

Milena Uzeda Garrão
Maria Carmelita Pádua Dias
PUC-Rio

Apresentação: Claudia Freitas

**Uma abordagem estatística para a identificação de
colocações verbais usando o projeto AC/DC em
www.linguateca.pt**

Objetivos

- 1) Identificação das colocações verbais (CVs) do tipo V+SN mais freqüentes na língua;
- 2) Proposição de detecção dessas CVs com base em corpus — através de um método estatístico.

Motivação

- A constatação da parca descrição dessas CVs tanto em lexicografia quanto em lexicografia computacional;
- A constatação de que o método geralmente utilizado para identificar essas CVs seja pouco preciso além de moroso;
- A crença de que a descrição de uma língua deve depender do falante desavisado (do *córpus*) e não das intuições interessadas do pesquisador.

O que nós não fizemos para identificar essas CVs

- Não partimos de uma visão representacionista da palavra;
- Não pressupomos uma visão imanentista de significado;
- Não contrastamos o sentido literal com o metafórico;
- Não utilizamos uma visão composicional do significado (opacidade versus transparência semântica).

Por que não utilizamos esta visão?

- Porque parece partir de um conjunto de critérios dedutivos que não prioriza usos reais na língua.

1) Critério de Não substituição das partes

[Bater as botas]

Como explicar, então, *Bater a caçuleta?*

2) Critério de Não inserção de constituinte

[Tomar partido]

Como explicar, então, *tomar muito partido.*

Que método foi utilizado?

- Método não-intuitivo com base em corpus. CETENFolha (Corpus Jornalístico de 24 milhões de palavras em PB) através do projeto AC/DC em www.linguateca.pt
- Método de base estatística que privilegia as CVs mais frequentes no PB.

Resumo do método utilizado:

- 1) uso de *corpus* etiquetado do PB como fonte de dados (Cetenfolha)
- 2) aplicação de um filtro para detecção de todos os padrões V+SN presentes no *corpus*;
- 3) aplicação de um teste estatístico ao filtro, chamado logaritmo de verossimilhança (Banerjee & Pedersen, 2003) para identificar as reais colocações em detrimento de combinações sintáticas casuais;
- 4) edição humana.

Aplicação do teste estatístico para detecção de CVs do tipo V + SN

- Extrator V+SN (Nogueira, 2004)
- Aplicação do teste de Logaritmo de Verossimilhança

$$H1: P(w_1 | w_2) = P(w_1 | \neg w_2)$$

$$H2: P(w_1 | w_2) \neq P(w_1 | \neg w_2)$$

$$H1: P(\text{fazer} | \text{sucesso}) = P(\text{fazer} | \neg \text{sucesso})$$

$$H2: P(\text{fazer} | \text{sucesso}) \neq P(\text{fazer} | \neg \text{sucesso})$$

Que verbos foram testados?

- Após obtenção dos 30 verbos mais frequentes no corpus, restringimos a busca para os 10 verbos mais freqüentes encabeçando a estrutura V+(det)+SN.
- O formalismo para chegar a esses verbos foi viabilizado pelo projeto AC]DC:
- [lema="fazer" & pos="V"] [pos="DET.*"]? [pos="N"] [classe="JOCF"]

São eles: *fazer, ter, dar, perder, usar, receber, deixar, tomar, ganhar, criar* .

Depois de aplicado o logaritmo de verossimilhança, optamos pelas 100 CVs mais frequentes para cada um dos 10 verbos: 1000 CVs

Fragmento do resultado gerado

■ **Teste 1: FAZER +(DET)+N**

fazer campanha,1,117

fazer parte,2,106

fazer sucesso,3,96

fazer sentido,4,94

fazer compras,5,73

fazer falta,6,57

fazer perguntas,7,51

fazer alguma coisa,8,44

fazer política,9,44

fazer as contas,10,42

Edição humana

- Precisão do método: 87, 2%.
- Tipos de ruídos:
 - i) Avaliação estrutural. Este tipo de erro pode ter sido cometido pelo método por duas razões principais: em função da etiquetagem equivocada no *corpus* (ETQ); em função de o método ter considerado uma janela sintática menor do que a expressão representa (JAN): *ter um papel*, por exemplo, foi detectado pelo método como uma colocação do padrão procurado quando, na verdade, sua estrutura vai além de V+(det)+N.
 - ii) Outros ruídos foram atribuídos exclusivamente ao *corpus*: colocações claramente datadas: como *criar a URV*, *usar a URV*, *tomar AZT* (DAT).
- Há outros dois tipos de interferência na detecção de colocações que não foram considerados propriamente ruídos. São eles: recursos coesivos, como a utilização de anáfora: alguns exemplos são *fazer a denúncia*, *dar a notícia*, *ter a doença* (COE) e omissões de artigo (ART) (tanto definido quanto indefinido), características de manchetes de jornal, como “Presidente da Shell deixa cargo amanhã”.

Conclusão

- A língua pode ser descrita como um fenómeno probabilístico;
- Essa perspectiva atenua uma visão chomskiana, focada na semântica do cálculo, e prioriza uma visão de língua inseparável da pragmática; isto é, enfatiza o teor eventivo do fenómeno lingüístico;
- O cópus, além de servir como base de dados para detecção de CVs, também tem um papel preditivo ao fornecer os ambientes lingüísticos tipicamente relacionados às CVs.

Contribuições

- Contribuição para semântica lexical;
- Contribuição para a Lexicografia das CVs;
- Contribuição para a Lexicografia computacional. (ex:tradução automática).

Obrigada, Claudinha!!!!!!!!!!!!!!!!!!!!!!