Corpus de Diálogo CORAL

Isabel Trancoso†, Maria do Céu Viana‡, Inês Duarte*, Gabriela Matos*

†INESC / IST ‡CLUL * FLUL

INESC, R. Alves Redol, 9, 1000 Lisboa, Portugal
E-mail: Isabel.Trancoso@inesc.pt
Tel.: +351 1 3100 268 Fax: +351 1 3145843

Resumo

Este artigo descreve as várias etapas da criação de um *corpus* de diálogo falado anotado a vários níveis: ortográfico, fonético-fonológico, e sintáctico-semântico.

Tal como em vários centros de investigação congéneres, o tema escolhido para o *corpus* foi a indicação de percursos em mapas. Foram desenhados 16 mapas distribuídos por 16 pares de locutores que, no seu total, participam em 32 diálogos. A tarefa de especificação dos elementos do mapa teve em conta, fundamentalmente, aspectos da fonética e fonologia cujo estudo é prioritário no contexto da fala espontânea em Português Europeu.

A gravação foi feita numa câmara insonorizada, directamente para cassete DAT. Cada diálogo foi posteriormente transcrito ortograficamente. Este nível de anotação inclui não só a transliteração em ortografia corrente do que foi dito, mas também um conjunto de anotações bastante elementares que permitem identificar diferentes tipos de unidades e localizar quer porções de sinal correspondentes a fala fluente quer porções onde ocorrem diferentes fenómenos típicos da situação de fala espontânea que têm de ser objecto de especial cuidado em fases posteriores de tratamento do *corpus*.

Os níveis de anotação seguintes são apenas efectuados sobre um subconjunto do *corpus*, dado serem, de momento, integralmente manuais ou, apesar da existência de ferramentas automáticas, necessitarem de correcções manuais (caso da anotação fonética).

As potencialidades de exploração de um *corpus* deste tipo são inúmeras, salientando-se, em particular, o mapeamento entre vários níveis de anotação (e.g. prosódia-sintaxe).

1 Introdução

O objectivo do projecto CORAL é a construção de um *corpus* de diálogo falado, com vários níveis de anotação: ortográfica, fonética, prosódica, sintáctica e semântica. Pretende-se constituir um *corpus* suficientemente representativo em termos de número de falantes, sobre um único tema escolhido de modo a limitar à partida o vocabulário usado. Este tipo de *corpus* é essencial para a 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, etc.. É também essencial para o estudo do diálogo propriamente dito, em particular da sua estruturação e interligação com o reconhecimento de fala.

O projecto não visa para já o estudo destes problemas, mas sim a criação de uma infraestrutura linguística que possibilite esse estudo em projectos a definir posteriormente por equipas interdisciplinares. É, portanto, essencial que, para além de incluir a transliteração do *corpus* completo, com a indicação de todos os fenómenos para-linguísticos, inclua também anotação a outros níveis - fonético, prosódico, sintáctico e semântico. Apesar da existência de algumas ferramentas automáticas para certos tipos de anotação, a sua fiabilidade com fala espontânea é bastante reduzida relativamente a fala lida, pelo que a maior parte deste trabalho é manual, exigindo recursos humanos fora do âmbito do projecto. Por este motivo, só um subconjunto

relativamente pequeno do *corpus* é etiquetado a todos os níveis.

A anotação multinível do *corpus* pressupõe a conjunção de esforços por parte de equipas interdisciplinares. Assim, o consórcio reune equipas do INESC (Instituto de Engenharia de Sistemas e Computadores), CLUL (Centro de Linguística da Universidade de Lisboa), FLUL (Faculdade de Letras da Universidade de Lisboa) e FCSH-UNL (Faculdade de Ciências Sociais e Humanas da Universidade Nova de Lisboa). Aliás, este *corpus* é o mais recente de um conjunto de *corpora* construídos no âmbito do convénio INESC-CLUL [3] e, embora preencha uma lacuna grande em termos dos recursos linguísticos que julgamos necessários para o estudo do Português Europeu, nomeadamente no que diz respeito a *corpora* de fala espontânea sobre um tema restrito, não esgota de modo algum estas necessidades.

