European Language Resource Coordination
Connecting Europe Facility

Dados e recursos linguísticos para a língua portuguesa

Amália Mendes
Centro de Linguística da Universidade de Lisboa (CLUL)

Recursos para uma língua internacional

- É necessário ter em conta a dimensão internacional da língua portuguesa quando falamos de recursos linguísticos para o português:
  - Língua falada por cerca de 220 milhões de falantes em 4 continentes
  - 3ª língua europeia mais falada no mundo
  - 5ª língua mais usada na internet
  - O uso do português *online* continua em expansão
  - Aumento dos aprendentes de português língua estrangeira

European Language Resource Coordination
Connecting Europe Facility

Recursos para uma língua internacional

- Situação frequente de contacto linguístico: línguas nativas ou crioulos como primeira língua (ex: países africanos com o português como língua oficial)
- Variação na codificação da norma: duas variedades endocêntricas (Portugal e Brasil) e outras variedades exocêntricas que seguem a norma do português europeu, mas esta é uma situação em mudança
- Recursos linguísticos para o português devem abarcar a diversidade de usos

European Language Resource Coordination
Connecting Europe Facility

Levantamento contrastivo de recursos linguísticos

Estudo contrastivo sobre o estado do desenvolvimento de recursos e ferramentas para o processamento das línguas naturais

- Levantamento exaustivo dos recursos linguísticos tendo em conta as variedades do português
- De acordo com determinadas categorias latas pré-estabelecidas
- Dando lugar a uma classificação do estado da arte para cada uma das línguas envolvidas

Branco et al. (2012) A língua portuguesa na era digital, Coleção Livros Brancos. Berlin, Heidelberg: Springer-Verlag.

European Language Resource Coordination
Connecting Europe Facility

Estado de desenvolvimento dos recursos linguísticos para o português

Recursos Linguísticos: Conjuntos de Dados e Bases de Conhecimento Linguístico	Quantidade	Disponibilidade	Qualidade	Cobertura	Maturidade	Sustentabilidade	Adaptabilidade
Corpora Escritos	3	3	4	4.5	4	4.5	4.5
Corpora de Fala	4	2	4	4	4	3	3
Corpora Paralelos	2	4	2	2	2	3	3
Recursos Lexicais	3.5	3	4.5	3	4	3	3
Gramáticas	1	4	5	2	2	2	2

8: Estado de desenvolvimento da tecnologia da linguagem para o português

Corpora escritos

- Representatividade vs. disponibilidade em *corpora* de grandes dimensões
  - CETEMPúblico (PT) (Linguateca)
  - Corpus de Referência do Português Contemporâneo (CLUL)
  - Corpus do Português (Brigham Young U./ Georgetown U.)
  - Banco do Português (BR) (PUC-SP)

- Recursos criados para apoiar a tecnologia da linguagem e/ou recursos desenvolvidos como fonte de dados para análise linguística
  - Razão pela qual existem conjuntos de dados importantes para o português em áreas diversas, que nem sempre estão associados a tecnologia da linguagem

Corpora de variedades geográficas

- *Corpora* comparáveis de variedades do português
  - Corpus África (CLUL)
  - CONDIVport (CEHUM)
  - VARPORT (UFRJ/CLUL)
  - VAPOR (CLUL)

- *Corpora* de variedades regionais de português europeu
  - CORDIAL-SIN (CLUL)
  - CPE-Var - Corpus de Português Europeu – Variação (CLUL)
  - Perfil Sociolinguístico da Fala Bracarense (CEHUM)

Corpora de aquisição e aprendizagem

- *Corpora* de aquisição do português L1
  - *Base de Dados de Aquisição do Português* (CLUL)
  - Corpus Freitas / Corpus Santos (CLUL)
  - Base de dados *LumaLIDa* (CLUL)

- *Corpora* de aprendizagem do português L2/LE
  - Corpus de Produções Escritas de Aprendentes de PL2 - PEAPL2 (CELGA-ILTEC)
  - Corpus de Aquisição de L2 - CAL2 (CLUNL)
  - Corpus de Português Língua Estrangeira/Língua Segunda - COPLE2 (CLUL)

Corpora anotados

- Anotação manual (ou manualmente revista)

  – Classe de palavras (*part-of-speech – POS*)

  + Flexão nominal e verbal
  + Lema — Corpora PAROLE / CORDIAL-SIN
  + entidades nomeadas — Corpus CINTIL (FCUL/CLUL)

  – Sintaxe — CINTIL-Treebank e Dependency bank (FCUL)
  Floresta Sintá(c)tica (Linguateca)
  CORDIAL-SIN
  PLN-Br (NILC)

Corpora anotados

- Anotação manual (ou manualmente revista)
  - Semântica e discursiva

forma lógica	CINTIL-LogicalFormBank (FCUL)
campos semânticos	CETEMPúblico
informação temporal	TimeBankPT (FCUL)
Informação modal	Corpus MODAL (CLUL)
relações discursivas	CSTNews Corpus (NILC)

- Necessário produzir treebanks de maiores dimensões e corpora com anotação semântica e discursiva mais alargada

