CORPUS DE AQUISIÇÃO DO PORTUGUÊS EUROPEU: A PRIMEIRA FASE

Isabel Hub Faria
Lab. Psicolinguística, DLGR, FLUL

Na sequência de um projecto de investigação anterior, do Laboratório de Psicolinguística, financiado pela JNICT (PCSH/C/LIN/154/91 – 'Produção, Compreensão e Aquisição em Português Europeu') foi submetido ao Programa Estímulo para as Ciências Sociais e Humanas da JNICT, um novo projecto, visando a construção de um banco de dados de linguagem infantil, com foco no Português Europeu.

Tal banco de dados tem como objectivo principal constituir uma base computacional credível que possa adequadamente permitir, num futuro próximo, o acesso de investigadores nacionais e estrangeiros a dados empíricos fiáveis – i.e., sistematicamente recolhidos, transcritos e codificados segundo princípios já amplamente testados – para elaboração de diversos estudos na área da aquisição do Português como língua materna.

Ao ser assegurada a internacionalização dos dados do Português Europeu através da entrada numa numa rede de dados internacional, acreditamos poder vir a encontrar, nas revistas científicas de maior divulgação internacional, e já num futuro próximo, referências correctas de e ao Português Europeu, aspectos de há muito não controlados, a avaliar pelas 'trapalhadas' que temos encontrado fazendo-se passar por dados do Português.

Tratando-se de registos de linguagem das crianças, optámos pela adesão à rede mais utilizada na área, a CHILDES Database[1].

Da autoria de Leonid Spektor (Carnegie Mellon), Brian MacWhinney (Carnegie Mellon), Catherine Snow (Harvard), Julia Evans (Carnegie Mellon) e Barbara Pan (Harvard), o sistema CHILDES é um instrumento computacional de apoio ao estudo da linguagem oral, com especial relevo para a análise da interacção verbal que constitui, naturalmente, pelas trocas entre mãe-criança, pai-criança, adulto-criança, criança-crianç, investigador-criança, a base empírica do estudo da aquisição e do desenvolvimento da linguagem.

Para além do CHAT (sistema de transcrição) e da Base de Dados constituída por registos de natureza longitudinal e transversal, monolingue e bilingue, normal e patológico, de várias dezenas de línguas, pussui um conjunto de programas de análise, CLAN, de que focamos os seguintes:

CED	editor de textos ASCII para trabalhar com ficheiros em formato CHAT
CHAINS	análiseador de discurso
CHECK	verificador da correcção dos ficheiros CHAT
CHIP	analisador de interacção verbal
CHSTRING	procura e substituição de cadeias de caracteres
COMBO	procura de expressões ou das suas combinatórias, ex: busca de palavras, classes de palavras, construções
FREQ	cálculo de frequências de formas lexicais, listagens. Pode funcionar em conjunto com o STATFREQ.
GEM	analisador de discurso, selecção de fragmentos especificados na transcrição
KEYMAP	análise de discurso, tabela de contingências para uma dada palavra ou código
KWAL	obtenção de linhas de produção que contêm uma palavra-chave
MAXWD	localização, medida e impressão da palavra ou da linha mais extensa
MLT	determinação da extensão média da intervenção

MLU	cálculo da extensão média do enunciado
MODREP	analisador de fala patológica
MOR	analisador morfológico automático ou semi-automático
PHONFREQ	cáculo de frequência de segmentos ffonológicos
REVCONC	concordância inversa, listagens de terminações de palavras, permite a análise de marcadores flexionais
STATFREQ	conversão do output de FREQ em dados estatísticos

O projecto foi aprovado pela JNICT (Projecto PCSH/C/LIN/ /524/93) com a designação de 'Corpus de Aquisição do Português Europeu – CHILDES Database', e teve início, no Laboratório de Psicolinguística da FLUL, em Janeiro de 1994. Desde essa altura, e até Outubro de 1995, foram levadas a cabo diversas tarefas que passamos a enunciar:

1. Elaboração da primeira versão de um pequeno manual em Português – *CHILDES: Uma Adaptação para o Português Europeu* – tarefa executada por Catarina Moraes, Dina Bárbara, Florbela Barreto, Iva Simas, Rita Veloso e Rosário Lopes, sob a orientação de Isabel Hub Faria e Hanna Batoréo.

2. Organização e realização de um *CHILDES workshop*, orientado pelos autores do CHILDES, Brian MacWhinney e Catherine Snow. O workshop teve lugar durante o *First Lisbon Meeting on Child Language*, encontro organizado em colaboração com a APL, Faculdade de Letras da Universidade de Lisboa, em Junho de 1994.

3. Recolha, transcrição e revisão das primeiras narrativas relativas ao corpus português da *'História da Rã'*, corpus internacionalmente ligado ao projecto de análise interlinguística coordenado por Dan Slobin e Ruth Berman[2], presentemente constituido por:

– narrativas de adultos que, no interior da análise, funcionam como grupo de controlo (27 adultos do sexo feminino e nível de instrução superior);

– narrativas de crianças (19 crianças igualmente do sexo feminino com idades compreendidas entre os 4;00 e os 10;00 anos, umas não escolarizadas, outras que frequentam os primeiros anos de escolaridade).

