Apresentação do Projecto

CORAL - Corpus de Diálogo Etiquetado

Consórcio:

- INESC
  - I. Trancoso
  - L. Oliveira

- CLUL
  - C. Viana
  - I. Mascarenhas

- FLUL
  - I. Duarte
  - G. Matos

- FCSH-UNL
  - H. Costa Campos
  - C. Correia

Duração do projecto: Janeiro de 1997 a Dezembro de 1998

Sumário

- Objectivos
- Especificação do corpus
- Gravação
- Anotação
- Trabalho futuro

Objectivos

Construção de um *corpus* de diálogo falado, com vários níveis de anotação:

- ortográfica
- fonética
- prosódica
- sintáctica
- semântica

O *corpus* será usado sobretudo em investigação em processamento de fala espontânea (caracterizada por toda uma série de fenómenos que dificultam sobremaneira a sua compreensão por parte de um computador - hesitações, recomeços, más construções, etc.) e também para o estudo do diálogo propriamente dito, em particular da sua estruturação e interligação com o reconhecimento de fala.

Especificação do *corpus*

Tema: descrição de trajectos em mapas.

Dimensionamento: 32 locutores distribuídos por 8 quartetos; 4 diálogos por locutor; 16 pares de mapas.

Principais aspectos a estudar: foram seleccionados 4 aspectos da fonética e fonologia, cujo estudo é prioritário no contexto da fala espontânea em P.E.:

- Velarização de /l/
- /s/ final de palavra seguido de fricativa coronal
- Sequências de oclusivas resultantes de queda de vogal
- Sequências de obstruintes resultantes (ou não) de queda de vogal

Especificação do *corpus* (cont.)

Cada mapa contem 6 elementos principais:

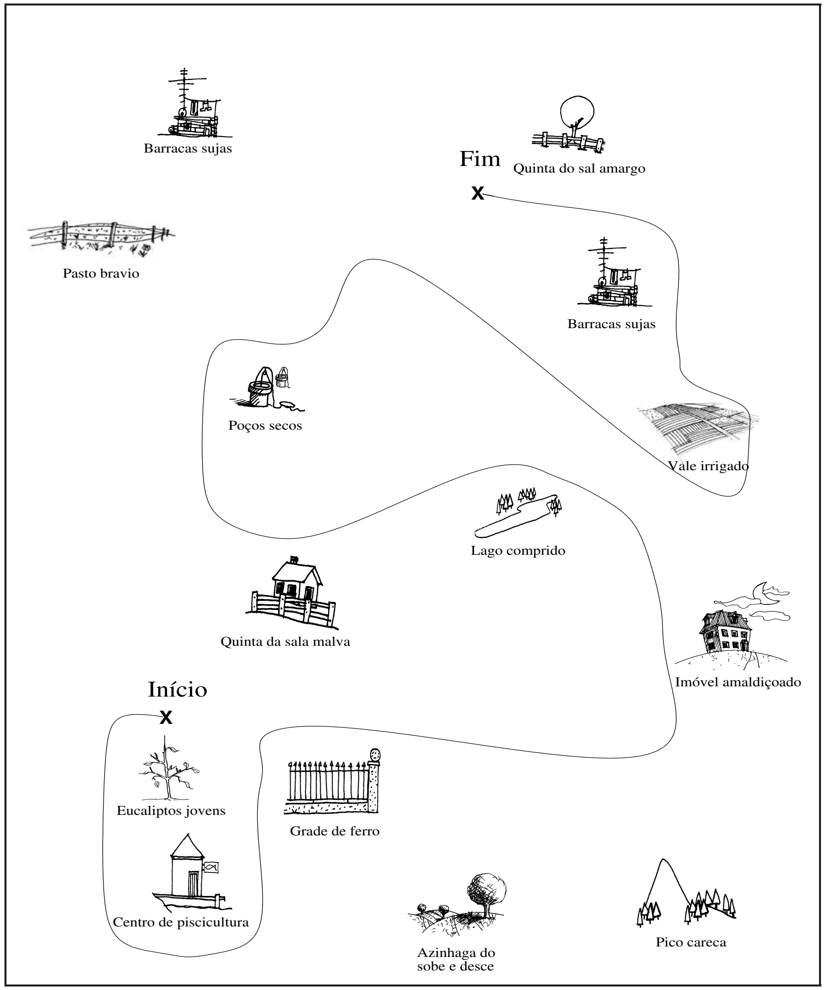
- Mestre (acordo e contraste)
- Duplicado
- Ausente/presente
- Nome modificado
- Comum
- Estranho

