Apresentação do Projecto
CORAL - *Corpus* de Diálogo Etiquetado

M. Céu Viana*, Isabel Trancoso†, Isabel Mascarenhas*, Inês Duarte‡, Gabriela Matos‡, Luís C. Oliveira†, Henriqueta C. Campos✧, Clara Correia✧

† INESC/IST, * CLUL, ‡ FLUL, ✧ FCSH-UNL

1. Introdução

O Projecto CORAL surge na sequência de um conjunto de trabalhos conducentes à recolha de materiais de fala para o Português Europeu em condições controladas, capazes de servir de suporte à investigação e desenvolvimento na área das tecnologias da fala.(cf. Martins et al., 1998). O seu objectivo é construção de um *corpus* de diálogo com vocabulário limitado, suficientemente variado em termos de número de falantes e com vários níveis de anotação ortográfica, fonética, prosódica, sintáctica e semântica.

À semelhança de trabalhos congéneres já realizados ou em curso para outras línguas, todos os diálogos incidem sobre um mesmo tema e têm por objectivo a resolução de uma tarefa específica. Estas restrições têm-se revelado eficazes no que diz respeito ao controle das dimensões do vocabulário e à obtenção de produções adequadas para a investigação em processamento de fala espontânea, nomeadamente no que se refere ao tratamento das disfluências, à modelização prosódica, ao desenho de interfaces e à estruturação dos diálogos propriamente ditos, em interligação com o reconhecimento de fala.

O estudo destas questões não cabe contudo no âmbito deste projecto que apenas visa a criação de uma infra-estrutura necessária para esse efeito. Pretende-se, assim, fundamentalmente, assegurar a transliteração completa de todos os materiais recolhidos, indicando cuidadosamente as disfluências e outros fenómenos para- ou extra-linguísticos, bem como a sua localização no sinal acústico. Embora seja também essencial a anotação muiltinível de todo o *corpus*, esta restringe-se apenas a um subconjunto relativamente pequeno, uma vez que, pelo menos de momento, a maior parte do trabalho é manual, exigindo recursos humanos muito para além dos disponíveis.

O primeiro ano de trabalho foi preenchido pela especificação do *corpus* e pela gravação e tratamento de um diálogo de teste, para uma melhor definição dos níveis e critérios de segmentação e etiquetagem a contemplar. Seguiram-se a recolha dos diálogos e, quase em paralelo, as várias tarefas de anotação.

2. Especificação do *Corpus*

Na selecção do tema dos diálogos foram tidos em consideração projectos congéneres para outras línguas, tendo a escolha recaído sobre a reconstituição de percursos em mapas, um tema que tem sido utilizado por várias equipas de investigação na Europa[1], na América[2] e no Japão. Entre os vários temas possíveis, este pareceu-nos o mais capaz de originar diálogos com alguma vivacidade e, simultaneamente, permitir a elicitação de um conjunto controlado de sequências sonoras em fronteira de palavra para o estudo de alguns aspectos de ordem fonético-fonológica.

De modo a possibilitar posteriores comparações, procurou-se seguir as principais linhas de orientação definidas no *MAP corpus* original recolhido pelo HCRC (*Human Computer Research Center*) da Universidade de Edimburgo. Houve, no entanto, que restringir o número de diálogos gravados para cerca de metade, dados os drásticos cortes relativamente à proposta inicialmente submetida.

O diálogo passa-se entre dois locutores que têm mapas semelhantes. O locutor que tem um trajecto desenhado entre os vários elementos constituintes do mapa actua como dador de informação e deverá dialogar com o seu interlocutor de modo a que este (o seguidor) consiga reconstituir o mesmo trajecto. O *corpus* é falado por 32 locutores, distribuídos por 8 quartetos, existindo apenas 16 pares de mapas, distribuídos por 4 quartetos. Os primeiros 16 falantes usam, por conseguinte, a totalidade dos mapas e os últimos 16 falantes tornam a usá-los. Cada falante actua duas vezes como dador (usando o mesmo mapa) e duas vezes como seguidor (usando mapas diferentes). Metade dos diálogos passam-se entre falantes conhecidos e metade entre falantes desconhecidos à partida. Os quartetos serão formados por 2 falantes do sexo masculino e 2 do sexo feminino.