Este corpus encontra-se transcrito e, apesar de estar longe de ser representativo relativamente à população infantil, pode desde já ser disponibilizado para teses de Mestrado sobre aquisição e desenvolvimento da linguagem.

4. *'Corpus Hanna Batoréo'* – registo audio e transcrição de cento e vinte narrativas relativas ao corpus português das *'História do Cavalo'* e *'História do Cão e do Gato'*, estímulos visuais utilizados por Maya Hickmann e que, à semelhança da 'História da Rã', têm sido amplamente utilizadas para a elaboração de análises interlinguísticas. Existem 60 narrativas de cada história (30 adultos, de ambos os sexos, amostra que funciona como grupo de controlo; e 30 crianças, também de ambos os sexos, com idades entre 4;06 e 10;00 anos). Este corpus só ficará disponível após a realização da tese de doutoramento de Hanna Batoréo.

5. Revisão da codificação do *'Corpus Dília Pereira'* – corpus recolhido e inicialmente transcrito por Dília Ramos Pereira com vista à elaboração da sua tese de mestrado[3] e, actualmente, objecto de estudo conducente à sua tese de doutoramento. Este corpus é constituído por 30 pares mãe-criança (10 crianças com idades de1;00 a 1;06, 5F e 5M; 10 crianças com idades de 1;07 a 2;00, 5F e 5M; 10 crianças com idades de 2;01 a 2;06, 5F e 5M). Para cada par mãe-criança existem registos de três situações (durante uma refeição, durante um banho, e a brincar), gravação audio em contexto intrafamiliar. Este corpus não se encontra disponível até à realização da tese de doutoramento de Dília Pereira.

6. Transcrição e codificação em CHAT do corpus longitudinal designado *'Corpus António Quintas Mendes'* constituído por registos vídeo de duas crianças: João Miguel (2;00-2;07) e Pedro Gil (2;07--3;04). Este corpus tinha já sido anteriormente transcrito pelo respectivo observador e, nessa versão, apareceu impresso como anexo à sua tese de mestrado (Quintas Mendes, FLUL, 1992). A transcrição em

CHAT foi realizada por Ernestina Carrilho, Fernanda Gonçalves e Maria Sousa Lobo que utilizaram estes dados para as respectivas teses de mestrado (cf. Fernanda Gonçalves, FLUL, 1994; Maria Sousa Lobo, FLUL, 1994; Ernastina Carrilho, FLUL, 1994). Irá proceder-se, em 1996, à sua revisão definitiva no interior do CHILDES.

7. *'Corpus Mª João Freitas'* – Foram registadas em vídeo, por Mª João Freitas, cerca de cem horas de corpus longitudinal, actualmente a ser transcrito e tratado no programa CHILDPHON (adaptação de C. Levelt e P. Fikkert, Max-Planck-Institute (Nijmegen)). São produções de 7 crianças com idades compreendidas entre os 0;10 meses e os 3;08 anos, em gravações realizadas mensalmente, de 30 a 60 minutos, em casa das crianças, em situação não estruturada, por vezes com a participação das respectivas mães.

Tem a seguinte constituição: João Pedro (24 sessões, de 0;10 a 2;10); Marta (12 sessões, de 1;03 a 2;03); Luís (12 sessões, de 1;10 a 2;11); Inês (12 sessões, de 0;11 a 2;00*); Pedro (12 sessões, de 2;07 a 3;07); Raquel (12 sessões, de 1;10 a 2;11); Laura (12 sessões, de 2;02 a 3;03). Este corpus não está, presentemente, totalmente transcrito no Programa CHILDES e só ficará disponível após a realização das provas de doutoramento de Mª João Freitas.

Com início em Setembro de 95, tem sido registada uma criança, Joana, a partir dos 0;03 meses de idade, registo que se prevê continuar por cinco anos.

Foram ainda realizadas, durante um ano, gravações com Catarina (4;00-5;00), corpus ainda por transcrever e que não será incluído na tese de doutoramento de M. J. Freitas.

8. *'Corpus recolhido por Isabel Leiria'* – dezoito sessões, gravadas em vídeo, de Sara (3;01 – 4;04), ainda por transcrever.

9. *'Corpus recolhido por Manuela Vasconcelos'* – dezasseis sessões, gravadas em vídeo, de João Gil (3;00 – 4;04). Doze das sessões já se encontram transcritas.

10. *'Corpus Isabel Hub Faria'* – Corpus de controlo constituído pelo registo (audio) de cem entrevistas, estruturadas por um guião de catorze perguntas, que foram efectuadas e gravadas nos locais de trabalho dos respectivos sujeitos. Uma primeira trancrição foi apresen-

tada como segundo volume da tese de doutoramento de I. H. Faria, *'Para a Análise da Variação Sócio-Semântica: Estrato sócio-profissional, sexo e local de produção enquanto factores reguladores, em Português Contemporâneo, das formas de auto-referência e de orientação para o significado'*, Lisboa: Universidade de Lisboa, 1983. Desta mesma tese foi publicado pelo INIC, em 1992, apenas o primeiro volume.

