

VARRA: Validação, Avaliação e Revisão de Relações semânticas no AC/DC

Cláudia Freitas (Linguatca-FCCN / PUC-Rio) Diana Santos (Linguatca-FCCN) Hugo Gonçalo Oliveira* (CISUC – Univ. de Coimbra) Violeta Quental (PUC-Rio)



Motivação

Ontologias, léxicos, taxonomias, tesouros, dicionários, por vezes até obtidos (semi)automaticamente

Como validar os resultados – as relações – gerados?

Pedir um julgamento humano sem contexto? Tarefa estranha para um ser humano!

VARRA: avaliar relações em contexto: mais do que eu "acho", observar se os falantes "usam"

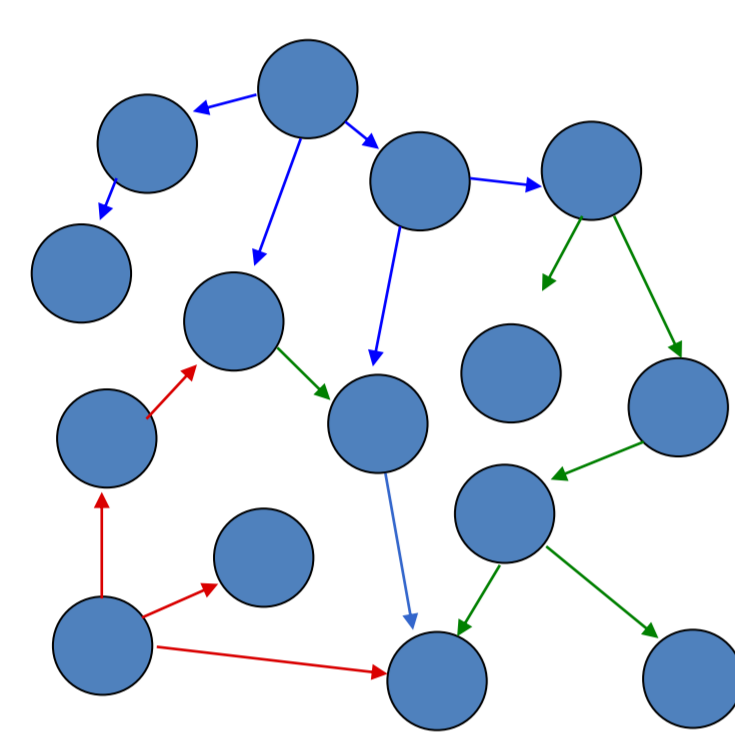
Acesso a Corpora / Disponibilização de Corpora

- 22 corpos em português, variantes portuguesa e brasileira
- Jornalístico, literário, texto didático, entrevistas, listas eletrônicas, web, ... (Costa et al., 2009)
- Cerca de 374 milhões de palavras, 16 milhões de frases
- Anotados gramaticalmente pelo analisador sintático automático PALAVRAS (Bick, 2000)
- Interface na web: <http://www.linguatca.pt/ACDC/>
- Anotação semântica em progresso (Santos, 2010)

