Tecnologia da Linguagem Humana

e o processamento computacional da língua portuguesa

António Branco

U LISBOA | UNIVERSIDADE DE LISBOA

1. **Que é a tecnologia da linguagem?**
2. Como está
3. Como avançar

Tecnologia da linguagem

- **Da forma linguística à representação do significado e vice-versa**
- **Processamento de fala**:
  - Obter uma representação discreta a partir de um sinal analógico
- **Processamento de linguagem**:
  - Obter uma representação do significado a partir de uma sequência de símbolos

Quebrar a última barreira comunicacional

- processamento das expressões linguísticas e do seu significado

- **Crucial**
  - Interagir com serviços e dispositivos artificiais em linguagem natural
    - desde robôs sofisticados a electrodomésticos banais
  - Comunicar com falantes de outras línguas
    - sem haver língua veicular comum

Impactos

- **Económico**
  - inovação e internacionalização no mercado digital global
- **Social**
  - cidadania plena na sociedade da informação
- **Cultural**
  - português, língua internacional de comunicação global

Impacto económico e social

- **Dois exemplos:**
- 1.1 mil milhões de euros
  - custo anuais da Comissão Europeia com multilinguismo

- 8.4 mil milhões de euros in 2008
  - gastos em localização de software a nível mundial

- **Linguagem natural e barreiras linguísticas**
  - grandes oportunidades por explorar em termos de inovação:

  - Tecnologia da linguagem em jogos, pacotes de eutainment, bibliotecas, ambientes de simulação e programas de treino, serviços de informação móveis, software de apoio à aprendizagem de línguas, ambientes de eLearning, software para deteção de plagiarismo, extração de opiniões e análise de sentimentos, busca na web avançada …

Impacto cultural

- **Língua portuguesa**
  - **5ª língua em número de falantes no mundo**
    - chinês, castelhano, inglês, árabe
  - 236 milhões de falantes em 4 continentes
    - crescimento para 335 milhões em 2050
  - linguagem de trabalho em 27 organizações internacionais
  - **era digital impõe choque tecnológico e necessidade de preparação científica e tecnológica**

Aplicações atuais

- **Interação homem-máquina**
  - **Interfaces com dispositivos e agentes artificiais**
  - Deteção de linguagem, autor, domínio,…
  - Classificação de textos
  - Agrupamento de textos
  - Busca de documentos
  - Extração de informação
  - Interfaces com bases de dados
  - Resposta a perguntas
  - Reconhecimento de fala
  - Síntese de fala
  - ...
- **Web**
  - Busca web avançada
  - Anotação de metadados
  - Gestão de ontologias
  - …

- **Interação multilingue**
  - **Tradução automática**
  - Agentes conversacionais
  - Publicação multilingue
  - ...
- **Produção e verificação de linguagem**
  - **Correção ortográfica**
  - Correção gramatical
  - Deteção de plágio
  - Linguagens controladas e sistemas de produção de documentação
  - Localização
  - Legendagem automática
  - Sistemas de ditado
  - Sumarização
  - Geração de relatórios
  - Ambientes de apoio à tradução
  - Simplificação de textos
  - ...
- **Aprendizagem da linguagem**
  - Avaliação de competências
  - Formação
  - …