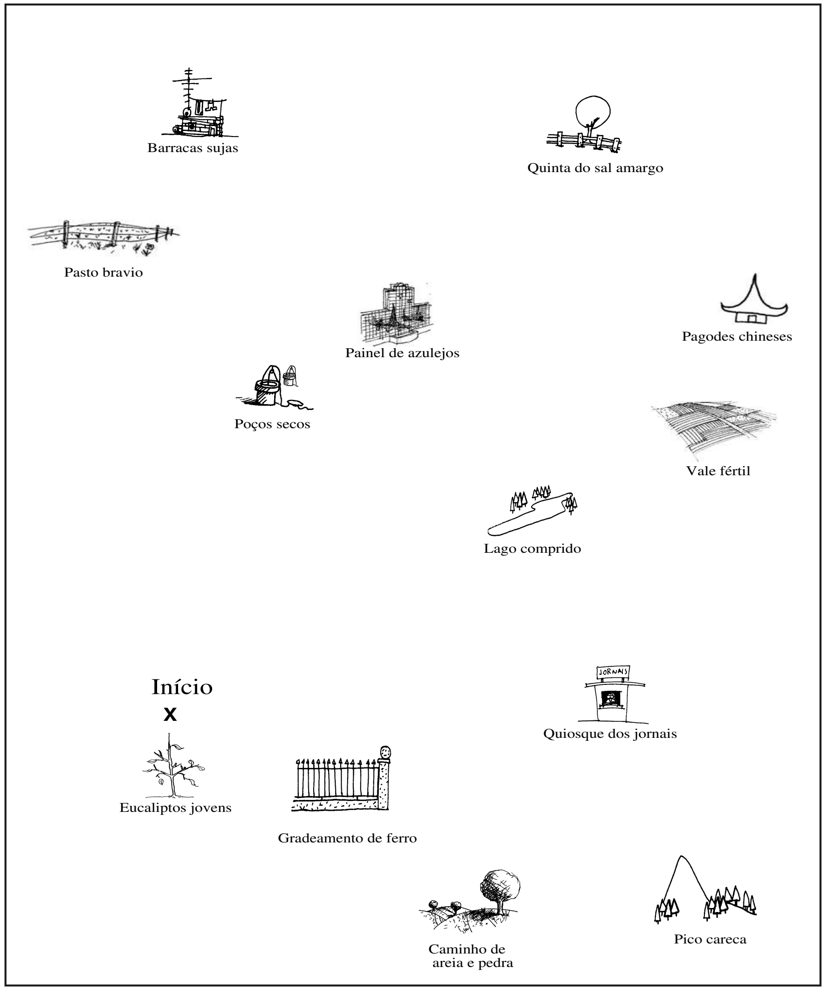
e um número variável (cerca de 10) de elementos secundários escolhidos de modo a complementar o estudo dos 4 aspectos acima mencionados, ou a variedade lexical ou ainda a prospecção de outros aspectos de ordem semântica ou sintáctica a analisar no contexto da fala espontânea.

Exemplo do mapa piloto

M1: quinta do sal amargo + quinta da sala malva / quinta do sal amargo
D2: barracas sujas
A1: - / painel de azulejos
N4: grade de ferro / gradeamento de ferro
C3: lago comprido
E2: - / pagodes chineses
S1: vale irrigado / vale fértil
S1: imóvel amaldiçoado / -
S2: centro de piscicultura / -
S2: eucaliptos jovens
S3: pasto bravio
S3: pico careca
S4: - / quiosque dos jornais
S4: poços secos
S5: azinhaga do sobe e desce / caminho de areia e pedra

Mapas do Dador e Seguidor





Gravação

- câmara insonorizada
- locutores separados por biombo sem contacto visual
- auscultador com microfone acopulado
- gravação em cassete DAT (um canal para cada locutor)
- digitalização a 16 kHz em stereo (amostras intercaladas dos 2 canais).

Anotação ortográfica

- Utilizou-se como ponto de partida a notação TEI/SGML, embora se tenham introduzido alterações significativas.
- Anotação das intervenções dos interlocutores por ordem cronológica em unidades independentes, localizáveis no sinal acústico.
- Marcação da sobreposição entre interlocutores
- Micro-anotações para dar conta de: pronúncias nâo esperadas, contracções, discontinuidades prosódicas, repetições com ou sem correcção de material lexical, pausas preenchidas, etc.
- Número reduzido de sinais de pontuação para demarcar fronteiras de unidades entoacionais (", . ? ! ...").

