Análise lexicográfica na FrameNet Brasil

Michele Monteiro de Souza*

RESUMO: O presente artigo traz uma pesquisa realizada no Projeto FrameNet Brasil, baseada na Semântica de *frames* e em dados colhidos em *corpora diversos*, que passam por um processo metodológico de análise e descrição lexicográfica de unidades lexicais que evocam o frame de "PLACING". O objetivo das anotações das UL's é a construção de um dicionário de frames da Língua Portuguesa.

PALAVRAS-CHAVE: FrameNet Brasil; Análise lexicográfica; Frame de *placing*; Unidade lexical.

ABSTRACT: This article presents a research conducted by the FrameNet Brasil Project, based on Frame Semantics and on data collected corpora, which go through a methodological process of lexicographic analysis and description of lexical units evoking the "PLACING" frame. The goal of this analysis is to build a dictionary of frames for the Portuguese language.

KEY-WORDS: FrameNet Brasil; Lexicographic analysis; Placing frame; Lexical units.

Introdução

O Projeto FrameNet Brasil objetiva criar um recurso lexical on-line disponível para pesquisas sobre o Português do Brasil, baseado na semântica de frames e sustentado por evidência colhida em corpus, além disso, as unidades lexicais devem ser coerentes em relação ao frame que evocam e perfilam. (FILLMORE, 1982; 1985; GAWRON, 2008).

A análise lexicográfica, dentro deste quadro teórico-analítico, consiste em levantar possibilidades combinatórias sintático-semânticas, por meio das realizações de valências das Unidades Lexicais.

No presente artigo, temos a análise de quatro Unidades Lexicais que evocam o frame de "PLACING", a saber: *colocar*, *guardar*, *ensacar* e *esconder*.

* Aluna graduanda em Letras na Universidade Federal de Juiz de Fora, e-mail para contato: michelemonteiro.mm@gmail.com

1 O Projeto FrameNet Brasil

A FrameNet Brasil é um projeto desenvolvido em parceria com o trabalho do Professor Charles Fillmore, a FrameNet, com sede na University of California, em Berkeley. A partir do projeto-matriz, realizamos nossa pesquisa na Universidade Federal de Juiz de Fora, visando, assim como é feito para a língua inglesa, construir um dicionário de frames da língua portuguesa.

O projeto baseia-se na Semântica de *Frames*, a partir da concepção dada por Fillmore de que "***significações são relativizadas a cenas"*** (FILLMORE 1977). Assim, temos "frame" como um "*esquema imagético*". Exemplifico essa perspectiva cognitiva com o clássico exemplo do **frame visual** do Triangulo Retângulo: só é possível compreender o lexema *hipotenusa* se a cognição evocar o frame do Triangulo Retângulo, ou seja, o *frame* torna possível o perfilamento do lexema.

A partir da escolha de um frame, selecionamos Unidades Lexicais (doravante UL's) e desenvolvemos um processo de análise lexicográfica, esse procedimento segue a metodologia da FrameNet americana, descrita em uma espécie de manual nomeado como "The Book" (RUPPENHOFER *et al*., 2006).

2 O Processo de Anotação

As averiguações partem da escolha de Unidades Lexicais que evocam determinado *frame*, as quais são buscadas em *corpora* do Português do Brasil disponíveis na Linguateca, Ancib, Nilc São Carlos e ECI-EBR; *corpora* públicos e usados com a devida autorização de seus organizadores, ou no Sketch Engine, Legendas de Filmes e Nurc; bem como *corpora* de acesso restrito.

Nos *corpora* realizamos uma busca por lexemas, a partir do lema ou da raiz. As sentenças obtidas são classificadas no programa Excel como (1) Verbo com Sentido Físico, (2) Verbo com Sentido Figurado, (3) Adjetivo, (4) Substantivo, (5) Contexto insuficiente ou ambíguo e (6) Outros. Após a classificação, processamos os dados no software "R", o qual separa todas as ocorrências de sentido físico, válidas para a anotação, e gera um relato estatístico de todas as demais.