Objetivo

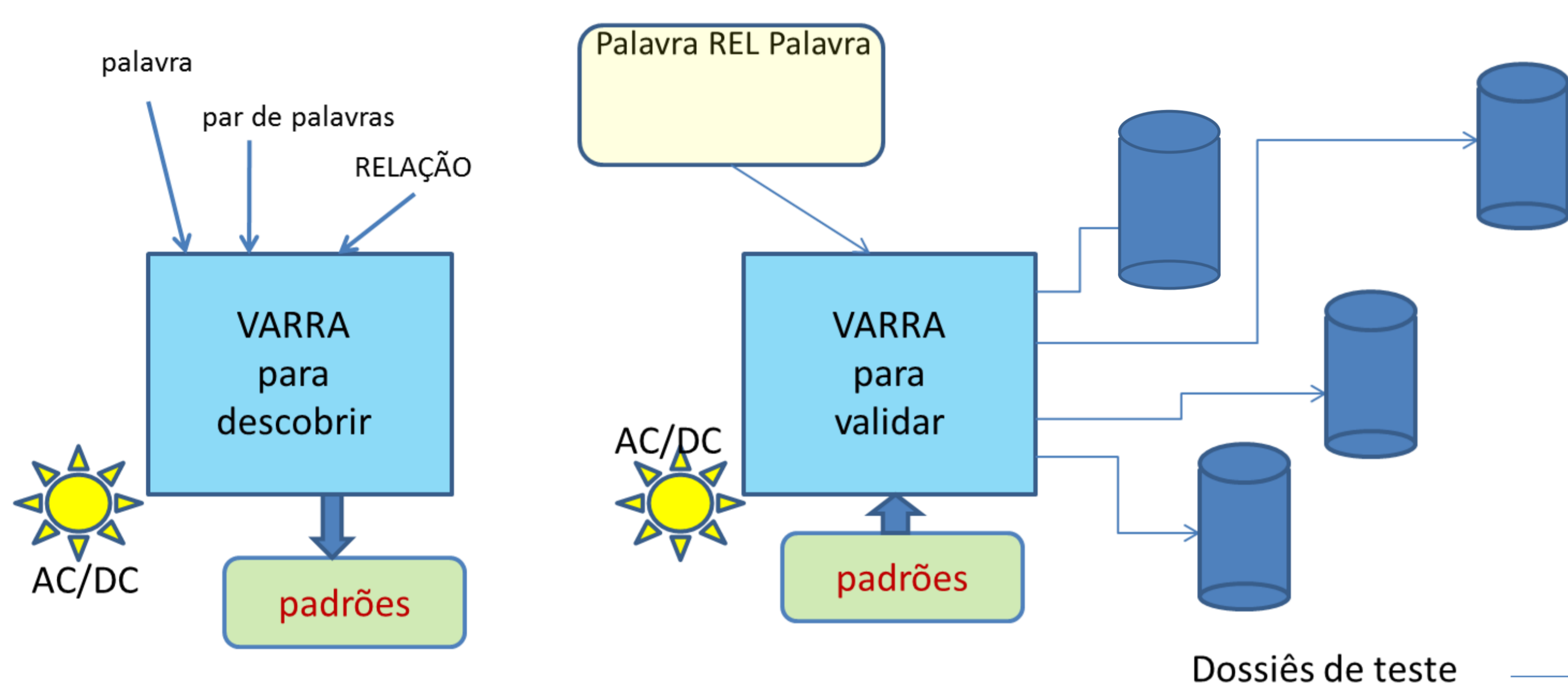
- Criação de questionários para validação de relações semânticas presentes em recursos semânticos (Santos et al., 2010)
- Teste e desenvolvimento de expressões de busca do próprio VARRA (em desenvolvimento)

lema="incluir|conter" & temcagr!="PCP" [pos="ADV.*"]* [pos="DET.*"]* tripo pretendido para validar: palavra1 PARTE_DE palavra2



TeP (Dias da Silva & Moraes, 2003), PAPEL (Gonçalo Oliveira et al., 2010),...

Utilizando o VARRA



Validação e avaliação de relações semânticas

Classificação dos resultados obtidos para a relação de SINONÍMIA entre **mentira** e **ilusão**:

1: Sim

Mudança maior, porém, vem do novo presidente do Supremo Tribunal Federal, ministro Sepúlveda Pertence, que afirmou: "Desde que se superou a mentira de que um juiz, particularmente um juiz constitucional, é um puro técnico capaz de extrair uma norma supostamente de um único sentido válido de um fato, desde que essa ilusão foi desfeita, a verdade é que o juiz é um homem, enquanto cidadão, com crenças, convicções, tendências conscientes e inconscientes."

