Português Antigo
Construção e Disponibilização de Recursos em Suporte Informático

M. Francisca Xavier
M. Lourdes Crispim
M. Graça Vicente
Centro de Linguística da Universidade Nova de Lisboa

Num momento em que a comunidade linguística reconhece a importância da chamada "linguística de *corpora*", e em que o interesse pela língua do passado aumenta mesmo em domínios teóricos que, não há muito tempo ainda, se debruçavam unicamente sobre as línguas no seu estado presente, tornou-se evidente, no caso do Português, a necessidade de recursos de fácil acesso e utilização por parte de estudantes e da comunidade científica, nacional e internacional.

A construção desses recursos passa pela existência e disponibilização de *corpora* informatizados e anotados, pela realização de estudos a diferentes níveis e pela elaboração de glossários, dicionários e terminologias, os quais poderão vir a ser fácil e rapidamente consultados através da *Web* e utilizados para novas investigações.

Foi com estes pressupostos que há uma década começou a ser desenvolvido no Centro de Linguística da Universidade Nova de Lisboa[1] um projecto que tem por objectivo a informatização e a divulgação de textos antigos – alguns deles de difícil acesso na edição em papel – e a construção de outros recursos linguísticos, nomeadamente um Dicionário, baseados naqueles textos – que permitam um melhor conhecimento do Português Medieval.

Ao fazer o balanço do trabalho desenvolvido neste período com vista à construção do CIPM – Corpus Informatizado do Português Medieval, é sobretudo gratificante o reconhecimento, por parte daqueles que o têm utilizado – estudantes de pós-graduação e outros estudiosos, nacionais[2] e estrangeiros[3], maioritariamente de linguística, mas também de história, de cultura e de literatura, de que se trata de um instrumento de grande interesse e utilidade para as suas investigações.

O CIPM contém actualmente cerca de 2,3 milhões palavras, extraídas dos mais antigos textos escritos em Português editados até agora, assim como de textos de

---

[1] A equipa deste projecto é constituída por linguistas e estudantes da Faculdade de Ciências Sociais e Humanas da Universidade Nova de Lisboa que integram a Linha de Investigação 1 – Linguística Comparada – do Centro de Linguística da UNL.

[2] Investigadores nacionais: Universidade Nova de Lisboa; Universidade do Minho; Universidade da Beira Interior; Universidade de Lisboa; Universidade de Coimbra; Universidade Católica Portuguesa.

[3] Investigadores provenientes de diferentes países europeus, nomeadamente Espanha, França, Itália, Inglaterra, Alemanha, Noruega e Dinamarca, assim como das Américas – Brasil e Estados Unidos.

***Actas do XVIII Encontro Nacional da Associação Portuguesa de Linguística*, Lisboa, APL, 2002, pp. 859-867.**

vários tipos e géneros até ao século 16. O acervo textual é constituído por uma importante colecção de documentos notariais (do século 12 ao século 16), foros, chancelarias, crónicas, textos didácticos e moralísticos e cantigas de escárnio e maldizer, todos eles editados por especialistas: uns filólogos – alguns deles linguistas –, outros historiadores e outros ainda investigadores das áreas da literatura e da cultura[4]. As edições não são homogéneas, como homogéneo não é o público que tem utilizado e continuará, certamente, a utilizar o CIPM.

Seja qual for o tipo de edição, porém, a preparação dos textos para o tratamento informático exige um conjunto de notações editoriais coerentes, que permitam uma correspondência biunívoca entre os sinais utilizados e as realidades que representam.

Cada texto é apresentado com um conjunto de referências que informam sobre data, proveniência geográfica do manuscrito, etc. São apresentadas também como comentário, assinaladas entre (()), informações como o assunto ou a mudança de linha do manuscrito.

Dada a diversidade de soluções gráficas que os editores adoptam para assinalar as suas intervenções ou os acidentes das fontes, procede-se a uma normalização gráfica das transcrições de acordo com os critérios estabelecidos para o CIPM. Considera-se que os textos informatizados, destinando-se principalmente a análises linguísticas, algumas automáticas, não necessitam de conter elementos que não são pertinentes para essas análises (mesmo sendo pertinentes do ponto de vista editorial). Assim, as informações sobre o aparato crítico das edições (textos introdutórios e notas) não são introduzidas nos textos informatizados, embora se tenha recorrido a elas para interpretar notações.

As normas de informatização estabelecidas são as seguintes:

1. Abreviaturas:
   (i) desenvolvimento transcrito entre ():
   Ex. m(orador)
   (ii) desenvolvimento duvidoso marcado no fim da palavra e sem espaço, por (?):
   Ex. fr(atre)s(?)
   (iii) não desenvolvimento marcado por (—?):
   Ex. Eo(—?)

