Desenho, construção e utilização de corpora

(e sua aplicação em ensino e extracção de conhecimento)

Diana Santos

Primeira Escola de Verão da Linguateca

O que é um corpus?

- Uma colecção **classificada** de **objectos linguísticos** para **uso** em PLN/LC/L
- Uso: estudo, medição, teste, avaliação
- Objectos linguísticos: textos, frases, palavras, entrevistas, erros ortográficos, entradas de dicionário, citações, pareceres jurídicos, filmes, imagens com legendas, traduções, correcções, telefonemas, WOZ, programas ...

Classificada...

- A nível dos parâmetros da recolha (que categorias considerar)
- A nível da escolha (todos, alguns, amostra,...)
- A nível dos fenómenos (tipo de erro, tipo de tradução, tipo de texto, ...)
- A nível dos constituintes (análise sintáctica, semântica, fonológica, discursiva, etc.)
- Avaliação

Conceptualmente

- Um “mesmo” corpus pode ser considerado vários corpora diferentes, conforme as escolhas e as classificações
- Um corpus pode sempre ser estendido
  - ao nível do tamanho em termos de objectos linguísticos
  - ao nível da classificação

Exemplos de corpora prototípicos

- Colecção de textos para estudo da língua geral (escritos, *Trésor*, falados, *PF*)
- Colecção de fragmentos de textos escritos para criar um dicionário (Sinclair) ou para estudar géneros (Biber)
- Colecção de textos literários e suas traduções para estudar a tradução ou a semântica
- Colecção de interacções com sistema de reserva de viagens (ATIS, Jurafsky/Martin)

Exemplos de corpora prototípicos

- colecção de notícias de uma agência classificada por assuntos (Reuters)
- colecção de frases analisadas sintacticamente (florestas sintácticas)
- colecção de traduções e suas correcções
- dígitos e nomes próprios gravados
- histórias contadas a partir de imagens, Slobin
- interacção com a família

Exemplos de corpora prototípicos

- um conjunto de textos técnicos em duas línguas para ensinar terminologia e tradução
- erros e sua correcção possível
- perguntas e respostas em contexto
- lixo electrónico (“spam”)

Outros corpora

- o conteúdo de uma biblioteca
  - se pretendermos fazer alguma coisa com ele
- um instantâneo da Rede
  - idem (estudar a língua na Web, os assuntos, etc.)
- Ou seja, em alguns casos a compilação é externa, mas desde que o conteúdo possa ser visto como objectos linguísticos e o uso seja de PLN e os classifiquemos...

O objectivo da compilação

- Quase nunca o objectivo da compilação é levado em conta por futuros utilizadores
- E: podem usar-se (engenhosamente) recursos criados com um objectivo para outros objectivos muito diferentes (nem sonhados)

Atitudes em relação a um corpus

- Corpus como objecto de respeito
  - A criação é como uma obra de arte, é para ficar para a posteridade, deve ser constantemente corrigido, mantido, e estendido
  - Respeito pelos autores, e pelo trabalho de criação
- Corpus para fazer e deitar fora
  - O que interessa é o que se faz com o corpus
  - Tudo se transforma, para o ano cria-se um mais moderno, maior

Há espaço para todos

- Um “mau” corpus para um pode ser “bom” para outro
- Um “bom” corpus pode não servir para o que queremos

- É apenas preciso não vender gato por lebre, e não ir procurar debaixo do candeeiro

Estudos com base em corpora

- Em vez de ... ? Conhecimento inato, conhecimento adquirido, perguntar ao vizinho, ir ao dicionário, consultar um advogado, ...
- Metodologia
  - com base em, inspirado por, ou arrastado por?
  - questões estatísticas e quantitativas vs. qualitativas
- Consulta ou obediência?
- Qual a autoridade? Qual o objectivo?