O processo de anotação é feito a partir da postulação de três camadas principais: Elemento do Frame, ou EF, que corresponde a uma função semântica microtemática, e também a Função Gramatical e o Tipo de Sintagma da realização lingüística do EF. A

anotação se estrutura formalmente em linhas, ou camadas, que são a sentença e as categorias da anotação, e as colunas, que são os termos anotados. Esse procedimento nos permitirá obter todas as possibilidades combinatórias de cada UL dentro do *frame*, produzindo como resultado todos os possíveis padrões sintáticos e semânticos de ocorrência da UL na variedade brasileira da Língua Portuguesa.

2.1. O *corpus* do Projeto FrameNet Brasil

É interessante apresentar os *corpora* do Projeto, pois são de onde provém toda a base de dados da pesquisa. Estes são selecionados de forma que sejam totalmente da variedade brasileira do português e que possuam dados dos mais variados gêneros textuais possíveis – vide Tabela 1 –, o que nos suscita um total de quase 72 milhões de palavras. Atualmente, trabalhamos para a expansão desses corpora. A divisão dos adiante se vê:

GÊNERO	TOTAL
Oral	1.186.336
Didático	1.388.660
Jurídico	761.852
Literário	2.425.955
Téc. & Cien.	1.767.565
Jornalístico	27.203.360
Universitário	1.027.908
Mensagem eletrônica	1.202.297
Legendas de filmes	34.800.000
Total de Palavras: 71.763.933	

Tabela 1: Distribuição dos *corpora* utilizados pelo projeto FrameNet BR no que tange aos gêneros.

3 O *frame* de "PLACING"

A escolha pelo frame de "PLACING" partiu da leitura do texto **FrameNet's Frames vs. Levin's Verb Classes** (BAKER & RUPPENHOFER, 2002), o qual compara a separação de lexemas em Classes Verbais pelo critério de alternâncias, feita por Levin, com a postulação semântica de unidades lexicais dentro de frames, que a FrameNet defende por ser a forma de produzir agrupamentos semânticos mais coerentes.

Nossa pesquisa segue sistematicamente a descrição dos *frames* feita pela FrameNet americana, que define o *frame* de "PLACING" como:

> "Generally without overall (translational) motion, an Agent places a Theme at a location, the Goal, which is profiled. In this frame, the Theme is under the control of the Agent/Cause at the time of its arrival at the Goal."

Os Elementos de *Frame* (EF's) Agent (agente) ou Cause (causa), Theme (tema) e Goal (alvo) são nucleares, fundamentais para a instanciação desse *frame*. Temos ainda os elementos periféricos, os quais são sempre informações extratemáticas, como Area , Beneficiary (beneficiário), Degree (grau), Depictive (depectivo), Distance (distância), Manner (modo), Means (meio), Place (lugar), Speed (velocidade) e Time (tempo). Todos os EF's são listados durante a análise lexicográfica, porém não são considerados na constituição dos padrões de combinações sintático-semânticos, por serem muito numerosos e não serem fundamentais para a valência das UL's.

3.1. Unidades Lexicais que evocam o *frame* de "PLACING"

Neste trabalho, selecionamos quatro UL's que evocam o frame de "PLACING", as quais buscamos nos *corpora* já citados anteriormente, realizando o processo metodológico do "The Book" para obter, criteriosamente, a análise lexicográfica de cada uma.

3.1.1. UL *colocar*

A busca pela UL *colocar* resultou 474 sentenças válidas dentre as 1.761 sentenças selecionadas para classificação. Verificando as ocorrências em porcentagem temos 33,35% de ocorrências com Sentido Físico (1), 17,33% com Sentido Figurado (2), 0,45% com Sentido Adjetivo (3), 0,35% com Sentido Substantivo (4), 3,15% com Sentido Insuficiente ou Ambíguo (5), e 44,95% classificadas como Outros (6).