2. Outras intervenções dos editores transcritas entre [ ]:
   (i) reconstituições de partes ilegíveis, palavras ou grafemas raspados ou atingidos por acidente do suporte:
   Ex. [Co]noç[u]da

[4] Ver no final lista dos textos que integram actualmente o CIPM e referências das fontes.

(ii) preenchimento de lacunas ou acrescentos correspondentes a grafemas ou palavras:
Ex. podero[so]; [por]
(iii) emendas por substituição ou permuta de caracteres:
Ex. patre > pa[rt]e; daras > [f]aras
(iv) não preenchimento de lacunas imputáveis quer ao escriba quer a deterioração do material, independentemente da sua extensão, indicadas por reticências: [...]
(v) símbolo gráfico não legível indicado por um ponto: [.]
(vi) leituras alternativas precedidas por ?:
Ex. [?vijr]

4. Grafemas ou palavras em letra diferente no manuscrito são indicados entre / /:

Ex. ant/e/

5. Leitura duvidosa de palavras ou símbolos assinalada imediatamente a seguir à palavra ou símbolo por /?/:
Ex. nahu~a/?/

6. Palavra com erro não corrigido ou forma estranha seguida de /sic/:
Ex. erda/sic/

7. Grafemas ou palavras presentes nos textos mas que os editores consideram não deverem ler-se, assim como repetições dos copistas conservadas e assinaladas no texto editado, representam-se entre | |:
Ex. demandado|r|

8. Entrelinhados (grafemas, palavras, frases) são transcritos entre ‖ ‖:
Ex. ‖domos‖

9. Grafemas ou palavras riscados figuram entre { }.
Ex. {M(a)r(avedi)}

10. Grafemas ou palavras borrados figuram entre // //.
Ex. //logar//

11. Excertos em latim são indicados entre {{ }}:
(i) presentes nos textos:
Ex. {{in secula seculoru~.}}
(ii) suprimidos pelo editor, com reticências: {{...}}

12. Adaptação grafemática

(i) Diacríticos que figuram sobre o grafema na edição encontram-se à direita deste (à excepção de ñ):
Exs. ã → a~
á → a'
ê → e^
y com barra sobreposta → y~
ỹ (= y nasal) → y~

(ii) Outros grafemas:
- nota tironiana → &
- σ → s
- ſ → s
- z visigótico → z
- γ → r
- ρ → r

13. Pontuação
- caldeirão → $
- ponto final de texto → %

14. Numeração romana
- X'(aspado) (= 40) → X^{L}
- milhares indicados por números romanos com barra sobreposta, representados por esses mesmos números seguidos de M em expoente: Ex. III com barra sobreposta → III^{M}

15. Supressões aquando da inclusão no CIPM dos textos editados
- ponto, nos numerais e nas séries: .III. → III; a.b.c. → abc
- signos notariais
- sinal de translineação
- sinal de corte de linha

Este corpus textual, informatizado de acordo com os critérios acima referidos, contém actualmente textos em diferentes fases de revisão, embora esteja desde há já alguns anos parcialmente disponível para os estudiosos que solicitam a sua utilização.

Um subcorpus do CIPM – cerca de 500.000 palavras, o que constitui apenas uma amostra do conjunto – encontra-se etiquetado, permitindo a identificação morfossintáctica das palavras[5]. Apresentam-se as etiquetas utilizadas:

[5] O etiquetador que tem sido utilizado é o apresentado em MARQUES, N.; G. Lopes (1996) "Using Neural Nets for Portuguese Part-of-Speech Tagging" in *Proceedings of the5th International Conference on the Cognitive Science of Natural Language Processing*, Dublin City University. Ver

N	substantivo	PCL	pronome clítico
NC	substantivo comum	POS	possessivo
NP	substantivo próprio	PI	indefinido
V	verbo	PD	demonstrativo
VINF	verbo no infinitivo	QU	pronome relativo / palavra-QU
VN	verbo no gerúndio ou no particípio presente	AD	artigo definido
VPP	verbo no particípio passado	AI	artigo indefinido
A	adjectivo	CARD	numeral cardinal
P	preposição	CARDR	numeral cardinal romano
ADV	advérbio	ORD	numeral ordinal
C	conjunção	I	interjeição
PES	pronome pessoal	[[ ]]	amálgama de palavras Ex. [[o]]A=_P=O_AD

São já significativos os trabalhos realizados com base nos textos do CIPM, envolvendo alunos da Universidade Nova de Lisboa, tanto no âmbito de projectos como a nível de licenciatura e de pós-graduação. Destacam-se os seguintes:

- sintaxe dos clíticos; construções transitivas e inacusativas – M. Alexandra Fiéis
- complementação infinitiva – M. Cristina V. da Silva
- orações adjuntas – Maria Lobo
- construções de posse – Ana Castro
- advérbios e movimento – Fátima Martins