O primeiro ano do projecto, recentemente completado, foi preenchido pela especificação do *corpus* e pela gravação e tratamento de um diálogo de teste. Seguem-se a recolha dos diálogos e, quase em paralelo, as várias tarefas de anotação. O projecto terminará em Dezembro de 1998 com o empacotamento do *corpus* em CD-ROM e um estudo preliminar do mapeamento entre os vários tipos de transcrição.

Este artigo descreve resumidamente o trabalho efectuado durante o primeiro ano do projecto, estando por conseguinte estruturado em três secções: a primeira dedicada à especificação do *corpus* e as seguintes à sua recolha e anotação.

2 Especificação do corpus

A primeira decisão a tomar neste contexto disse respeito ao tema (ou domínio) do *corpus*. Foram analisados vários domínios possíveis, tendo em conta, nomeadamente, projectos congéneres para outras línguas. A decisão recaiu sobre mapas, um tema de diálogo utilizado por várias equipas de investigação na Europa[1], América[2] e Japão. O diálogo passa-se entre dois locutores que têm mapas semelhantes entre si. 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.

Procurou-se seguir as mesmas linhas de orientação definidas no *MAP corpus* original recolhido pelo HCRC (Human Computer Research Center, Universidade de Edimburgo), de modo a possibilitar posteriores comparações. Houve, no entanto, que restringir o número de diálogos gravados, dados os drásticos cortes orçamentais do projecto, relativamente à proposta submetida.

O *corpus* é falado por 32 locutores, distribuídos por 8 quartetos. Existem apenas 16 pares de mapas que são distribuídos igualmente em 4 quartetos. Por conseguinte, os primeiros 16 falantes usam 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.

Tal como no *MAP corpus* original, 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

Cada par de mapas inclui pelo menos um elemento para o estudo de cada um desses aspectos. Esses elementos, que designaremos por ”principais”, são na realidade 6 por mapa. De modo a provocar o diálogo em torno destes 6 elementos principais, estes são organizados em 6 tipos diferentes, respectivamente:

- Mestre (M) - Cada mapa inclui um ou dois elementos de um par de elementos ”mestres”. A designação deve-se ao facto de existirem apenas 4 pares de elementos mestres, cada um dos quais com um trajecto correspondente e sempre na mesma

[1] http://www.cogsci.ed.ac.uk/hcrc/wgs/dialogue /dialog/maptask.html
[2] http://www.ldc.upenn.edu/readme_files/dciem .readme.html

localização. Cada par de elementos mestres é partilhado, por conseguinte, por 4 pares de mapas.

- Duplicado (D) - Elemento que existe duas vezes no mapa do dador (numa das vezes é relevante para o trajecto e na outra irrelevante) e apenas uma vez no mapa do seguidor (a irrelevante). Os elementos deste tipo não são repetidos em pares de mapas diferentes.

- Ausente/presente (A) - Elemento que está presente num dos mapas do par e ausente no outro. Os elementos deste tipo não são repetidos em pares de mapas diferentes.

- Nome modificado (N) - Elemento que, embora tendo o mesmo desenho e localização em ambos os mapas do par, apresenta um nome diferente (por exemplo: capela / ermida). Os elementos deste tipo não são repetidos em pares de mapas diferentes.

- Comum (C) - Elemento comum no par de mapas do dador e do seguidor, tanto em termos de desenho, como nome e localização. Os elementos deste tipo não são repetidos em pares de mapas diferentes.

- Estranho (E) - Elemento que parece fora do contexto em relação ao ”cenário” projectado para cada mapa (por exemplo, um ”ministério das finanças” inserido num cenário rural). Os elementos deste tipo não são repetidos em pares de mapas diferentes.

O total de elementos principais diferentes é de 88: 4 pares de elementos mestre e 16 conjuntos de 5 elementos (duplicado, ausente/presente, nome modificado, comum e estranho). Isto dá um total de 22 elementos diferentes para o estudo de cada fenómeno seleccionado.