---

[1] http://www.cogsci.ed.ac.uk/hcrc/wgs/dialogue/dialog/maptask.html

[2] http://www.ldc.upenn.edu/readme_files/dciem.readme.html

Cada mapa contém 6 elementos, ditos *principais*, em torno dos quais se organiza o trajecto e é fomentado o diálogo. Dado que a quantidade de elementos por mapa é relativamente pequena, o número de fenómenos de ordem fonético-fonológica cujas ocorrências podem ser objecto de algum controle é também limitado (4 no *MAP corpus* original). Para o Português Europeu, por razões que se prendem fundamentalmente com o desenvolvimento do sistema de síntese, são elicitadas sequências sonoras para o estudo (1) da velarização de /l/, (2) de /s/ seguido de fricativa coronal, (3) de sequências de oclusivas resultantes de queda de vogal e (4) de sequências de obstruintes resultantes ou não de queda de vogal. Cada par de mapas inclui pelo menos um elemento principal para o estudo de cada um destes casos.

Existem apenas 4 trajectos a que correspondem quatro pares de elementos ditos *mestre* (M) ligados a cada um dos fenómenos a analisar. Estes encontram-se sempre na mesma localização e repetem-se em cada quarteto, embora com ligeiras variações: o mapa do dador pode conter os dois elementos do par ou apenas um (*+/contraste*) e pode ou não haver concordância entre o seu mapa e o do seguidor (*+/- acordo)*.

De modo a provocar o diálogo, há ainda outros elementos principais que não coincidem exactamente nos dois mapas: um elemento *duplicado* (D) no mapa do dador que apenas aparece uma vez no do seguidor, um elemento a*usente/presente* (A) que, como o nome indica, está presente num dos mapas e ausente no outro e um elemento que, embora tenha o mesmo desenho e localização nos dois mapas, tem o *nome modificado* (N). Existem ainda dois outros tipos de elementos por par de mapas: um elemento *comum* (C) tanto em termos de desenho como nome e localização e um elemento *estranho* (E), isto é, fora do contexto em relação ao "cenário" projectado para cada mapa, que também pode estar ausente num dos mapas.

Para além dos 6 elementos principais, cada par de mapas contém ainda um número variável de elementos *secundários* (cerca de 10). Estes elementos foram seleccionados para complementar o estudo dos fenómenos acima mencionados (S1 a S4) ou para a prospecção de outros aspectos de ordem sintáctica e semântica (S5). Para não prejudicar a espontaneidade da interacção entre dador e seguidor, os elementos S5 contemplam apenas dois tipos de casos: (1) expressões nominais envolvendo diferentes colocações do adjectivo e constituindo pares

mínimos (ex: *estrada antiga das minas* vs *antiga estrada das minas*; *estrada das minas antigas* vs *estrada das antigas minas*), que poderão fornecer dados interessantes sobre a correlação entre fraseamento sintáctico e prosódico; (2) pares de sinónimos (ex: *fraga* vs *penha*) e parónimos (*tonel* vs *túnel*) que, para além de permitirem avaliar o grau de precisão vocabular dos intervenientes também podem contribuir para fomentar o diálogo.

Cada trajecto é desenhado de modo a começar e acabar num elemento comum a ambos os mapas. Os elementos intermédios podem ser comuns ou diferir em alguns dos aspectos acima mencionados, havendo pelo menos dois elementos que só aparecem no mapa do dador e dois no do seguidor.

A lista seguinte, corresponde ao par de mapas na figura1, os utilizados no diálogo 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

O elemento mestre destina-se ao estudo da velarização de /l/, com contraste e sem acordo entre dador e seguidor. As diferenças entre os mapas são assinaladas por uma barra inclinada "/" e a inexistência de um elemento por um traço "-".

3. Condições de gravação

Os diálogos são gravados em câmara insonorizada. Os falantes encontram-se sentados em mesas independentes e estão separados por um biombo que impede o contacto visual e reduz a incidência directa do som no outro microfone. Cada falante dispõe de um auscultador com um microfone acoplado que está ligado a um canal independente da mesa de mistura e, através desta, a um gravador digital DAT. O armazenamento é feito em

dois canais, em cassete de fita magnética de 4 mm, a uma frequência de 48 KHZ e a digitalização da onda sonora a 16 kHz em stereo (amostras intercaladas dos 2 canais).

