DOS PROBLEMAS DE CONSTITUIÇÃO ÀS POTENCIALIDADES DE UTILIZAÇÃO DE *CORPORA*: O CASO DO CIPM

Maria Francisca Xavier
Universidade Nova de Lisboa – F.C.S.H.

O CIPM – *Corpus* Informatizado do Português Medieval – foi objecto de uma comunicação apresentada por Xavier, Brocardo e Vicente ao X Encontro Nacional da Associação Portuguesa de Linguística, realizado em Évora, no ano passado. A informatização de textos medievais com vista à constituição do *corpus* tinha sido iniciada havia apenas alguns meses e este *corpus* está a ser desenvolvido como um subprojecto do Projecto "A Gramática do Português Medieval. Contributos para a sua caracterização (GPM)", da responsabilidade de uma equipa da Faculdade de Ciências Sociais e Humanas da Universidade Nova de Lisboa, subsidiado pela JNICT.

Apresentamos agora, um ano mais tarde, a constituição actual do CIPM, juntamente com uma breve reflexão sobre o trabalho feito e os planos da equipa para a sua ampliação e tratamento automático.

Os textos medievais até agora seleccionados e que actualmente integram o CIPM foram editados segundo princípios e critérios rigorosos e são todos eles considerados de grande interesse para estudos linguísticos históricos.

Ao procedermos à informatização das edições, procuramos manter, no essencial, os critérios de transcrição utilizados pelos editores. No entanto, cada texto passa por um trabalho prévio de pequenas adaptações, tendo em vista alguma simplificação e uniformização de

critérios, que são exigidas, essencialmente, pelo suporte lógico utilizado.

Alguns textos foram informatizados por leitura óptica, o que obriga a um esforço paciente de revisão e correcção dos erros produzidos pela deficiente interpretação automática de caracteres; outros textos foram digitados pelos próprios editores e foram-nos oferecidos, tendo estes últimos apenas sido submetidos a adaptações pontuais de critérios antes de serem incorporados no CIPM.

Tivemos até agora várias ofertas de edições em suporte informático, que muito agradecemos, porque vieram contribuir para o enriquecimento rápido e económico do *corpus*. A primeira oferta que tivemos foi o Arquivo de Textos do Português Antigo (AOPT – Archive of Old Portuguese Texts) de Stephen Parkinson (1983), da Universidade de Oxford, que colabora connosco desde o início do projecto; também os textos editados por M. Teresa Brocardo (1994) e M. de Lourdes Crispim (1995), membros da equipa do projecto, foram oferecidos para integrar o CIPM e, ainda, na sequência da nossa comunicação apresentada no ano passado, em Évora, M. Helena Garvão (1992) e M. Celeste Rodrigues (1992) ofereceram-nos espontaneamente os textos por elas editados.

A selecção dos textos e as decisões sobre a simplificação e uniformização de critérios necessárias para a sua inserção no *corpus* foi objecto de trabalho regular dos membros da equipa ao longo deste ano e meio. E o trabalho de informatização, correcção e adaptação de critérios foi sendo feito, paralelamente, por tarefeiras estudantes de Linguística.

Para preparação de listas, índices, estatísticas e concordâncias de palavras utilizamos o Programa de Concordâncias de Oxford – OCP (Oxford Concordance Program), que também nos foi oferecido por Stephen Parkinson. O OCP corre em MS-DOS para IBM pelo que os textos são convertidos em linguagem ASCII.

Temos ainda utilizado o seguinte suporte lógico e equipamento informático disponível no Gabinete de Informática da Faculdade:

Suporte Lógico
- OmniPage Professional;
- Word for Windows 2.0b.

Equipamento
- computador DIGITAL com processador 486 SX/25, 4 Mb RAM, ligado à rede de computadores;
- scanner HP Scanjet IIc with HP Accupage;
- impressora EPSON Page Printers EPL-4300.

Contamos, igualmente, com a assistência constante dos colaboradores daquele Gabinete. Porém, trabalhar tranquilamente num serviço da Faculdade que dá apoio a docentes e discentes torna-se muitas vezes difícil, e por vezes mesmo impossível, pelo que estamos a concorrer ao Praxis XXI e à JNICT para tentarmos conseguir o apoio financeiro imprescindível para melhorar as condições de trabalho e, simultaneamente, podermos desenvolver o CIPM.

A constituição actual do CIPM é a seguinte:

Século XIII

	Nº/palavras
História do Galego-Português (Maia (1986))[1]	28 804
Chancelaria de D. Afonso III (Duarte (1986))[1]	16 866
Notícia de Torto (Cintra (1990))[2]	779
Testamento de D. Afonso II (Costa (1979))[2]	
Manuscrito L	1 432
Manuscrito T	1 437
Tempos dos Preitos (Ferreira (1986))[3]	1 591
Afonso X, Foro Real (Ferreira (1987))[3]	49 679
Arquivo de Textos do Português Antigo (Parkinson)[4]	19 504
Foros de Garvão (Garvão (1992))[5]	7 395
	127 487

Século XIV

História do Galego-Português (Maia (1986))[1]	33 242
Afonso X, Primeyra Partida (Ferreira (1980))[6]	174 512
Dos Costumes de Santarém (Rodrigues (1992))[7]	37 021
	244 775

Século XV	
História do Galego-Português (Maia (1986))[1]	30 162
O Livro das Três Virtudes (Crispim)[8]	56 272
Crónica do Conde D. Pedro de Meneses (Brocardo 1994))[8]	136 661
	223 095
Século XVI	
História do Galego-Português (Maia (1986))[1]	3 107
Cópias Tardias	
Vidas de Santos de um Manuscrito Alcobacense (Castro (1982-1983; 1984-1985))[9]	29 566

Se compararmos a dimensão actual do *corpus* com a de há um ano atrás, verificamos que aumentou consideravelmente. Relativamente ao séc. XIII, que era o que naturalmente mais nos preocupava por existirem menos edições deste período, quase duplicámos o número de palavras – passámos de 68.822 para 127.487. A este número podemos ainda acrescentar, embora com algumas reservas por se tratar de cópias do séc. XV de textos originais do séc. XIII, 29.566 palavras. Do séc. XIV tinhamos apenas 33.242 palavras e temos agora 244.775. E aos textos do séc. XV não acrescentámos nada, mantemos as 223.095 palavras.