Tecnologia nuclear

Ferramentas de processamento

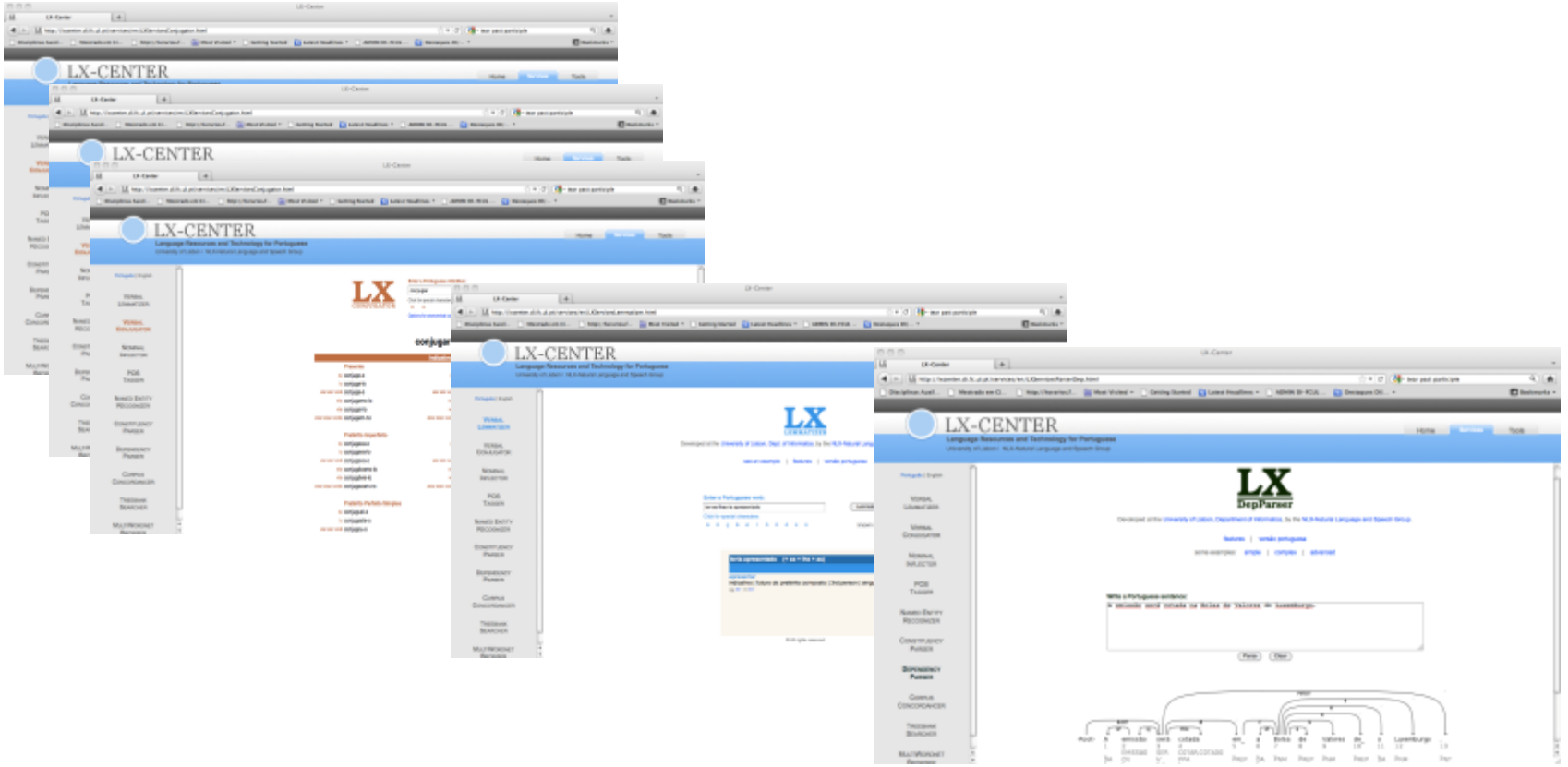
- Separador de frases
- Separador de palavras
- Etiquetador morfossintáctico
- Lematizador
- Analisador morfológico
- Reconhecedor de nomes de entidades
- Desambiguador de aceções de palavras
- Analisador de constituência sintáctica
- Analisador de dependências gramaticais
- Etiquetador de papéis semânticos
- Gramática para processamento linguístico profundo (análise semântica)
- …

Recursos linguísticos

- Corpora anotados
- Corpora paralelos e alinhados
- Bases de dados de fala
- Listas de palavras
  - abreviaturas, nomes próprios, palavras funcionais,…
- Vocabulários
- Léxicos
- Ontologias lexicais
- Terminologias
- Treebanks
- Propbanks
- …