Há 4 tipos de usos principais

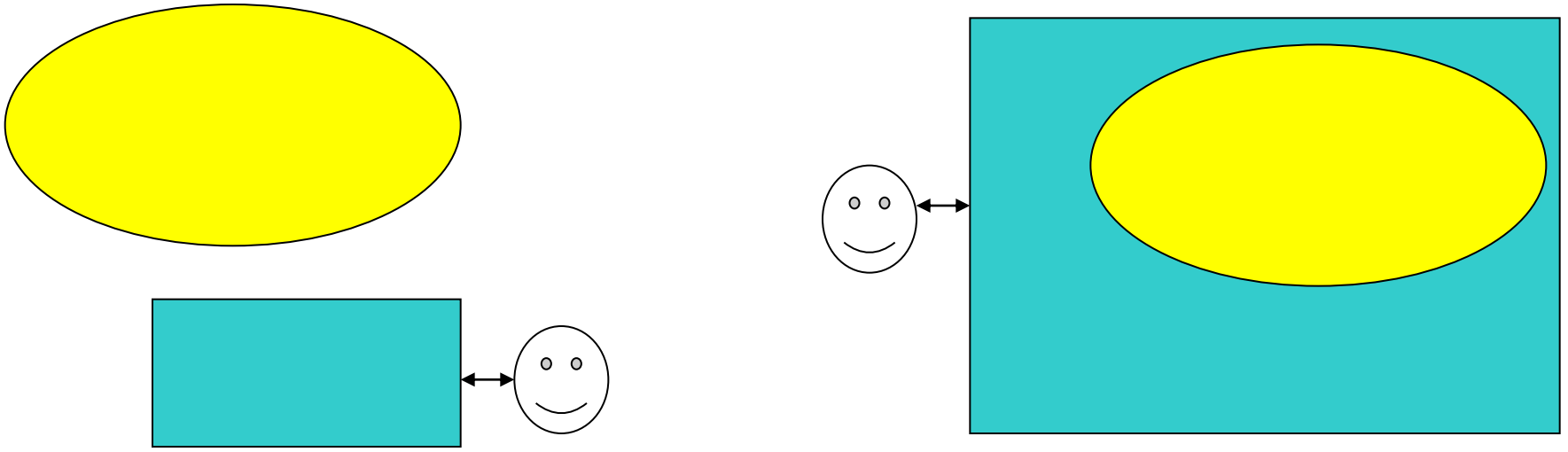
1. Ter uma ideia do problema/conhecer
   - consultor
   - familiarizador
   - treinador
2. Medir
3. Avaliar
   - uma hipótese
   - um sistema
   - um método

Recursos baseados em corpora

4. Criar outras coisas
   - dicionários ou estruturas de conhecimento
   - materiais de teste
   - sistemas de RAP
   - sistemas de ensino e jogos
   - terminologias, almanaques e catálogos
   - sistemas de detecção (de plágio, de spam, ...)

Aplicações baseadas em corpora

- vs. aplicações sobre corpora ou colecções
- "baseadas" no conhecimento extraído de corpora, e não "incluindo" os corpora



Medir: para quê? o quê?

- Questões de frequência

- Decidir o que fazer primeiro
- Pesar de maneira diferente fenómenos diferentes (implementação)
- Estimar um problema (risco)
- Planear a amostragem, o tamanho futuro
- Comparar através das medidas (identificar)

Que tipo de medidas (em corpora)

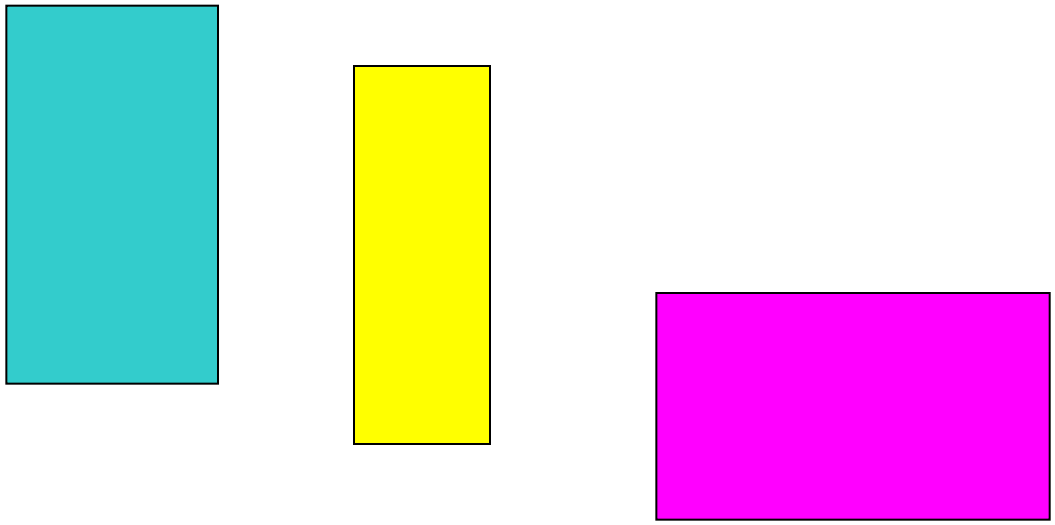
- Frequência absoluta
- Frequência relativa
- Ordem/posto (“ranking”)
- Distribuição
- Entropia
- Co-ocorrências
- Agrupamento (clusters)
- tfidf, LS