O fato de termos uma quantidade maior de sentenças classificadas como Outros se justifica no uso do lexema "colocar" na Língua Portuguesa, que não evoca apena o frame de PLACING (ver definição no tópico 4 deste artigo), mas suscita também noções, na maioria dos casos verificados na presente pesquisa, de posicionamento, "-Isso o COLOCARIA na frente da maioria dos sujeitos ao teu redor" (retirado do corpus Legenda de Filmes), no caso, pertencente a outro frame.

Quanto aos padrões de valência, temos um total de 26 padrões, sendo 23 totais com a realização dos Elementos de Frame na ordem Agente - Alvo – Tema – somando 471

sentenças anotadas –, 2 totais na forma Agente - Alvo - Tema -Pronome Relativo – com 2 sentenças – e 1 total na forma Causa - Alvo – Tema – com apenas 1 ocorrência. Abaixo apresentamos exemplos dos padrões, os quais nos permitem ver também o processo final de anotação lexicográfica na FrameNet Brasil.

(i) O médico COLOCOU um livro de medicina na mesa e passou 15 minutos lendo-o atentamente .

Camadas	O médico	COLOCOU	um livro de medicina	na mesa
EF	Agente		Tema	Alvo
FG	Ext		Obj	Dep
TS	SN		SN	SP

(ii) O bruxo espera e o vice-bruxo sai de cena voltando logo em seguida com uma cadeira-trono que COLOCA no meio da cena . IND[1]

Camadas		cadeira-trono	que	COLOCA	no meio da cena
EF	Agente=IND	Tema	Tema		Alvo
FG		Obj	Obj		Dep
TS		SN	Relativo		SP

(iii) A incrível capacidade de mobilização da Aliança no Rio de Janeiro COLOCOU nas ruas dezenas de milhares de pessoas que se deslocavam do Estádio para a Feira, da Feira para a Sede da Aliança, a poucos quarteirões de distância uns dos outros, em busca de um lugar para ouvir a carta do Cavaleiro da Esperança.

Camadas	A incrível capacidade de mobilização da Aliança no Rio de Janeiro	COLOCOU	nas ruas	dezenas de milhares de pessoas
EF	Causa		Alvo	Tema
FG	Ext		Dep	Obj
TS	SN		SP	SN

3.1.2. UL *guardar*

Na busca pela UL *guardar*, obtivemos 339 sentenças anotadas, dentro de um total de 1.186 sentenças geradas nos *corpora*, o que, em termos percentuais, corresponde a 30,42% de sentenças válidas, 30,24% com Sentido Figurado, 4,91% com Sentido Adjetivo, 7,18% com Sentido Substantivo, 2,73% classificadas como Contexto insuficiente ou ambíguo, e, por fim, 24,64% pertencentes a outros frames.

[1] Temos no exemplo (ii) um caso com Agente IND (Instanciação Nula Definida), o que ocorre quando o EF não se realiza na sentença, porém é inferível no contexto.

A explicação por termos quase a mesma quantidade de sentenças com Sentido Figurado e com Sentido Físico se remete ao fato de termos inúmeras construções metafóricas da UL *guardar* já canonizadas na Língua Portuguesa como “guardar lembrança” ou “guardar na cabeça”.

O fato mais interessante no processo de anotação da UL *guardar* foi a grande quantidade de padrões resultantes, exatamente 39 realizações de valência possíveis, foi o maior número dentre as UL’s pesquisadas. Contudo, é plausível existirem UL’s com as possíveis variações na grade argumental ainda em maior quantidade, pois, quanto mais oralizados são os gêneros dos *corpora* investigados, maior a flexibilidade para ocorrência de instanciações nulas nos elementos de frame.

Abaixo exemplos do resultado de análise lexicográfica da UL em questão:

(iv) - O material nós GUARDÁVAMOS dentro dessa carteira.