Exemplos

- **LX-Center** `http://lxcenter.di.fc.ul.pt`
- Demos e serviços linguísticos online
  - Uni de Lisboa, Fac. Ciências, Dep. Informática, Grupo de Fala e Linguagem Natural



1. Que é

2. Como está a língua portuguesa preparada para a era digital?

3. Como avançar

Livro Branco

- **eBook**
  - http://metanet4u.eu/ebook
- **Língua portuguesa**
  - 10 coautores; Portugal e Brasil
- **Coleção**
  - 30 línguas; 200+ peritos
  - divulgação, avaliação, recomendação
- **Rede europeia de excelência em I&D**
  - META-NET; 60 centros; 34 países

Preparação tecnológica dos **falantes**?

- **5ª** língua em nº de utilizadores na internet
  - 5ª língua em nº de falantes no mundo
    - 236 milhões de falantes em 4 continentes
    - crescimento para 335 milhões em 2050
    - Chinês, Castelhano, Inglês, Árabe

- **3,8%** utilizadores de internet no mundo
  - 3,7% população mundial

- **34,1%** penetração da internet nos países da CPLP
  - 32,7% penetração no mundo

- 31/12/2012, internet world stats (inglês, chinês, castelhano, japonês)

Porém alguns dados perigosamente ilusórios

- 3ª língua em nº de utilizadores no **Twitter**
  - inglês e castelhano
    - 31/01/2012, semiocast
- 3ª língua em nº de utilizadores no **Facebook**
  - inglês e castelhano
    - 13/11/2012, socialbakers
- **Perigosamente ilusórios**
  - estas são tecnologias largamente independentes da linguagem, e facilmente extensíveis a diferentes idiomas
  - **apenas reflectem o número de utilizadores que são falantes do português com acesso à internet, não a preparação da língua portuguesa para a era digital**

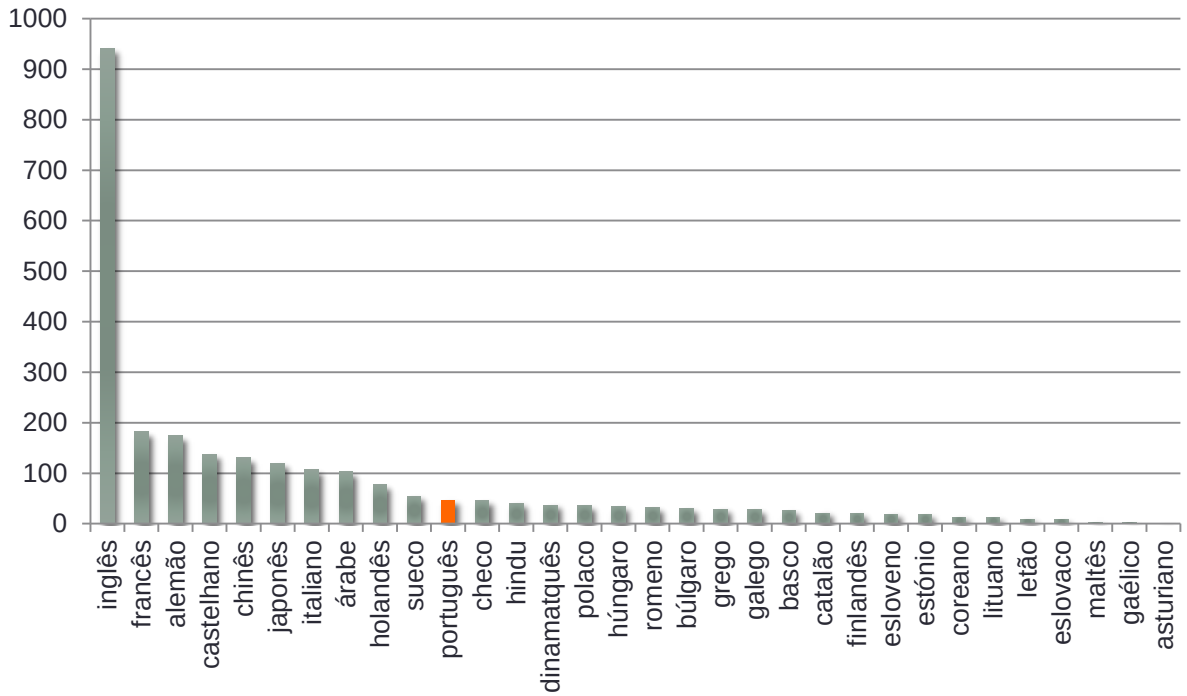
Preparação tecnológica da **língua**? (avaliação qualitativa)