Não

2: É compatível, mas não exactamente

Uma fábula sobre as relações do real e da ilusão, sobre a mentira, o tempo que volta para trás, a beleza da paternidade e o enigma da fraternidade quebrada.

3: O texto é completamente não relacionado
*Relação de sinonímia entre **grana** e **dinheiro**

**O que não se pode ainda é começar orações com pronomes átonos, nem meter a mão na grana do Orçamento, nem no bolso do povo com novo IPMF e outras marotagens, nem receber dinheiro de lobistas para se eleger.*

4: Pelo contrário, invalida-a
Não era uma mentira, era ilusão, o cinema possui esse poder de criar ilusão.

5: Não sei
Distingue, nas promessas que lhe fazem, o que é mentira ou ilusão.

Referências

Eckhard Bick. *The Parsing System "Palavras": Automatic Grammatical Analysis of Portuguese in a Constraint Grammar Framework*. Dr.phil. thesis. Aarhus University. Aarhus, Denmark: Aarhus University Press. Novembro de 2000.

Luis Costa, Diana Santos & Paulo Alexandre Rocha. "Estudando o português tal como é usado: o serviço AC/DC". In *The 7th Brazilian Symposium in Information and Human Language Technology (STIL 2009)* (São Carlos, SP, Brasil, 8-11 de setembro 2009).

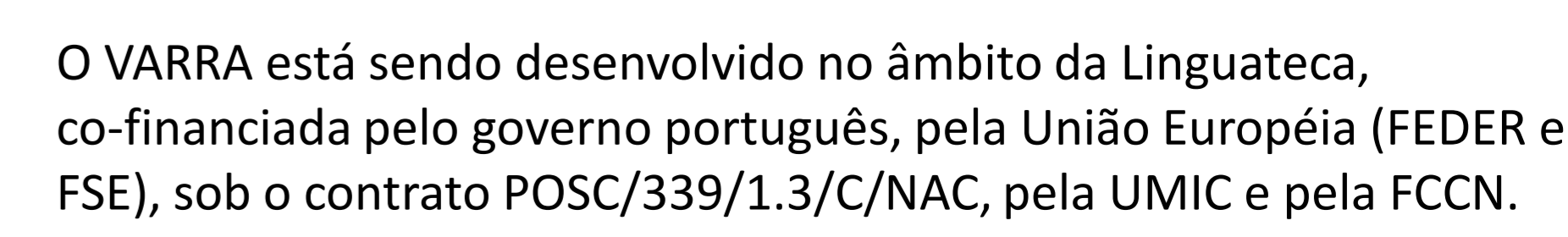
Bento Carlos Dias-da-Silva & Helio Roberto de Moraes. "A construção de um thesaurus eletrônico para o português do Brasil". *ALFA* 47, 2, 2003, pp. 101-115.

Hugo Gonçalo Oliveira, Diana Santos & Paulo Gomes. "Extracção de relações semânticas entre palavras a partir de um dicionário: o PAPEL e sua avaliação". *Linguística* 2, 1, 2010, pp. 77-93.

Diana Santos. "Linguatca's infrastructure for Portuguese and how it allows the detailed study of language varieties". *OSLA: Oslo Studies in Language* 2 (2010).

Diana Santos, Anabela Barreiro, Cláudia Freitas, Hugo Gonçalo Oliveira, José Carlos Medeiros, Luís Costa, Paulo Gomes & Rosário Silva. "Relações semânticas em português". In *Textos seleccionados apresentados ao XXV Encontro Nacional da Associação Portuguesa de Linguística*. APL, 2010.

Para saber mais, consulte o catálogo de publicações da Linguatca com as marcas ACDC e PAPEL



* Hugo Gonçalo Oliveira é Hugo Gonçalo Oliveira é apoiado pela bolsa FCT SFRH/BD/44955/2008, co-financiada pelo FSE.