Camadas	O material	nós	GUARDÁVAMOS	dentro dessa carteira
EF	Tema	Agente		Alvo
FG	Obj	Ext		Dep
TS	SN	SN		SP

(v) a casa era velhíssima agora a copa era era a geladeira e acho que tinha o armário que ela GUARDAVA louça assim... essas coisas mas eu acho que não tinha mais nada tudo era muito pequeno... sabe... agora vê que mancada o que tinha IND

Camadas	ela	GUARDAVA	louça	
EF	Agente		Tema	Alvo=IND
FG	Ext		Obj	
TS	SN		SN	

(vi) ... é difícil perguntar pra eles assim... o quê que você fazia de gostoso na infância... nesse estilo assim que a gente né... GUARDAVA aquele dinheirinho né... pra poder comprar uma bala né... um sorvete... eu gostava era de sorvete de limão... que era o mais barato IND INI

Camadas		GUARDAVA	aquele dinheirinho	
EF	Agente=IND		Tema	Alvo=INI
FG			Obj	
TS			SN	

Temos nos exemplos (v) e (vi) casos com Alvo IND (Instanciação Nula Definida), Agente IND e Alvo INI (Instanciação Nula Indefinida), essas instanciações nulas aparecem quando um EF nuclear não está presente na oração, mas é definível (IND) ou indefinível (INI) no contexto em análise.

3.1.3. UL *ensacar*

Na UL *ensacar*, obtivemos um total de 31 ocorrências, sendo, portanto, a UL de menor incidência dentre as pesquisadas, porém o número já é suficiente para analisarmos o comportamento dessa UL na Língua Portuguesa. As ocorrências em Sentido Físico configuram uma média de 56,29%; em Sentido Figurado, 5,55%; como Adjetivo temos uma média de 11,85%; 17,41% como Substantivo e 8,89% classificadas como Outros.

A quantidade de sentenças anotadas nessa UL gerou apenas 5 padrões, o que nos permite, nesse momento, apresentar a finalização do processo de análise de uma UL na FrameNet Brasil, o que não foi possível nas demais unidades devido a abundância de padrões, que formam tabelas que ocupam várias páginas. Temos, então, a Tabela 2, que apresenta o Sumariamento de *ensacar*:

	UL ENSACAR		
Número anotado	**Padrões**		
5 Totais	Agente	Alvo	Tema
(6)	Ext. SN	Incorporado	Obj SN
(1)	INC	Incorporado	Obj SN
(3)	INC	Incorporado	Obj SN
(1)	IND	Incorporado	IND
(1)	Ext. SN	Incorporado	IND

Tabela 2: Sumariamento da UL *ensacar*

A observação da tabela nos revela também uma importante característica da unidade lexical *ensacar*, o elemento de *frame* nuclear Alvo se instancia como “Incorporado”, o que significa que a UL incorpora o elemento Alvo sintática e semanticamente.

O fator sintático se explica pela morfologia da Língua Portuguesa: em *ensacar* temos o prefixo *en-* que indica movimento para dentro (cf. Dicionário Eletrônico Houaiss da Língua

Portuguesa), assim, temos a idéia de *ensacar* como ‘por-no-saco’. Essa particularidade da UL em questão a distingue das demais. Vejamos, então, como se configura a anotação desta:

(vii) -Os homens cavavam e as mulheres ENSACAVAM a terra.

Camadas	as mulheres	ENSACAVAM	a terra
EF	Agente	Alvo (incorporado)	Tema
FG	Ext		Obj
TS	SN		SN

3.1.4. UL *esconder*

A UL *esconder* resultou 954 ocorrências, dentre as quais 51,15% são de Sentido Físico; 14,04%, Figurado; 21,49%, Adjetivo; 2,35%, Substantivo; 0,05% com Contexto ambíguo ou insuficiente; e 10,91% como Outros.