Apoio excelente	Apoio bom	Apoio médio	Apoio fragmentário	Pouco/nenhum apoio
	Inglês	Alemão Espanhol Francês Italiano Neerlandês	Basco Búlgaro Catalão Checo Dinamarquês Eslovaco Esloveno Finlandês Galego Grego Húngaro Norueguês Polaco **Português** Romeno Sueco	Croata Estónio Irlandês Islandês Letão Lituano Maltês Sérvio

10: Análise do Texto: estado da tecnologia da linguagem para 30 línguas europeias

Preparação tecnológica da **língua**? (avaliação quantitativa)

- **Esforço de investigação**
- Referências a línguas em artigos científicos da área, 10 conferências de topo, 2010-2012
  - ≈ Basco, Romeno, ...
  - Holandês 2 x mais, Italiano 2,5 x, Castelhano 3 x, Alemão e Francês 4 x mais
  - **Inglês 22 x mais**



1. Que é
2. Como está
3. **Como fazer avançar o processamento computacional da língua portuguesa?**

Comunidade internacional

- **pilares para I&D**
  - número crescente de publicações e doutorandos
  - grupos e centros especializados em:
    - **Brasil:** Univ São Paulo (Inst Ciências Matemáticas e Computação); Univ Federal São Carlos; PUCRS; UFRGS; Univ Vale do Rio dos Sinos; PUCRJ; Univ Fed Pará; Univ Fed Minas Gerais,…
    - **Portugal:** Univ Lisboa, UNL, Évora, Algarve, Coimbra, Porto, Minho, …

- **comunidade internacional de I&D**
  - alguns projetos de intercâmbio
  - Conferência bienal, “pendular” entre Brasil e Portugal
    - PROPOR International Conference on the Computational Processing of Portuguese
    - São Carlos/2014, Coimbra/2012, Porto Alegre/2010, … Lisboa/1993

Infraestrutura de investigação

- **Infraestrutura de Investigação de Interesse Estratégico**
  - CLARIN
  - disponibilizará recursos e tecnologia a investigadores de todas as áreas disciplinares cujos temas de pesquisa, desenvolvimento ou inovação dizem respeito à língua portuguesa
  - Aprovada em fev 2014; 1º sem 2015 início da implementação

Tecnologia emergente

- **Ferramentas e recursos**
  - Grande complexidade científica e tecnológica
  - Competências de investigação de combinação difícil
  - Custos muito elevados
  - Grande morosidade de construção

- **Potenciar o progresso**
  - Transferência de tecnologia entre academia e indústria
  - Plataforma de distribuição de recursos
  - Procurar efeito “bola de neve” investigação-inovação

Assumir como desafio estratégico prioritário:

- **Processamento computacional da língua portuguesa**

- **Importante**
  - assegurar cidadania na era digital
  - promover inovação e internacionalização na economia digital global

- **Necessário**
  - ninguém fará por nós: investigação com resultados públicos potenciadora de inovação por PMEs

- **Oportuno**
  - juntos faremos mais e melhor por um objetivo comum do que isoladamente

Estímulo interdisciplinar específico

- **Necessários programas coordenados de estímulo** a esta área interdisciplinar

  - Para cobrir:
    - investigação básica
    - desenvolvimento tecnológico
    - empreendedorismo inovador

  - Envolvendo um leque amplo de instrumentos e.g:
    - chamadas de propostas de projetos dirigidas à área
    - bolsas de pós-graduação agrupadas para esta temática
    - ... etc

Obrigado

António Branco

U LISBOA | UNIVERSIDADE DE LISBOA