- terminologia – Ana Castro e Ana C. Reis
- etimologias – Sandro Dias
- marcas morfológicas arcaizantes vs. modernizantes – João Loureiro

- etiquetagem morfossintáctica automática – Nuno Marques
- análise sintáctica automática – Vitor Rocio
- extracção automática de unidades multipalavra – Joaquim F. da Silva

Havia a consciência, desde há muito, de que a melhor maneira de tornar os textos acessíveis a todos passava, necessariamente, pela sua disponibilização na Internet. Uma grande parte dos textos do CIPM, bem como dois módulos de um Dicionário – o *Dicionário de Verbos do Português Medieval* (DVPM) *e o Dicioná-*

também Xavier, M.F.; M.G. Vicente; M.C. Silva (1997) "Aplicações de um Etiquetador Morfossintáctico a Textos Portugueses Medievais", Workshop "Taggers para o Português", I.L.T.E.C.

rio de Terminologia Portuguesa Medieval (Jurix) –, apesar de estarem ainda em construção, encontram-se já publicados em http://cipm.fcsh.unl.pt/.

O projecto visa a disponibilização on-line dos textos do CIPM, em versões com e sem anotações, bem como de um Dicionário, de Glossários e de estudos que tenham por base dados daqueles textos, ou de outros, sobre o Português Medieval. No sentido de melhorar e aumentar o corpus textual e os dicionários e de optimizar o acesso à informação disponibilizada, estão a ser desenvolvidas e melhoradas, em colaboração com investigadores e estudantes de diferentes áreas, nomeadamente Linguística, História, Cultura, Literatura e Informática, as funcionalidades de consulta, a visualização gráfica e as ferramentas computacionais para o tratamento automático dos sub*corpora* seleccionados para estudos específicos.

Finalmente, o projecto procura também levar os professores e os alunos de Linguística Histórica e de História da Língua Portuguesa a trabalharem, em aula ou isoladamente, a informação que se disponibiliza na Rede. Este é um objectivo que conduzirá, certamente, ao efeito desejado – o desenvolvimento do conhecimento e o aprofundamento daqueles domínios. De facto, algumas experiências de propostas de utilização da informação disponibilizada na Internet, quer do CIPM quer de outras páginas, tem motivado claramente os alunos da FCSH-UNL, levando-os a uma maior participação activa na aquisição e desenvolvimento dos conhecimentos, especialmente em História da Língua Portuguesa.

TEXTOS DO CIPM[6]

Século 12

CHP – Textos Notariais, MARTINS (1994)
DN – Textos Notariais, MARTINS (2000)

Século 13

CA – Documentos Portugueses da Chancelaria de D. Afonso III, DUARTE (1986)
CEM – Cantigas de Escárnio e Maldizer, LOPES (2002)*
CHP – Textos Notariais, MARTINS (1994)
CS – Dos Costumes de Santarém, RODRIGUES (1992)
DN – Textos Notariais, MARTINS (2000)
FG – Foros de Garvão, GARVÃO (1992)
FR – Foro Real, FERREIRA (1987)
HGP – Textos Notariais da Galiza e do Noroeste de Portugal, MAIA (1986)
NT – Notícia de Torto, CINTRA (1990)
TL e **TT – Testamento de D. Afonso II** (ms. de Lisboa e de Toledo), COSTA (1979)
TOX – Textos Notariais, PARKINSON (1976-1978)*
TP – Tempos dos Preitos, FERREIRA (1986)

Séculos 13/14

CEM – Cantigas de Escárnio e Maldizer, LOPES (2002)*
VS – Vidas de Santos de um Manuscrito Alcobacense (cópias do século XV), CASTRO (1985)

Século 14

CAXL (ms. **L**) e **CAXP** (ms. **P**) – **Crónica de Afonso X**, CINTRA (1951)
CEM – Cantigas de Escárnio e Maldizer, LOPES (2002)*
CGE – Crónica Geral de Espanha de 1344, CINTRA (1951)
CHP – Textos Notariais, MARTINS (1994)
CS – Dos Costumes de Santarém, RODRIGUES (1992)
DN – Textos Notariais, MARTINS (2000)
DSG – Demanda do Santo Graal, NUNES (2001)*
FG – Foros de Garvão, GARVÃO (1992)
HGP – Textos Notariais da Galiza e do Noroeste de Portugal, MAIA (1986)
PP – Primeira Partida, FERREIRA (1980)
TOX – Textos Notariais, PARKINSON (1996-1998)*

[6] O asterisco indica aqueles que ainda não estão disponibilizados *on-line*.