Para além dos 6 elementos principais, existem em cada par de mapas elementos secundários (S), que podem ser em número variável (cerca de 10). O critério para a escolha destes elementos foi variado: a complementação do estudo das 4 aspectos acima mencionados (elementos S1 a S4), ou a variedade lexical e a prospecção de outros aspectos de ordem semântica ou sintáctica a analisar no contexto da fala espontânea (elementos S5).

Tratando-se de um *corpus* de diálogo, adoptou-se uma posição muito prudente quanto à elicitação de dados com objectivos estritamente sintácticos, uma vez que uma grande sobrecarga de texto a ler nos mapas prejudicaria necessariamente a espontaneidade da interacção entre dador e seguidor. Assim, limitou-se tal elicitação a dois casos:

- Expressões nominais envolvendo diferentes colocações do adjectivo, 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 *bracketing* sintáctico e *phrasing* prosódico.

- Pares de sinónimos (ex: *fraga* vs *penha* e parónimos (ex: *tonel* vs *túnel*) na designação de elementos constantes dos mapas que, para além de permitirem avaliar o grau de precisão vocabular dos intervenientes, poderão gerar situações de impasse geradoras de diálogo.

O par de elementos mestre de cada característica aparece de forma diferente nos mapas do dador e do seguidor de modo a fomentar o diálogo. A forma como diferem pode ser de dois tipos:

- Contraste - se o mapa do dador contiver o par de elementos mestre contrastante, será marcado como ”+contraste”; se o mapa do dador tiver apenas um dos elementos do par, será marcado como -contraste”.

- Acordo - se o mapa do seguidor estiver de acordo com o do dador em termos de contraste (se o dador tiver o par, o seguidor também o terá; se o dador tiver apenas um, o seguidor também só tem um), será marcado como ”+acordo”; se o mapa do seguidor estiver em desacordo com o do dador em termos de contraste (se o dador tiver o par, o seguidor tem apenas um e viceversa), será marcado como -acordo”.

Cada trajecto é desenhado de modo a começar e acabar num elemento comum a ambos os mapas. Os elementos intermédios ao longo do trajecto alternam entre elementos comuns e elementos que diferem de algum modo (acima descrito)

e há pelo menos dois elementos que só aparecem no mapa do dador e dois elementos que só aparecem no mapa do seguidor.

No total, existem cerca de 270 nomes de elementos diferentes nos 16 pares de mapas. Dado que o mesmo desenho foi utilizado várias vezes com nomes diferentes, o total de figuras distintas foi de cerca de 180.

Apresentamos de seguida uma lista de elementos possível - neste caso, a utilizada no par de mapas do diálogo piloto. Para além dos 6 elementos principais, foram escolhidos 9 elementos secundários. Sempre que há diferenças de nomes nos dois mapas do par, isso é marcado com uma barra inclinada (nome do elemento no mapa do dador / nome do elemento no mapa do seguidor). A inexistência de um elemento num dos mapas é marcada com um traço. O par de mapas é o segundo do segundo quarteto, sendo o elemento mestre escolhido para estudar o aspecto de velarização do /l/, com contraste e sem acordo entre dador e seguidor.

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

As figuras 1 e 2 ilustram respectivamente os mapas do dador e seguidor do diálogo piloto.

3 Gravação

Tanto o diálogo piloto, como os restantes diálogos gravados até à data (cerca de metade), decorreram numa câmara insonorizada. Os falantes encontram-se sentados com uma mesa de apoio para o mapa, virados para a parede da câmara e 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 microfone acoplado. Cada microfone está ligado a um canal independente da mesa de mistura ligada por sua vez a um gravador digital DAT. A digitalização é efectuada a uma frequência de amostragem de 48 KHz e o armazenamento é efectuado em cassete de fita magnética de 4mm. O sinal do microfone do falante 1 é gravado no canal esquerdo enquanto que o do falante 2 se grava no canal direito. Uma vez que ambos os falantes se encontram na mesma câmara, o microfone de um falante capta também o sinal acústico do outro falante, embora com um nível muito inferior.