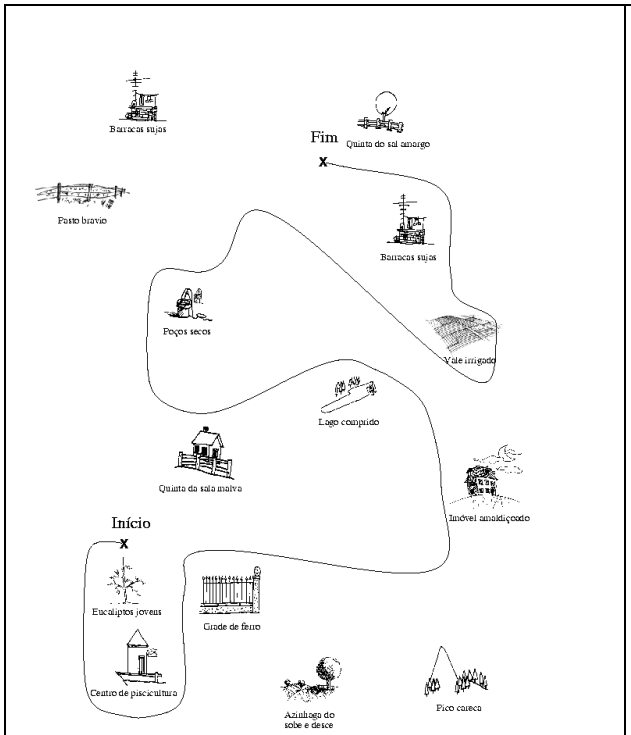
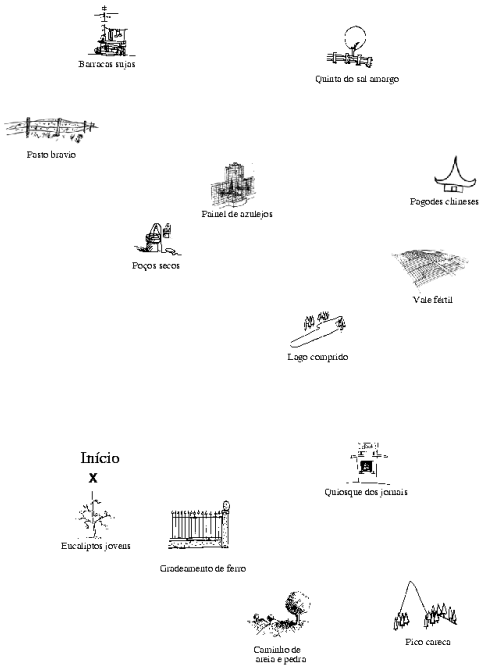


Fig. 1 – (a) Mapa do dador



(b) Mapa do seguidor

4. Anotação

No tratamento dos materiais de fala recolhidos (ou a recolher) no âmbito deste projecto, são contemplados diferentes níveis de representação, alinhados entre si e com o sinal acústico.

A representação ortográfica, assegurada para a totalidade do *corpus*, tem por objectivo permitir um acesso fácil da generalidade dos utilizadores aos conteúdos dos ficheiros de sinal e ainda facilitar o processamento automático ou semi-automático posterior.

A cada intervenção de um locutor corresponde uma unidade iniciada por duas linhas em formato SGML, delimitadas por parêntesis angulares. A primeira permite identificar o falante e o número da unidade e a segunda o número da amostra no ficheiro de sinal em que esta se inicia. Para indicar se a uma mudança de unidade corresponde ou não uma tomada de palavra, esta inicia-se por maiúscula ou minúscula, respectivamente. As sobreposições são marcadas entre parêntesis rectos e um conjunto de micro-anotações (entre chavetas e facilmente removíveis) permite dar conta de determinado tipo de fenómenos (ex: descontinuidades prosódicas, repetições com ou sem correcção de material lexical,, pausas preenchidas, contracções, etc). Neste exemplo, trata-se de uma anotação fonética (código *ph*), indicando uma pronúncia inesperada em alfabeto SAMPA[3]. Os sinais de pontuação procuram reflectir apenas a continuidade do fluxo discursivo (","), terminalidade (".") e apelo à intervenção do interlocutor ("?").

ex:
<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.

Para cada ficheiro de fala são produzidos automática ou semi-automaticamente dois níveis de transcrição fonética gerados com o módulo de conversão grafema-fone do

[3] http://www..phon.ucl.ac.uk/home/sampa/home.html

sintetizador DIXI {Oliveira et al., 1992): forma canónica e transcrição fonotípica.

Com base no reconhecimento de segmentos fonéticos com modelos de Markov não observáveis (HMM - Hidden Markov Models) e utilizando o pacote de programas HTK (HMM Toolkit), desenvolvido pela Universidade de Cambridge (UK) e comercializado pela ENTROPIC, procede-se então ao alinhamento das transcrições fonotípicas com o sinal de fala. É a partir deste alinhamento que é gerada uma transcrição fonética estreita que descreve com maior aproximação o que foi efectivamente dito. Para esse efeito, é obrigatório proceder à correcção manual da segmentação e etiquetagem produzidas pelo reconhecedor, o que envolve a observação cuidada do sinal acústico.

