GEOGRAPHIC BRASIL: <http://nationalgeographic.abril.uol.com.br/home/index_0711.shtml>.
²Website: <http://www.ims.uni-stuttgart.de/projekte/Corplex/TreeTagger/>.
³Website: <http://www.lexically.net/downloads/version5/HTML/index.html>.
⁴Website: <http://www2.lael.pucsp.br/corpora/alinhador/index.html>.

RESULTS

- 91 of the 165 NN compounds translated present the structure formed by noun, preposition "de" and noun. • Other structures: "N adjective", "N para N", "N em N", "N d' N", "N verb N", "N de V" and "N para V". Sometimes one noun in Portuguese. • SERVES TO: many translation equivalents are formed by noun and adjective, which sounds much more natural than a NN construction.

Many, however, still live on the fringes of society, relegated to manual labor and barred from obtaining business licenses , government jobs, or access to higher education.	Mas muitos ainda vivem à margem da sociedade, relegados ao trabalho braçal e impedidos de obter licença para abrir negócios , de ter emprego público ou acesso à educação superior.	Paradise nowadays is finding a free spot in the crowded car park .	NO TRANSLATION.	The son never cut down his father's coffin tree to have it made into a coffin.	O filho não abatera a árvore funerária do pai para fazer um caixão.	Here's what I discovered: First, empty the day pack of everything, except for the sandwich, trail mix, and water.	Eis o que eu descobri: primeiro, tire tudo da mochila , menos o seu sanduíche, seu lanche e água.
The farmers tallied their losses: homes, pigs, farm tools , grain sheds, and the woven clothes and silver heirlooms of grandmothers and mothers.	Os agricultores avaliam seus prejuízos: casas, porcos, ferramentas agrícolas , depósitos de grãos, as roupas tecidas e as heranças de prata de suas avós e mães.	Ethanol and biodiesel are now made from food crops like corn and soybeans, but in principle any plant material will do.	O etanol e o biodiesel, produzidos a partir de cereais alimentícios como milho e soja, em princípio poderiam ser feitos com qualquer vegetal.	'He's wearing his school uniform ,' says Shawrieh.	'Ele está de uniforme escolar ,' diz Shawrieh.	And robots don't need space suits , radiation shields, toilets, exercise bikes, a bail-out system during launch, or any consumables to speak of except energy.	Além disso, robôs não necessitam de trajes espaciais , escudos anti-radiação, banheiros, sistemas de escape em caso de acidente nem de comida, exceto energia.

CONCLUDING REMARKS

The preposition "de" in Portuguese can represent several semantic relations and deserves more studies on its polysemy. An important aspect on the translation is the cultural influence on the translator options. **Contributions:** to improve the language processing tasks, such as machine translation systems.

REFERENCES

- BARKER, Ken; SZBAPAKOWICZ, Stan. Semi-Automatic Recognition of Noun Modifier Relationships. In: COLING-ACL '98, 1998, Montreal. Proceedings... Montreal: ACL, 1998. p. 96-102.
- BERBER SARDINHA, Antonio Paulo. *Linguística de Corpus*. São Paulo: Manole, 2004.
- HERNÁNDEZ, Chantal Pérez. A Pilot Study on Translation Equivalence between English and Spanish. *International Journal of Lexicography*, v. 9, n. 3, p. 218-237, 1996.
- JOHNSTON, Michael; BUSA, Federica. *Qualia Structure and the Compositional Interpretation of Compounds*. In: VIEGAS, Evelyn (org.). *Breath and Depth of Semantic Lexicons*. London: Kluwer, 1999. p. 167-187.
- FILLMORE, Charles J. *Frame Semantics*. In: LINGUISTICS Society of Korea (ed.). *Linguistics in the Morning Calm*. Seoul: Hanshin, 1982. p. 111-137.
- PUSTEJOVSKY, James. *The Generative Lexicon*. Cambridge: MIT, 1995. 298p. RYDER, Mary Ellen. *Ordered Chaos: The Interpretation of English Noun-Noun Compounds*. Berkeley: University of California, 1994. 449p.
- TEIXEIRA, Lilian Figueiró. *A semântica dos compostos nominais: um estudo de corpus paralelo inglês/português*. 2009. 209 f. Dissertação (Mestrado em Linguística Aplicada) - Programa de Pós-Graduação em Linguística Aplicada, Universidade do Vale do Rio dos Sinos, São Leopoldo, RS, 2009. Orientação de Rove Luiza de O. Chishman.

UNISINOS

CNPq

Santander