O monitor da gravação, instalado na antecâmara, dispõe também de um microfone ligado à mesa de mistura para comunicar com os falantes antes e depois da sessão de gravação. Ambos os falantes recebem o mesmo som de retorno enviado pela mesa de mistura que incluiu o sinal do seu próprio microfone, o do microfone do outro falante e, no caso de se encontrar ligado, do microfone do monitor.

Antes de iniciar o diálogo, os dois falantes identificam-se, dizendo o nome e a data (dador) ou o nome e a hora de gravação (seguidor). Depois de concluir o diálogo, os dois falantes lêm as listas anexas contendo os nomes de todos os elementos presentes nos respectivos mapas.

A digitalização da onda sonora é feita a 16 kHz em stereo (amostras intercaladas dos 2 canais).

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. Devido à morosidade do trabalho de anotação e a um conjunto de restrições de ordem material e humana, apenas um pequeno subconjunto dos diálogos recolhidos poderá ser objecto de uma análise mais fina.

Uma das tarefas que tem vindo a ser objecto de especial cuidado é a da representação ortográfica que é assegurada para a totalidade do *corpus*. Para além da transliteração em orto-

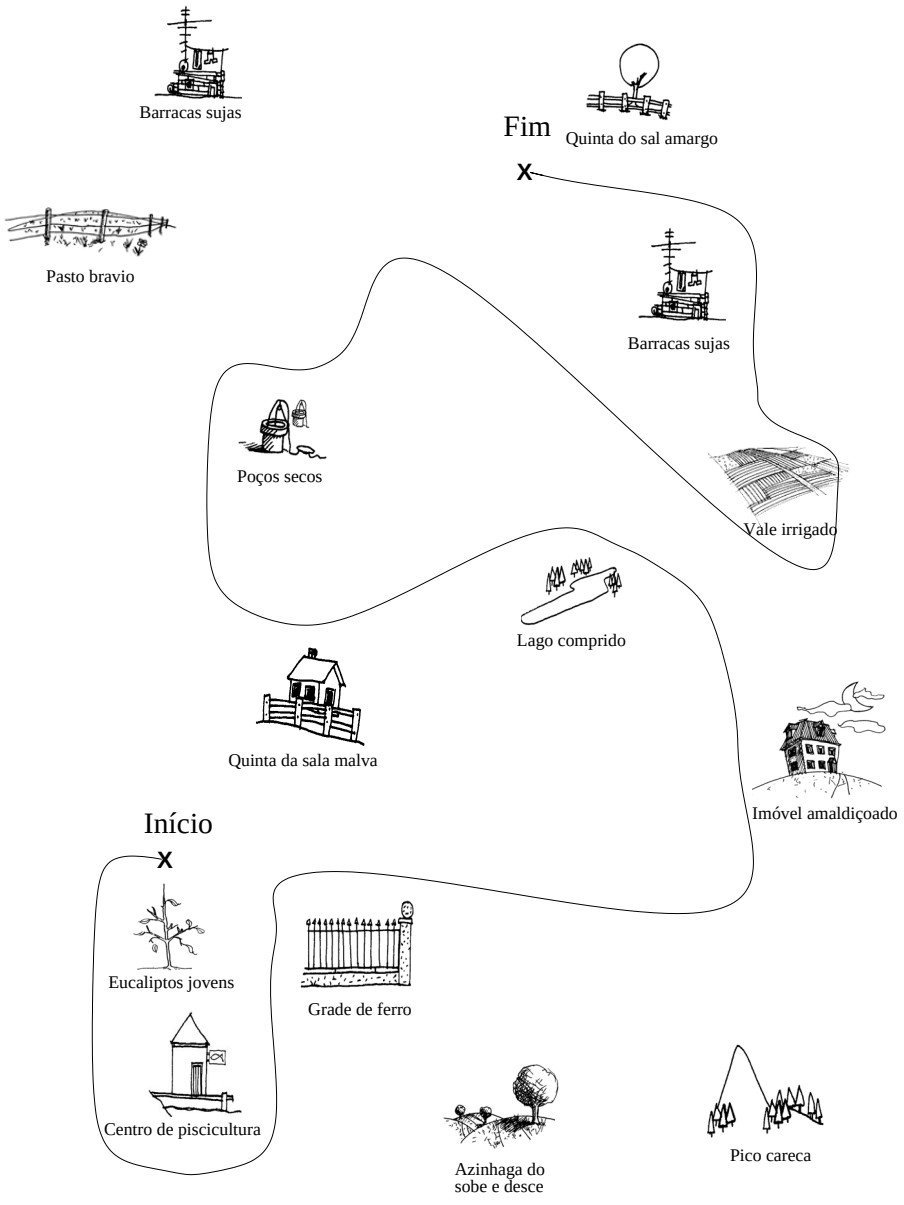

Figura 1: Mapa do dador

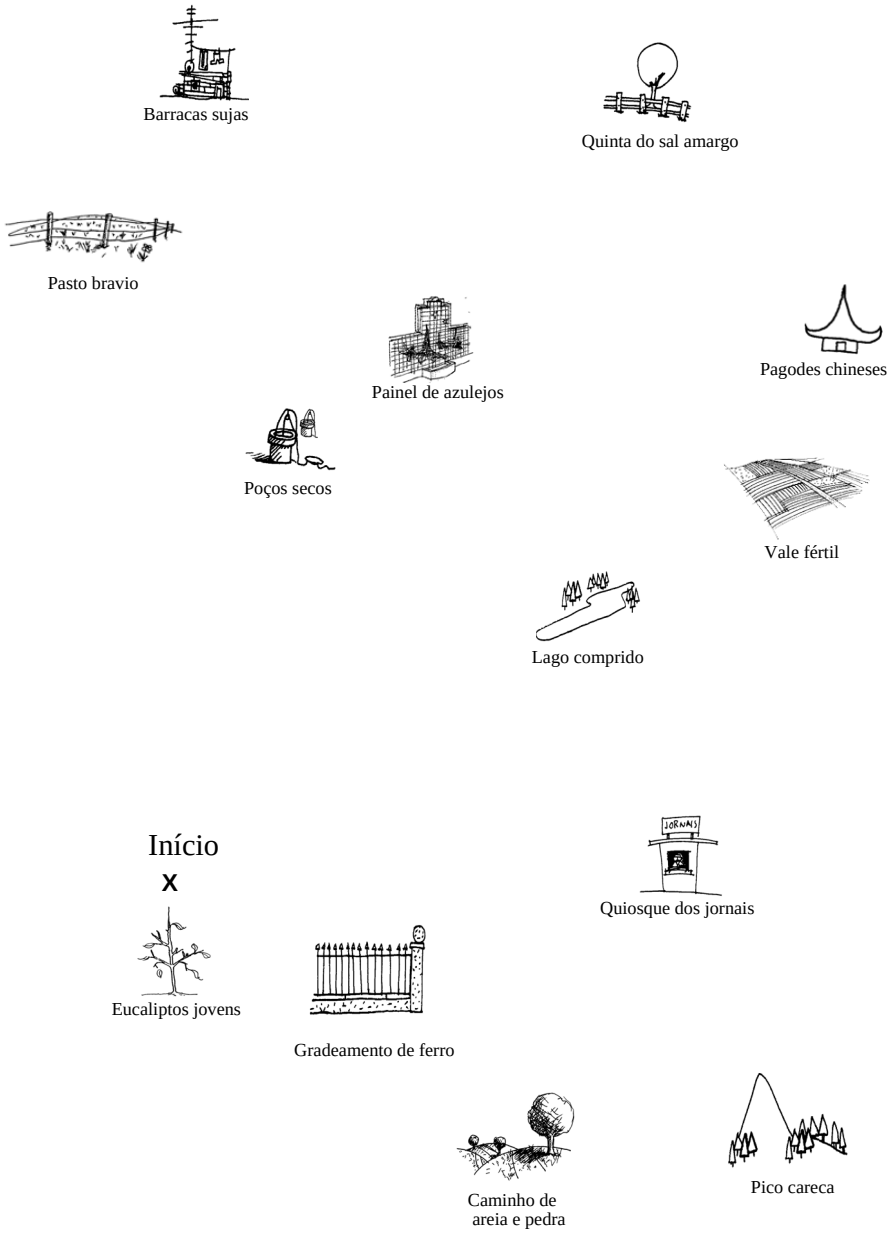

Figura 2: Mapa do seguidor

grafia corrente do que foi dito, este nível de representação inclui um conjunto de anotações que se destinam a permitir um acesso fácil da generalidade dos utilizadores aos conteúdos dos ficheiros de sinal e ainda a facilitar o processamento automático ou semi-automático posterior.