As sentenças válidas se dividiram em 27 padrões de valência, de forma que 22 desses padrões eram com sentenças dispostas sintaticamente como Agente - Alvo – Tema; 3 como Agente - Alvo - Tema - Pronome Relativo 2 como Agente - Alvo - Tema - Pronome Relativo outros 2 como Causa - Alvo - Tema e um como Causa - Alvo - Tema - Pronome Relativo. Uma diferença que distingue essa UL das anteriores é a de termos 2 padrões com o EF ‘causa’, sendo um total de 15 sentenças em que o ‘tema’ está sob o controle da ‘causa’, e não do ‘agente’, vejamos exemplos dessas análises:

(viii) Farsas dentro de celas tentam ESCONDER crime. INI

Camadas	Farsas dentro de celas	ESCONDER	crime.	
EF	Causa		Tema	Alvo=INI
FG	Ext.		Obj	
TS	SN		SN	
Verbo	tentam			

(ix) Estou ESCONDIDA. INC IND INI

Camadas		ESCONDIDA		
EF	Agente=INC		Tema=IND	Alvo=INI
FG				
TS				
Verbo	Estou			

No exemplo (viii) o ‘tema’ está sob o controle da ‘causa’, o ‘alvo’ é INI pois não é inferível dentro do contexto. Já no exemplo (ix) temos um caso de Agente=INC, Instanciação Nula Constitucional. Esse tipo de instanciação ocorre nos elementos ‘agente’ ou ‘causa’ quando se trata de uma construção passiva ou imperativa. Outra característica desses exemplos é a presença de mais uma camada de anotação, que surge da necessidade de representar verbos auxiliares, suporte ou cópula.

Considerações Finais

O trabalho de descrição lexicográfica do *frame* de PLACING contribui para o objetivo final do Projeto FrameNet Brasil, que é o de descrever as estruturas conceptuais da variedade brasileira do Português. Esse propósito é de imensa extensão e estamos apenas no início, considerando que a FrameNet americana já possui um total de mais de 960 *frames* descritoscom 11.600 UL’s listadas.

A pesquisa representa a importância de termos uma estrutura de análise dos esquemas imagéticos própria da Língua Portuguesa, já que, observando o comportamento do *frame* “PLACING” na língua, verificamos várias disparidades em comparação com o mesmo *frame* descrito no Inglês. Exemplos de tais diferenças seriam o processo morfológico de incorporação de EF’s no Português, a enorme quantidade de combinações na grade temática, gerando mais padrões do que os encontrados para a UL correspondente na Língua Inglesa, e ainda a possibilidade de instanciações INI consideravelmente maior que no Inglês.

A próxima etapa do projeto é a expansão de nossos corpora, principalmente do gênero oral, o que nos dará subsídios para uma análise ainda mais criteriosa do comportamento dos *frames* na língua, assim como pesquisar UL’s mais típicas da língua oralizada.

A principal perspectiva é a expansão tanto em quantidade de *frames* quanto em quantidade de UL’s, para que a construção de um dicionário de *frames* da Língua Portuguesa seja cada vez mais satisfatória no âmbito qualitativo.

Referências Bibliográficas

BAKER, Collin F. & RUPPENHOFER, Josef. *FrameNet’s Frames vs. Levin’s Verb Classes*. International Computer Science Institute & University of California, Berkeley, 2002.

FILLMORE, C.J. The case for case reopened. *In:* P. COLE; J. SADDOCK (eds.), *Grammatical relations*. New York, Academic Press, 1977, p. 59-81.

FILLMORE, C.J. Frame semantics. In *The Linguistics Society of Korea. Linguistics in the morning calm*. Seoul, Hashin, 1982, p.111-137.

FILLMORE, C.J. Frames and the semantics of understanding. *Quaderni di semantica*, 1: *6*, 1985, p.222-254.

GRAWRON, Jean Mark. *Frame Semantics*. 2008.
Disponível em http://www.icsi.berkeley.edu/~framenet. Acesso em 27 agosto de 2009.

RUPPENHOFER, Josef; ELLSWORTH, Michael & outros. 2006. *The Book*. Disponível em http://www.icsi.berkeley.edu/~framenet. Acesso em 21 de maio de 2009.

SALOMÃO, Maria Margarida Martins. *FrameNet Brasil: um trabalho em progresso*. 2009. Disponível em http://www.framenetbrasil.ufjf.edu.br. Acesso em 27 agosto de 2009.

Aceito para publicação em 15 de novembro de 2010.