# VARRA: **V**alidação, **A**valiação e **R**evisão de **R**elações semânticas no AC/DC

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

Fundação para a Computação Científica Nacional

University of Coimbra
Faculty of Sciences and Technology
Department of Informatics Engineering

Knowledge and Intelligent Systems Laboratory
Cognitive and Media Systems Group
Centre of Informatics and Systems of the University of Coimbra

Pontifícia Universidade Católica do Rio de Janeiro

## Motivação

Ontologias, léxicos, taxonomias, tesauros, 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!

↓

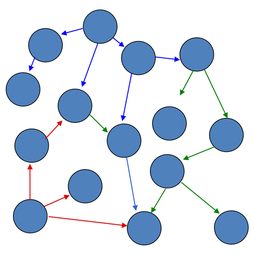
**VARRA**: avaliar relações **em contexto**: mais do que eu "acho", observar se os falantes "usam"

## Acesso a **C**orpora / **D**isponibilização de **C**orpora

- 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.linguateca.pt/ACDC/
- Anotação semântica em progresso (Santos, 2010)

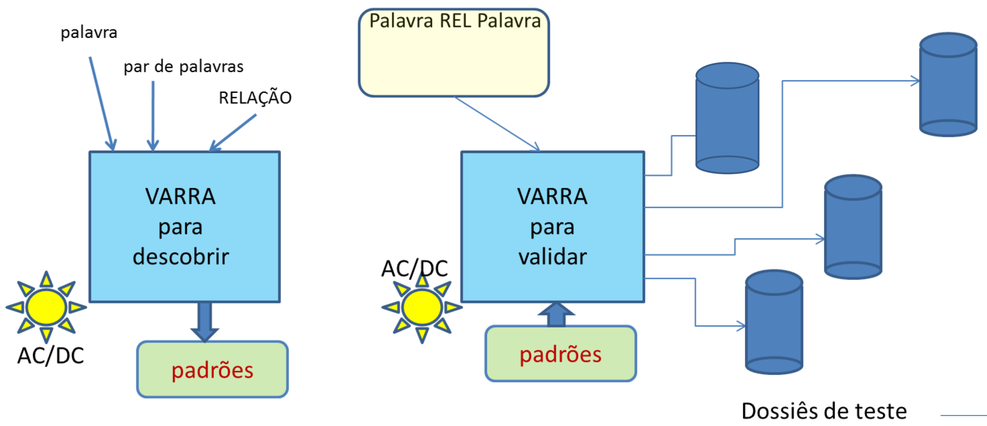
## Objetivo

- 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)

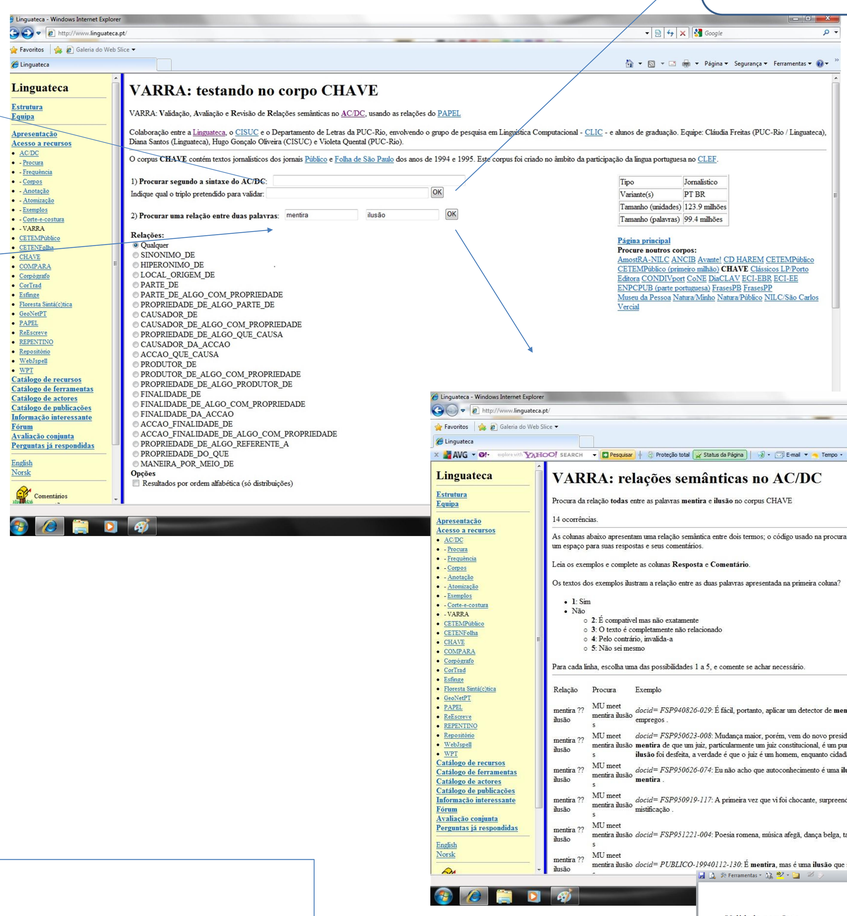


**TeP** (Dias da Silva & Moraes, 2003),
**PAPEL** (Gonçalo Oliveira et al., 2010),…

# Utilizando o VARRA

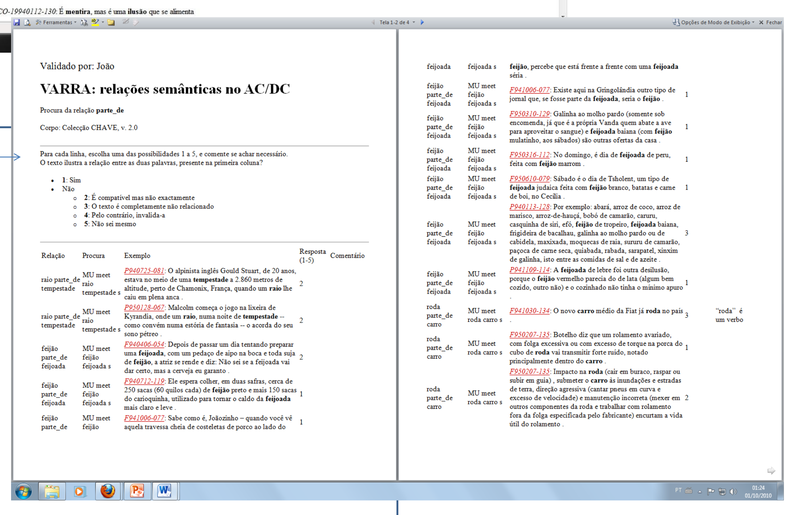



```
lema="incluir|conter" & temcagr!="PCP"] [pos="ADV.*"]* [pos="DET.*"]*
```
triplo pretendido para validar: `palavra1 PARTE_DE palavra2`










## **V**alidação 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** .*

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

## 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.

Luís 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". *Linguamática* **2**, 1, 2010, pp. 77-93.

Diana Santos. "Linguateca'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.

O VARRA está sendo desenvolvido no âmbito da Linguateca, co-financiada pelo governo português, pela União Européia (FEDER e FSE), sob o contrato POSC/339/1.3/C/NAC, pela UMIC e pela FCCN.

POS_Conhecimento · União Europeia FEDER · UMIC · FCT Fundação para a Ciência e a Tecnologia

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