À semelhança do que tem vindo a acontecer em projectos congéneres, procurou-se tornar o texto tão legível quanto possível, utilizando apenas anotações bastante elementares, com formato simples e invariante e facilmente removíveis. Estas permitem identificar, no entanto, diferentes tipos de unidades e localizar quer porções de sinal correspondentes a fala fluente quer porções onde ocorrem diferentes fenómenos típicos da situação de fala espontânea que têm de ser objecto de especial cuidado em fases posteriores de tratamento do *corpus*. Embora o formato das anotações se tenha baseado, em parte, em recomendações do projecto TEI (Text Encoding Initiative [3]) e no standard SGML (Standard Generalized Mark-Up Language), o conhecimento destas normas não é de todo necessário para a compreensão da anotação, bastando para isso uma breve explicação.

As tomadas de palavra são unidades fundamentais na análise do discurso espontâneo, correspondendo ao intervalo de tempo de fala de um interlocutor, até este passar a palavra a outro ou a palavra lhe ser retirada por outro. Como é do conhecimento geral, nem sempre se verifica, no entanto, uma alternância clara de tomadas de palavra: os interlocutores podem começar a falar ao mesmo tempo (ou quase ao mesmo tempo) e podem interromper-se um ao outro com diferentes intuitos, tomando a palavra ou