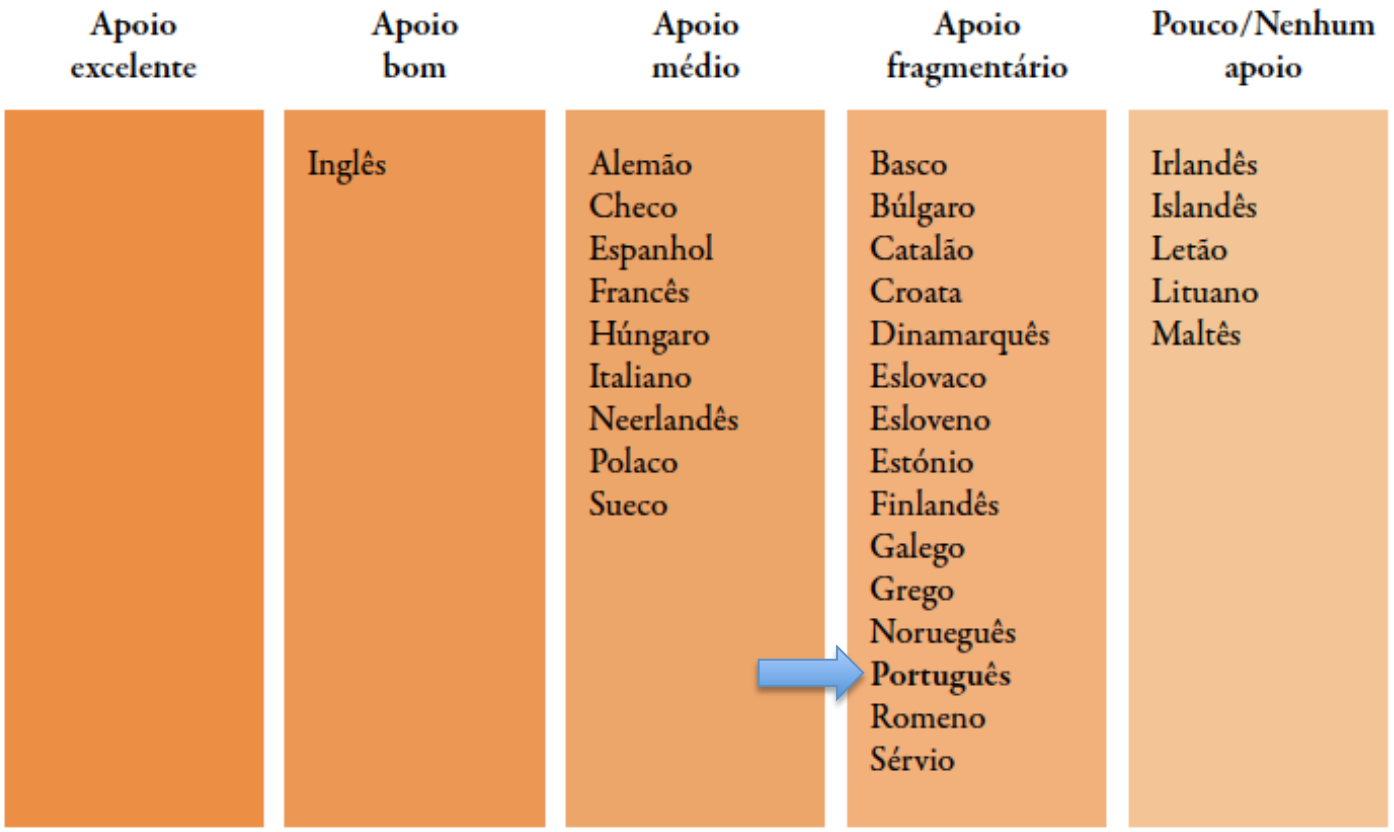
Recursos lexicais e gramáticas

- Léxicos e wordnets
  - Mordebe (CELGA-ILTEC)
  - Léxico Parole e SIMPLE (CLUL)
  - WordNet-PT (CLUL)
  - MultiWordNet (FCUL)

- Os recursos lexicais são ainda de dimensões reduzidas para aplicações computacionais, ou traduzidos a partir de recursos para o inglês.
- As gramáticas computacionais são escassas.

European Language Resource Coordination
Connecting Europe Facility

Estado de desenvolvimento dos recursos linguísticos para o português



12: Recursos linguísticos escritos e orais: estado da tecnologia da linguagem para 30 línguas europeias

European Language Resource Coordination
Connecting Europe Facility

Conclusão

- Existe um conjunto já importante de recursos para o português, embora nalgumas áreas estes continuem a ser escassos ou de dimensões reduzidas e longe de corresponder à dimensão internacional do português;
- Quanto mais conhecimento linguístico está envolvido, menos recursos estão disponíveis (sintaxe, discurso, texto); o suporte para níveis mais avançados de processamento da linguagem fica assim comprometido;
- A cobertura de recursos que exigem informação linguística profunda está dependente de esta ser considerada uma área prioritária por todas as entidades envolvidas.