Dada a extrema morosidade das correcções manuais e o facto de estas exigirem o recurso a anotadores especializados, este e os restantes níveis de anotação apenas são contemplados para um pequeno subconjunto dos materiais do *corpus*

Na transcrição prosódica seguem-se basicamente as propostas do sistema ToBI (*de Tone and Break Indices*) {cf. Silverman et al., 1992; Beckman et al., 1994[4]; Pitrelli et al. 1994) e as de Shriberg (1994) no que diz respeito ao tratamento independente das disfluências. São assim contemplados os seguintes níveis de anotação: (1) *ortográfico* (com segmentação palavra a palavra), (2) *tonal* (3) *índices de ruptura,* (4) *disfluências* e (5) *miscelâneo*, restringindo-se este último à ocorrência de fenómenos de carácter para- extra-linguístico.

As anotações de carácter sintáctico-semântico são também muiti-lineares, contemplando tanto o domínio frásico como o transfrásico e atribuindo etiquetas aos seguintes níveis: (1) *palavra* (ex: N,V, A...); (2) *sintagmático* (ex: NP, VP, PP...); (3) *predicação* (ex: SUJ, PRED); (4) *identificação de elipses*, sendo identificada a unidade que contém o antecedente de um elemento elíptico e caracterizado o tipo de elipse. No que diz respeito ao discurso, é considerado (5) um nível *informacional* e (6) um nível *temático* (TOP - Tópico marcado; COM-comentário). Finalmente, é contemplado um nível (7) de *reformulações e descontinuidades*, em que é identificado o ponto inicial das reformulações e rupturas sintácticas devidas quer a dificuldades do

[4] http://ling.ohio-state.edu/Phonetics/E_ToBI/etobi_homepage.html

planeamento on-line quer a erros imputáveis a violações de propriedades de regência dos itens lexicais.

4. Conclusões e trabalho futuro

O artigo descreve um *corpus* de diálogo cujo objectivo é o estudo de vários fenómenos típicos da fala espontânea num domínio restrito. De particular interesse para este objectivo é a anotação multilinear do *corpus* que se procurará prosseguir quer ainda no âmbito deste projecto quer de outros que se venham a propor futuramente.

As potencialidades de exploração de um *corpus* deste tipo são inúmeras, tanto de um ponto de vista da investigação fundamental como aplicada. Embora a correlação entre as parentetizações sintáctica e prosódica seja, de momento, objecto de especial atenção, procurar-se-á alargar progressivamente o estudo a outras áreas, nomeadamente ao mapeamento fonética/fonologia. É de salientar, também, a possibilidade de utilização de *corpora* deste tipo para o estudo de esquemas de codificação da estrutura dos diálogos. De facto, o *MAP corpus* original tem sido usado recentemente para o estudo de esquemas de codificação a três níveis (movimentos, jogos e transacções), com particular ênfase na replicabilidade deste tipo de codificação subjectiva (Carletta, 1997). Na vanguarda desta área, estão os estudos que procuram tirar partido destes métodos de codificação para derivar predições estatísticas sobre o tipo do próximo movimento que é esperado do utilizador em sistemas de diálogo falado, usando reconhecimento e síntese de fala.

Referências

Carletta, J., A. Isard, S. Isard, J. Kowtko, G. Doherty-Sneddonn e A. Anderson (1997) - "The Reliability of a Dialogue Structure Coding Scheme", *Computational Linguistics*, 23.

Martins, Ciro, I. Mascarenhas, H. Meinedo, J. Neto, L. Oliveira, C. Ribeiro, I. Trancoso e C. Viana (1998) - "Spoken Language *Corpora* for Speech Recognition and Synthesis in European Portuguese", *Proc. RECPAD'98 - 10th Portuguese Conference on Pattern Recognition*. Lisboa.

Oliveira, L., C. Viana e I. Trancoso (1993) –"DIXI: sistema de síntese de fala a partir de texto para o Português", *Proc. EPLP'93*, Lisboa.

Pitrelli, J., M. Beckman e J. Hirschberg (1994) - “Evaluation of prosodic transcription labeling reliability in the ToBI framework", *Proc. ICSLP94*, Yokohama, Japão.

Shriberg, E. (1994) - “Preliminaries to a theory of speech disfluencies", PhD. Diss., Univ. California-Berkeley.

Silverman, K., M. Beckman, J. Pitrelli, M. Ostendorf, C. Whightman, P. Price, J. Pierrehumbert e J. Hirschberg (1992) - “ToBI: a standard for labeling English prosody", *Proc. ICSLP'92*, Banf, Alberta.