não. Dada a complexidade destas situações, é fundamental procurar anotar quem disse o quê e quando, identificando os falantes e reflectindo de alguma maneira a ordem cronológica, sem obscurecer (ou quebrar) a continuidade inerente a uma tomada de palavra.

Uma vez que um dos objectivos deste tipo de *corpus* é justamente o de vir a permitir o estudo das estratégias utilizadas na gestão das tomadas de palavra, optou-se pela anotação das intervenções de cada um dos interlocutores por ordem cronológica e em unidades independentes, localizáveis no sinal acústico.

Cada unidade é iniciada por duas linhas em formato SGML, delimitadas por parêntesis angulares, a primeira permitindo identificar o falante e o número da unidade de intervenção e a segunda indicando o número da amostra no ficheiro de sinal que corresponde ao início dessa mesma unidade.

<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.

Os excertos de fala sobreposta são indicados entre parêntesis rectos, obrigando a uma mudança de unidade, em que é localizado o início da sobreposição. No exemplo acima, o falante ”F” produziu um ”sim” de assentimento, enquanto o falante ”G” dizia ” e os”. O início da sobreposição foi marcado em <sfo samp=2732665>.

Como os interlocutores alternam obrigatoriamente nestes casos, são utilizadas outras convenções para indicar se a uma mudança de unidade corresponde ou não uma tomada de palavra (maiúscula ou minúscula, respectivamente, no início da primeira palavra da unidade).