Consideramos que para a descrição e análise que estamos a desenvolver no âmbito do Projecto GPM o *corpus* se encontra completo. Segundo Kroch, linguista habituado a trabalhar com *corpora* informatizados, é possível começar a fazer análise morfossintáctica com base em 50.000 palavras, pelo que nos sentimos confortáveis com a dimensão actual do CIPM para poder dar cumprimento à análise linguística prevista naquele projecto.

O interesse suscitado pelo CIPM levou à constituição de uma equipa interdisciplinar que inclui investigadores das áreas da Linguística, da Informática, da História e da Cultura Medievais. O projecto elaborado por esta equipa tem como principal objectivo contribuir para um melhor acesso a fontes textuais (em termos qualitativos, de rapidez

e de economia), através da implementação de um sistema automatizado de análise que permita optimizar a investigação nas diferentes áreas sobre *corpora* do português medieval.

Assim, se conseguirmos os apoios imprescindíveis, continuaremos o trabalho de informatização de *corpora* textuais em português medieval e desenvolveremos a análise morfossintáctica nesessária à concepção do suporte lógico que venha a permitir a automatização da etiquetagem e da segmentação sintagmática, bem como a construção de um léxico electrónico.

Notas

1 Informatização de Susana Pereira e Teresa Oliveira.

2 Informatização de António Emiliano.

3 Informatização de Susana Pereira e Alexandra Fiéis.

4 O arquivo, informatizado por Stephen Parkinson, é composto por revisões críticas de edições já publicadas e transcrições originais de inéditos (v. Parkinson (1983)).

5 Adaptação da informatização do editor por Alexandra Fiéis.

6 Informatização de Alexandra Fiéis e Cristina Silva.

7 Adaptação da informatização do editor por Alexandra Fiéis e Cristina Silva.

8 Informatização do editor.

9 Textos originais do s. XIII; cópias do s. XV; informatização de Carlos Rocha e Carla Laranjeira.

Referências

BROCARDO, M. T. (1994) *Crónica do Conde D. Pedro de Meneses de Gomes Eanes de Zurara. Edição e Estudo.* Faculdade de Ciências Sociais e Humanas da Universidade Nova de Lisboa (dissertação de doutoramento).

CASTRO, I. (ed.) (1982-1983); (1984-1985) "Vidas de Santos de um Manuscrito Alcobacense", *Revista Lusitana*, Nova Série, 4,5, Lisboa, Centro de Estudos Geográficos, pp. 5-52; 43-71.

CINTRA, L. F. L. (1990) "Sobre o mais Antigo Texto Não-Literário Português: a 'Notícia de Torto' (Leitura Crítica, Data, Lugar de Redacção e Comentário Linguístico)", *Boletim de Filologia* 31, pp. 22--77.

COSTA, A. J. (1979) "Os mais Antigos Documentos em Português. Revisão de um Problema Histórico-Linguístico", *Revista Portuguesa de História* 17, pp. 263-340.

CRISPIM, M. L. (1995) *Edição Crítica de O Livro das Três Virtudes de Christine de Pizan.*Faculdade de Letras da Universidade de Lisboa (dissertação de doutoramento).

DUARTE, L.F. (1986) *Os Documentos em Português da Chancelaria de D. Afonso III (Edição)*, Faculdade de Letras da Universidade de Lisboa (dissertação de mestrado).

FERREIRA, J. A. (1986) "Edição e Estudo Linguístico dos 'Tempos dos Preitos'" in J. Roudil, *Jacobo de Junta. Summa de los Nueve Tiempos de los Pleitos. Édition et Étude d'une Variation sur un Thème*, Paris, Klincksieck.

FERREIRA, J. A. (1987) *Afonso X. Foro Real, Edição, Estudo Linguístico e Glossário*, 2 vols., Lisboa, INIC.

HOCKEY, S. M. (1988, 1993) *Micro-OCP User Manual*, Oxford University Computing Services, Oxford, Oxford University Press.

KROCH, A. S. (1994b) "Morphosyntactic Variation", in K. Beals et al. (eds.) *Papers from the 30th Regional Meeting of the Chicago Linguistics Society: Parasession on Variation and Linguistics Theory (to appear).*

MAIA, C. A. (1986) *História do Galego-português. Estudo Linguístico da Galiza e do Noroeste de Portugal do Século XIII ao Século XVI*, Coimbra, INIC.

PARKINSON, S. (1983) "Um Arquivo Computorizado de Textos Medievais Portugueses", *Boletim de Filologia* 28, pp. 241-252.

XAVIER, M. F.; M. T. Brocardo e M. G. Vicente (1994) "CIPM – um *Corpus* Informatizado do Português Medieval" in *Actas do X Encontro da Associação Portuguesa de Linguística*, Lisboa, pp. 599-612.