

## Pontes no trabalho de investigação: simpósio doutoral Abril 2006

Diana Santos  
Linguateca  
[www.linguateca.pt](http://www.linguateca.pt)

## Motivação

- Juntos podemos constituir o **círculo da Linguateca**
- Problemas comuns
- Instanciação diferente
  - quer nos problemas investigados
  - quer nos métodos utilizados
- Seguir caminhos paralelos, mas podendo usar simbioticamente o trabalho dos outros e sobretudo o resultado das suas inquirições
  - Recursos
  - Métodos e medidas de avaliação
  - Anotação dos recursos

## Problemas gerais com que todos se defrontam

- Relação entre texto e conhecimento
- O que é uma ontologia (estruturação de conhecimento)?
- Como avaliar o uso da ontologia/EC?
- Como avaliar a própria ontologia/EC?
- Como avaliar os textos/ a anotação/ compreensão dos textos com a própria ontologia?
- É preciso definir questões práticas e questões teóricas definidas
- É preciso pensar a sério no que pode ser implementado, e no que pode ser contrariado (refutado), e como

## Tipos de aplicações em PLN

- LN->LN: Extrair algo
- LN->X->LN: Arrumar algo para subseqüente organização
- X->LN: Ajudar a criar algo
- Extra: descobrir plágio, descobrir um autor, autenticação, etc.
- tarefas com sentido
  - RAP, EI, TA, RI
- tecnologias
  - análise sintáctica, POS tagging,...
- recursos e métodos
  - dicionários, ontologias, aprendizagem automática, prospecção (data mining), ...

## Exemplos de pontes: tarefas semelhantes

- Obtenção de palavras semelhantes
  - em texto livre
  - em dicionários
  - em ontologias
  - em textos paralelos
- Obtenção de relações entre conceitos
  - geográficas
  - metonímicas
  - hiponímicas
- População de estruturas de conhecimento
  - aumentar a base

## Exemplos de pontes: problemas semelhantes

- O que são palavras? (conceitos, termos)
  - desambiguação de sentidos
  - expressões multipalavra
  - radicalização
- Como detectar relações?
  - co-ocorrência
  - intrínseco: comida come-se
  - implícito:
- Qual o limite para a população?
  - às vezes, sempre, todos, relevantes, frequentes
  - erros

## Ambiguidade e vagueza

- Pressupõem uma tradução/classificação
- Ambiguidade e vagueza
  - mais de uma tradução para a mesma entidade
  - a diferença está na relação entre as traduções
- Ambiguidade ou vagueza sintáctica: mais do que uma análise sintáctica
- Ambiguidade ou vagueza semântica: mais do que um sentido
- Em geral as duas coincidem, mas não necessariamente
  - a mesma análise sintáctica mas dois sentidos: *sentei-me no banco*.
  - a mesma análise semântica mas duas estruturas: *sentei-me num banco na cozinha*
- Ambiguidade lexical (que está no léxico)

## Exemplos de pontes: tarefas semelhantes

- Avaliação
  - escolher as especificações antes de trabalhar
- Utilização
  - escolher um aplicação (mesmo que ideal) na qual raciocinar
- Implementação
  - questões práticas
- Idealmente
  - reuso de programas e métodos
  - documentação e depuração de ambos
  - crítica e comparação

## Qual a vantagem de pertencer ao círculo?

- Testadores entusiasmados
- Críticos ferozes
- A possibilidade de comparação com outros domínios (citação cruzada, reutilização dos dados)
- Criação de fundo comum
  - Adamastor
  - WPT03-05/BACO
  - GeoNet, PAPEL
  - Ferramentas: NATools, Esfinge, SIEMES, ...
- A minha orientação

## Problemas

- Separação geográfica
- Pouca comunicação científica (embora bastante técnica!) entre o pessoal júnior
- Desconfiança mútua
- Falta de experiência em co-citação (publicidade em conjunto)
- Espírito de competição
- Co-orientadores diferentes
  - não é garantido que queiram ouvir sequer falar dos outros 6...
- Estilos de trabalho diferentes
- Imprevistos

## Casos concretos, pontes concretas

- Se o Marcirio classificar o BACO com o SIEMES
  - depura o dito
  - cria um novo recurso
- Se o Nuno usar o BACO-siemesado para extrair relações entre entidades
  - depura o dito
  - pode criar classes associadas a relações no PAPEL
  - pode fazer um BACO-papelizado
- Se o Alberto usar o BACO-sieme- ou papelizado para obter exemplos
  - depura ambos
  - cria novas classes que podem ser usadas pelo Marcirio e Nuno Seco
  - pode fazer um BACO-generalizado

## Casos concretos, pontes concretas

- Se o Alberto cria um conjunto de exemplos com base em corpora bilingues
  - pode ser testado pelo Nuno Seco
  - pode ser testado pelo Luis Sarmento
  - pode ser usado para adicionar informação ao BACO
- Se a Anabela descobre regras de parafraseação
  - podem ser testadas para descobrir mais exemplos
- Se a Cristina descobre regras de datação
  - podem ser usadas para avaliar / anotar entradas de dicionário/PAPEL
  - podem ser usadas para indexar textos
  - para prever a evolução

## Casos concretos

- Se o Nuno Seco descobre um método de identificar conceitos semelhantes no dic.
  - pode ser testado pelo Marcirio
  - pode ser testado pelo Luis Sarmento
  - pode ser usado pela Anabela
- Se o Luís Sarmento descobre um método de agrupar contextos
  - pode ser usado pelo Alberto
  - pode ser usado pela Anabela
  - pode ser usado pelo Marcirio
- Se o Marcirio descobre um método de comparar uma ontologia com um conjunto de textos

## Algumas medidas/soluções concretas

- Encontros semestrais; Encontros informais 2 a 2 quando der jeito
- Envio dos artigos/relatórios/ideias para o molhe
  - já feito pelo Nuno Seco e pelo Luís Sarmento, mas eu sugiro os Jânicos...
- Partilha das apresentações com os outros (para isso sugiro o catálogo da Linguateca, com palavra-chave)
- Sugestão de artigos conjuntos (são sempre melhores do que individuais!)
- Criação de um curso em comum
  - métodos estatísticos
  - métodos de avaliação
  - problems típicos