Século XV

CD – Chancelarias Portuguesas: D. Duarte, DIAS (1998, ...)*
CHP – Textos Notariais, MARTINS (1994)
CP – Castelo Perigoso, NETO (1997)
DN – Textos Notariais, MARTINS (2000)
HGP – Textos Notariais da Galiza e do Noroeste de Portugal, MAIA (1986)
HRP – História dos Reis de Portugal, CINTRA (1951)
LC – Leal Conselheiro, PIEL (1942)*
LE – Livro da Ensinança de Bem Cavalgar Toda Sela, PIEL (1944)*
LTV – O Livro das Tres Vertudes, CRISPIM (1995)*
OE – Orto do Esposo, MALER (1956)
ZPM – Crónica do Conde D. Pedro de Meneses, BROCARDO (1994)

Século XVI

CHP – Textos Notariais, MARTINS (1994)
CRB – Crónica dos Reis de Bisnaga, LOPES (1897)
DN – Textos Notariais, MARTINS (2000)
HGP – Textos Notariais, MAIA (1986)

Referências das fontes

BROCARDO, M. Teresa (1994) *Crónica do Conde D. Pedro de Meneses de Gomes Eanes de Zurara. Edição e estudo.* Dissertação de Doutoramento, Lisboa, F.C.S.H.

CASTRO, Ivo *et alii* (eds.) (1985) *Vidas de Santos de um Manuscrito Alcobacense* (Cod. Alc. CCLXVI / ANTT 2274). Lisboa, I.N.I.C.

CINTRA, Luís Filipe Lindley (1990) "Sobre o mais Antigo Texto Português", *Boletim de Filologia*, vol. XXXI.

CINTRA, Luís Filipe Lindley (1951) *Crónica Geral de Espanha de 1344. Edição crítica.* Lisboa, I.N.C.M.

COSTA, Pe. Avelino Jesus da (1979) "Os mais Antigos Documentos Escritos em Português", *Revista Portuguesa de História*, 17.

CRISPIM, M. de Lourdes (1995) *Christine de Pizan. O Livro das Tres Vertudes* (edição digitalizada).

DIAS, João Alves (ed.) (1998,...) *Chancelarias Portuguesas: D. Duarte.* Lisboa, Centro de Estudos Históricos da U.N.L.

DUARTE, Luiz Fagundes (1986) *Os Documentos em Português da Chancelaria de D. Afonso III (Edição)*, Dissertação de Mestrado, F.L.U.L.

FERREIRA, José de Azevedo (1980) *Alphonse X, Primeyra Partida. Edition et Etude.* Braga, I.N.I.C.

FERREIRA, José de Azevedo (1987) *Afonso X, Foro Real. Vol. I. Edição e Estudo Linguístico.* Lisboa, I.N.I.C.

FERREIRA, José de Azevedo (ed.) "Tempo dos Preitos" in ROUDIL, Jean (1986) *Summa de los Neuve Tiempos de los Pleitos. Édition et étude d'une variation sur un thème*. Paris, Klincksieck.

GARVÃO, M. Helena (1992) *Foros de Garvão. Edição e Estudo Linguístico*. Dissertação de Mestrado, Lisboa, F.L.U.L.

LOPES, David (ed.) (1897) *Chronica dos Reis de Bisnaga*. Lisboa, Imprensa Nacional.

LOPES, Graça Videira (2002) *Cantigas de Escárnio e Maldizer dos Trovadores e Jograis Galego-Portugueses. Edição*. Lisboa, Estampa.

MAIA, Clarinda de Azevedo (1986) *História do Galego-Português*. Coimbra, I.N.I.C.

MALER, Bertil (ed.) (1956), *Orto do Esposo*. Rio de Janeiro, Ministério da Educação e Cultura, Instituto Nacional do Livro.

MARTINS, Ana M. (1994) *Clíticos na História do Português – Apêndice Documental – Documentos notariais dos séculos XIII a XVI do Arquivo Nacional da Torre do Tombo*, Dissertação de Doutoramento, vol. 2, Lisboa, F.L.U.L.

MARTINS, Ana M. (2000) *Documentos Notariais dos Séculos XII a XVI* (edição digitalizada).

NETO, João A. Santana (1997) *Duas Leituras do Tratado Ascético-Místico Castelo Perigos*. Dissertação de Doutoramento, São Paulo, Faculdade de Filosofia, Letras e Ciências Humanas, U.S.P.

NUNES, Irene (2001) *A Demanda do Santo Graal* (edição digitalizada).

PARKINSON, Stephen *Arquivo de Textos Notariais em Português Antigo*. Oxford (edição digitalizada).

PIEL, Joseph M. (ed.) (1942) *Leal Conselheiro*. Lisboa, Bertrand (com emendas de João Dionísio).

PIEL, Joseph M. (1944) *Livro da Ensinança de Bem Cavalgar Toda Sela. Edição crítica*. Lisboa, Bertrand (com emendas de João Dionísio).

RODRIGUES, M. Celeste Matias (1992) *Dos Costumes de Santarém*. Dissertação de Mestrado, Lisboa, F.L.U.L.