O corpus é constituído da segunte forma: 100 sujeitos adultos portugueses pertencentes a diferentes estratos sócio-profissionais (5 estratos sócio-profissionais, 20 sujeitos por estrato, 10 M e 10F). Metade dos sujeitos são do sexo feminino e a outra metade do sexo masculino (50M e 50F) e 50 gravações foram efectuadas na FLUL e outras 50 foram efectuadas na SCC.

Trata-se de um extenso conjunto de dados que, pela forma como foram estruturados na recolha, permite a extracção para análise de subconjuntos correspondentes às variáveis estrato sócio-profissional, sexo e local de trabalho. Uma parte da codificação inicial foi de Rita Veloso, estando, actualmente, a ser levada a cabo por Carla Soares. Prevê-se a conclusão da codificação para o final de 1996.

Os objectivos de constituição de um Corpus de Aquisição do Português Europeu não se encontram, como é natural, esgotados no actual projecto de investigação. Independentemente de conseguirmos levar a bom termo todas as tarefas já iniciadas, outras foram surgindo que deverão ser objecto de novo projecto. A próxima etapa deverá alargar a base de dados longitudinal monolingue, iniciar a base de dados patológica, e avançar para dados de bilingues com foco no Português Europeu como L2.

Notas

[1] Sobre Base de Dados CHILDES ver Faria & Batoréo (1994) in *Revista Internacional de Língua Portuguesa, 11*: 137-145.

[2] Para uma perspectiva da importância e extensão dos estudos interlinguísticos levados a cabo com base em recolhas da 'História da Rã' ver Berman, R. A. & Slobin, D. I. 1994. *Relating Events in Narrative: A Crosslinguistic Developmental Study*. Hillsdale, New Jersey: Lawrence Erlbaum. Pode encontrar-se neste volume uma referência ao corpus do Português Europeu, na página 671.

[3] Dília M. M. Ramos Pereira. A Linguagem Dirigida à Criança em Fases Iniciais da Aquisição do Português Europeu como Língua Materna: Aspectos Lexicais e Enunciativos.

Referências

BERMAN, R. A. & SLOBIN, D.I. 1994. *Relating Events in Narrative: A Crosslinguistic Developmental Study*. Hillsdale, >New Jersey: Lawrence Erlbaum.

CARRILHO, E. M. R. 1994. *A Topicalização e a Construção de Objecto Nulo no Desenvolvimento Sintáctico do Português Europeu: A produção espontânea de duas crianças dos 2;00 aos 3;03 anos*. Dissertação de Mestrado em Linguística Portuguesa Descritiva (Psicolinguística). Fac. Letras, Universidade de Lisboa.

FARIA, I. H. 1993. A Aquisição da Noção de 'Agente' e a Produção de Sujeitos Sintácticos por Crianças Portuguesas até aos Dois Anos e Meio. *Revista Internacional de Língua Portuguesa, 10*, 16-50.

FARIA, I.H. & BATORÉO, H. J. 1994. Corpus de Aquisição do Português Europeu: Base de Dados CHILDES. *Revista Internacional de Língua Portuguesa, 11*, 137-145.

GONÇALVES, F. M. R. 1994. *Negação Frásica em Português. Caracterização Sintáctica com Referência ao Processo de Aquisição*. Dissertação de Mestrado em Linguística Portuguesa Descritiva (Psicolinguística). Fac. Letras, Universidade de Lisboa.

HICKMANN, M. 1990. *The Development of Discourse Cohesion: Coding Manual*. Nijmegen: Max Planck Institute for Psycholinguistics.

LOBO GONÇALVES, M. F. H. S. 1994. *Para uma Redefinição do Parâmetro do Sujeito Nulo*. Dissertação de Mestrado em Linguística Portuguesa Descritiva (Psicolinguística). Fac. Letras, Universidade de Lisboa.

MACWHINNEY, B. 1991. *The CHILDES Project.Tools for Analyzing Talk*. Hillsdale, New Jersey: Lawrence Erlbaum.

MACWHINNEY, B. & SNOW, C. 1985. The Child Language Data Exchange System. *Journal of Child Language, 12*, 271-296.

MACWHINNEY, B. & SNOW, C. 1990. The Child Language Data Exchange System. An Update. *Journal of Child Language, 17*, 459-471.

QUINTAS MENDES, A. 1992. *A Referência Temporal no Discurso Conversacional aos 2/3 Anos de Idade*. Dissertação de Mestrado em Linguística Portuguesa Descritiva, Fac. Letras, Universidade de Lisboa.

RAMOS PEREIRA, D. 1992. *A Linguagem Dirigida à Criança em Fases Iniciais da Aquisição do Português Europeu como Língua Materna: Aspectos Lexicais e Enunciativos*. Dissertação de Mestrado em Linguística Portuguesa Descritiva, Fac. Letras, Universidade de Lisboa.

SOKOLOV, J.L. & SNOW, C. E. 1994. *Handbook of Research in Language Development Using CHILDES*. Hillsdale, New Jersey: Lawrence Erlbaum.