As linhas de texto contêm ainda micro-anotações (entre chavetas) que podem ser parcial ou totalmente removidas. Dentro das chavetas é introduzido obrigatoriamente um código que permite identificar a ocorrência de um determinado tipo de evento. Se esse evento não afecta nenhum material lexical não é acrescentada mais nenhuma informação. Se afecta, coloca-se o separador ”|” e especifica-se. Essa especificação pode ser simplesmente ortográfica ou conter indicações de ordem fonética. Neste último caso, a transcrição fonética é dada em primeiro lugar, separada da forma ortográfica pelo separador ”=”. No exemplo acima, {pp} indica simplesmente a ocorrência de uma pausa e {ph|p”OSu=poços} corresponde a uma anotação fonética usando o alfabeto SAMPA (Speech As-

[3] http://www.uic.edu/orgs/tei/

sessment Methods Phonetic Alphabet [4]) que indica que a palavra "poços" foi pronunciada de forma não esperada.

As micro-anotações dão conta de ocorrência de determinados tipos de disfluências (descontinuidades prosódicas, repetições com ou sem correcção de material lexical, pausas preenchidas, etc.) e, em conjunto com os sinais de pontuação, fornecem indicações essenciais para a interpretação sintáctico-semântica do material produzido.

É utilizado um número reduzido de sinais de pontuação para demarcar fronteiras de unidades entoacionais que apenas procuram reflectir um conjunto de categorias básicas funcionais: continuidade do fluxo discursivo (","), terminalidade (".") e apelo à intervenção do interlocutor ("?"). Julgou-se ainda conveniente utilizar dois sinais adicionais: um para marcar descontinuidades prosódicas que não são cobertas pelas micro-anotações ("...") e outro para assinalar as unidades de carácter exclamativo ("!").

A par com o ficheiro de anotação ortográfica, é também produzido um outro ficheiro que contem apenas as marcas temporais do início e fim de cada tomada de palavra, incluindo as sobreposições. Embora apenas a marca de início seja copiada para o ficheiro de anotação ortográfica, a marcação de ambas permitirá, por exemplo, o estudo de tempos de reacção num futuro próximo.

Para cada ficheiro de fala são produzidos automatica ou semi-automáticamente diferentes níveis de transcrição fonética. O primeiro corresponde à forma como as palavras são pronunciadas quando produzidas isoladamente e em estilo cuidado (forma canónica ou forma de citação). Esta representação é gerada automaticamente com o módulo de conversão grafema-fone do sintetizador DIXI [4]. Utilizando também o mesmo módulo, é gerado um segundo nível de representação (transcrição fonética larga ou fonotípica), em que são tidos em conta diferentes processos (assimilação, elisão, supressão, etc.) tanto na palavra como na frase.

Com base no reconhecimento de segmentos fonéticos com modelos de Markov não observáveis (HMM - Hidden Markov Models) e utilizando o pacote de programs HTK (HMM Toolkit), desenvolvido pela Uiversidade 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.

O recurso às informações fonéticas contidas nas micro-anotações é de grande utilidade para assegurar um alinhamento aceitável, uma vez que são frequentes situações em que a 6 ou 7 segmentos da transcrição fonotípica, apenas correspondem três acusticamente identificáveis (ex: {ct|p"ot=para o outro}).

Dada a extrema morosidade das correcções manuais e o facto de estas exigirem o recurso a anotadores especializados, este nível de anotação apenas é contemplado para um subconjunto dos materiais do *corpus*. O mesmo se verifica com as anotações prosódicas, integralmente manuais.

Na transcrição prosódica seguem-se basicamente as propostas do sistema ToBI (de *Tone and Break Indices*) [7] [1] [5], representando diferentes aspectos da prosódia em fiadas paralelas alinhadas com a representação ortográfica e o sinal acústico.

Na sequência dos trabalhos que têm vindo a ser realizados para várias línguas e dos testes efectuados sobre materiais do Português Europeu, foram no entanto introduzidas algumas alterações, nomeadamente no que diz respeito ao tratamento das disfluências em que se seguem de perto as propostas de [6]. As disfluências são, por conseguinte, consideradas num nível independente, reservando-se o nível miscelâneo para a anotação de outros fenómenos de carácter para- ou extra-linguístico. É também feito, por outro lado, uso sistemático de um conjunto de diacríticos que asseguram um alinhamento mais preciso dos tons com o material segmental e o sinal acústico.