A rede (Web) como corpus

- Mais um slogan ☺
- Amostras da rede
- Fragmentos da rede
- A rede como
  - confirmadora
  - estimadora
  - consultora
- A rede como fonte de todo o conhecimento

Comparação de corpora

- vs. comparação de estudos baseados em corpora vs. contraste de corpora
- validação: quão generalizável é um estudo, ou uma avaliação?



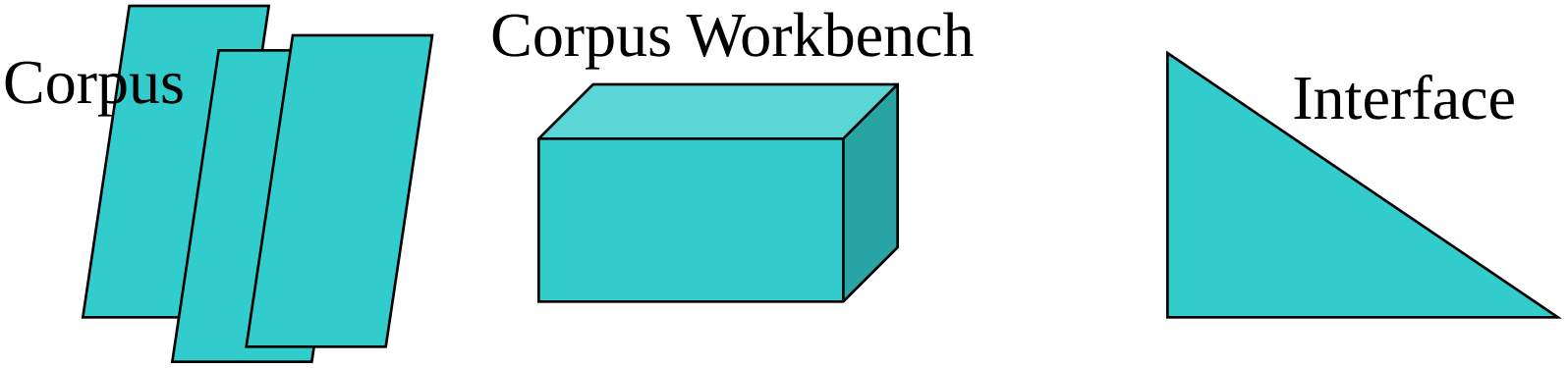
Exemplos de perguntas a um corpus

- Há correlação entre duas propriedades observáveis?
- Há correlação entre uma categoria de classificação e uma propriedade observável?
- Diferenças e semelhanças entre dois objectos

Manutenção de corpora

- A questão das versões
- A questão da correcção de erros
- A questão das funcionalidades da interface

- Um corpus em geral é um trio (Santos 1998)



Classificação: anotação e marcação

- É o que diferencia um corpus de um conjunto de textos
- A escolha corresponde a uma marcação implícita (segundo os critérios de selecção)
- Geralmente distingue-se
  - marcação (por partes): extralinguística
  - anotação (por palavras): linguística
- Questão de implementação (misturar ou separar)

Usar um corpus

- Avaliar a sua adaptação à questão
- Como fazer (perguntar) para sabermos o que queremos
- Como ajuizar os resultados
- Como melhorá-lo

- Processo **muito** iterativo...