Exemplo de anotação ortográfica

<u who=G n=78>

<sfo samp=2692447>

Sim. O pasto bravio fica-te à tua esquerda,{pp} [e os]

<u who=F n=79>

<sfo samp=2732665>

[sim].

<u who=G n=80>

<sfo samp=2738908>

{ph|p”OSu=poços} secos à tua direita.

Anotação fonética

- Vários níveis produzidos semi-automaticamente (conversão grafema-fone + alinhamento):
  - forma de citação
  - transcrição fonética larga ou fonotípica
- Correcção manual da transcrição fonética estreita com base na observação do sinal acústico.
- Utilização das micro-anotações para assegurar um alinhamento aceitável.
- As anotações fonéticas, prosódicas, sintácticas e semânticas são apenas produzida para um subconjunto do corpus.

Anotação prosódica

- Seguem-se basicamente as propostas do sistema ToBI (de *Tone and Break Indices*), representando diferentes aspectos da prosódia em fiadas paralelas alinhadas com a representação ortográfica e o sinal acústico.
- Alterações no tratamento das disfluências em que se seguem de perto as propostas de [Shriberg:94], passando a ser consideradas num nível independente, reservando-se o nível miscelâneo para a anotação de outros fenómenos de carácter para- ou extra-linguístico.
- Uso sistemático de um conjunto de diacríticos que asseguram um alinhamento mais preciso dos tons com o material segmental e o sinal acústico.

Anotação sintáctica

- Domínio frásico:
  - Nível da palavra: atribuição de etiquetas categoriais a cada um dos itens presentes no *corpus* (e.g., N, V, A,...);
  - Nível sintagmático: identificação dos constituintes principais de cada unidade e correspondente etiquetagem categorial (e.g., NP, VP, PP, ...);
  - Nível da predicação: identificação dos constituintes com a relação gramatical de sujeito e de predicado e correspondente etiquetagem relacional (SUJ, PRED);
- Domínio transfrásico (frases complexas, pares pergunta-resposta e tomadas de palavra que continuam o discurso anterior)
  - Nível de identificação de elipses: identificação da unidade que contém o antecedente de um elemento elíptico e caracterização do tipo de elipse (E; VPE, GAP, STRIP...);

Anotação sintáctica (cont.)

- Discurso:
  - Nível informacional: identificação da estrutura informacional de cada enunciado e correspondente etiquetagem (GIV; INF FOC, Q FOC,...);
  - Nível temático: identificação da estrutura temática de cada enunciado e correspondente etiquetagem categorial (TOP/COM);
- Reformulações e disfuncionalidades:
  - Identificação do ponto inicial de reformulações e rupturas. sintácticas.

Nível de anotação inteiramente manual, tal como o semântico.

Anotação semântica

- Representação semântica de ocorrências nominais. A análise dos sintagmas nominais terá de abranger todos os constituintes que convergem para a construção da ocorrência.

- A análise da ocorrência de uma noção não se restringe a uma o tomada de palavra mas abrange uma sucessão de tomadas de palavra constituindo uma unidade discursiva.

- A representação semântica etiquetada é uma proposta que irá sendo sucessivamente adequada às necessidades e limites da computação e à correspondente projecção na representação sintáctica.

Exemplo de anotação semântica

Dador: $<<$DX1 MAPA $\supseteq$ Op DDEF / DX1[O início] $\supseteq$ Op [é junto a] DINDEF / EXIST [uns eucaliptos jovens] $>$

- Princípio de unidade discursiva
- Construção da primeira ocorrência no mapa do dador: predicação de existência
- Artigo indefinido, plural, N com funcionamento discreto
- Localizador: o início
- Operador de localização: é junto a

Seguidor: $<$ADV / ANAFFR [sim também] DX2 $[pro]_1$ $\supseteq$ Op $[\text{tenho}]_1$ DØ [eucaliptos jovens] $>$

- Construção da primeira ocorrência no mapa do seguidor: predicação de existência
- Determinante Ø, plural, N com funcionamento discreto
- Localizador: [pro] (DX2: referência deíctica, localizador = sujeito enunciador)
- Operador de localização: tenho
- também - anáfora proposicional, confirma a anterioridade da existência de N.

Dador: $<$Pronto$>>_{(FIM)}$

Trabalho futuro

Mapeamento entre os diversos níveis de anotação:

As potencialidades de exploração de um *corpus* deste tipo são inúmeras. Começámos já a tentar estabelecer correlações entre o *bracketing* sintáctico e *phrasing* prosódico, por um lado, e o *bracketing* sintáctico e semântico, por outro.