As anotações de carácter sintáctico-semântico são, de momento, integralmente manuais, estando a ser asseguradas, também, apenas para o mesmo subconjunto do *corpus* que está a ser tratado com maior detalhe.

De modo a permitir uma clara identificação

[4] http://www.phon.ucl.ac.uk/home/sampa/home.htm

das propriedades sintácticas a ter em conta na análise do *corpus*, foi desenvolvida uma metodologia de anotação que considera os seguintes níveis distintos:

- 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 e 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.
- Discurso:
  - Nível informacional: identificação da estrutura informacional de cada enunciado e correspondente etiquetagem;
  - Nível temático: identificação da estrutura temática de cada enunciado e correspondente etiquetagem categorial (TOP - Tópico marcado; COM - Comentário).
- Reformulações e disfuncionalidades:
  - Identificação do ponto inicial de reformulações e rupturas sintácticas devidas quer a dificuldades do planeamento sintáctico on-line quer a erros imputáveis a violações de propriedades de regência dos itens lexicais.

Como esperado, ocorrem nos diálogos já gravados fenómenos e construções tidos como recorrentes no diálogo espontâneo: vários tipos de construções elípticas (e.g., elipse de sintagma verbal, despojamento), estruturas de coordenação, construções com focos informacionais e quantificacionais marcados, interrogativas focalizadas, interrogativas-*tag*, e, de uma forma geral, o relaxamento ou o recuo de construções hipotácticas em favor da parataxe.

A análise do *corpus* recolhido permitirá uma melhor compreensão de alguns dos fenómenos e construções referidos, quer do ponto de vista da sua representatividade, quer do ponto de vista das suas propriedades, uma vez que a existência de um contexto bem caracterizado possibilitará não só a identificação rigorosa dos antecedentes das expressões anafóricas e elípticas como também uma análise fundamentada da estrutura informacional e temática dos fragmentos discursivos relevantes. A análise do *corpus* a anotar poderá ainda suscitar novos problemas a aprofundar.

5 Conclusões e trabalho futuro

O artigo descreve um *corpus* de diálogo falado cujo objectivo principal é o estudo de vários fenómenos típicos da fala espontânea, num domínio restrito. De particular interesse para este objectivo é a anotação do *corpus* a vários níveis: ortográfico, fonético, prosódico, sintáctico e semântico.

As potencialidades de exploração de um *corpus* deste tipo são inúmeras, tanto do ponto de vista de investigação fundamental como aplicada. De momento, no entanto, o estabelecimento de correlações entre a parentização sintáctica e o fraseamento prosódico tem vindo a ser objecto de especial atenção. Saliente-se também a utilização de *corpora* deste tipo para o estudo de esquemas de codificação da estrutura de 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 [2]. 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

[1] M. Beckman e G. Ayers, "Guidelines for Tobi_Labeling", Columbus, Ohio State University, 1994. (http://ling.ohio-state.edu/Phonetics/E_ToBI/etobi_homepage.html)

[2] J. Carletta, A. Isard, S. Isard, J. Kowtko, G. Doherty-Sneddonn e A. Anderson, "The Reliability of a Dialogue Structure Coding Scheme", Computational Linguistics, Volume 23, pp. 13-31, 1997.

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

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

[5] J. Pitrelli, M. Beckman e J. Hirschberg, "Evaluation of prosodic transcription labeling reliability in the ToBI framework", Proc. ICSLP94, Yokohama, Japão, pp. 123-126, 1994.

[6] E. Shriberg, "Preliminaries to a theory of speech disfluencies", PhD. Diss., Univ. of California at Berkeley, 1994.

[7] K. Silverman, M. Beckman, J. Pitrelli, M. Ostendorf, C. Whightman, P. Price, J. Pierrehumbert e J. Hirschberg, "ToBI: a standard for labeling English prosody", Proc. ICSLP'92, Banf, Alberta, pp.867-870, 1992.