

1. Data de 1949 a primeira aplicação da informática ao tratamento de textos, com a decisão do P. Roberto Busa de constituir um Index Thomisticus recorrendo à mecanografia e ao computador. A partir de então, um pouco por toda a parte, foram surgindo projectos e realizações semelhantes; raro será o país (1) ou a língua de cultura (2) que não conheçam experiências deste tipo. O avanço tecnológico dos últimos anos, quer na concepção dos computadores (aumento da capacidade de memória, particularmente) quer nas linguagens de programação, na entrada de dados, no recurso a memórias auxiliares, na forma de exploração, de obtenção de resultados, etc., favoreceu essas realizações ao mesmo tempo que iam sendo reconhecidas as vantagens de aliar a informática às ciências humanas. O domínio das ciências da linguagem, de resto, pode considerar-se privilegiado já que lhe pertencem trabalhos tão audaciosos e marcantes como os que conduziram à tradução automática ou às Estruturas sintácticas de N. Chomsky, os quais só por si definem um período de investigação. Se desencantamento houve da parte de alguns sectores ao verem abandonadas as experiências que haviam estado na origem de tais realizações, não deixará de ser significativo que a consciência das limitações da máquina, pelo lado dos cientistas que dela se serviam, e o ruir do mito do computador máquina-milagre, pelo lado dos que alguma vez haviam sonhado com um substituto do cérebro humano, contribuíram para um mais exacto reconhecimento da alteridade humana e do que de específico apresentam as suas manifestações perante a robótica. Não será demais recordar que a distância entre uma e outra fica claramente determinada na possibilidade ou impossibilidade de reconhecer e distinguir, numa sequência, linguística ou outra, o que é significativo do que não é significativo, de apreender uma situação na sua globalidade concreta e determinada, de julgar de uma activida-

de simbólica e simbolizante, de aceitar o implícito como forma de comunicação. No domínio linguístico, não era apenas a criatividade humana que ficava a constituir resíduo inacessível à actividade repetitiva da máquina; se é possível formalizar tipos de frase, por ex., a sua realização implica uma transformação situada e não previsível e como tal não transmissível à máquina em algoritmo previamente estabelecido.

O acentuar das diferenças trazia como resultante o reconhecimento da necessidade da intervenção humana a completar o trabalho mecânico na análise dos diversos actos de comunicação linguística, sejam eles de índole prática ou de natureza simbólica.

A informática surgia assim na sua verdadeira função de subordinada e não de substituto do homem. Filólogos e documentalistas (por razões bem diferentes e bem mais modestamente do que a investigação anterior - teórica ou prática - quer na disponibilidade de meios técnicos quer na definição imediata de objectivos) assegurariam a conciliação do automatismo e da análise singular, num trabalho interdisciplinar notável e com resultados que representam para o nosso tempo qualquer coisa de semelhante ao aparecimento de instrumentos de trabalho na época das universidades medievais. (3) O progresso tecnológico das várias gerações (o termo é metafórico, mas recobre uma realidade bem conhecida) de computadores pôs, de resto, nas suas mãos um leque considerável de possibilidades.

Difícil, se não impossível, fazer um balanço (por imfeito que fosse) dos trabalhos já realizados ou em curso. Na verdade, muitos deles, no todo ou em parte, não vêm a público, e nem todos se inscrevem estritamente no âmbito de investigação particular dos centros onde tiveram origem. (4)

Embora a metodologia de base seja idêntica ou semelhante, as aplicações concretas variam de caso para caso, segundo a diversidade e o tipo de formação dos elementos que constituem as equipas de investigação, dos objectivos perseguidos, das disponibilidades técnicas ou financeiras, do grau de risco de inovação consentido em relação ao problema apresentado, etc..

Não é, por isso, nosso intuito ou pretensão referir tais experiências, mas tão somente chamar a atenção para o apoio que a informática pode oferecer no estabelecimento ou no avanço de uma investigação que pretende constituir-se e na criação de instrumentos de trabalho indispensáveis para a sua prossecução. (5)

2. Desfaçamos, no entanto, um equívoco, ou antes, um mito. Nem tudo é possível em computador. Deixámo-lo já implícito, mas importa acentuá-lo para compreensão da metodologia adequada.

Como máquina, o computador não tem intuição nem conhecimento do mundo exterior nem acesso a dados implícitos. R. MOREAU resumia recentemente a sua capacidade: "L'ordinateur peut simuler certaines activités intelligentes, il ne s'agit que d'un nombre restreint d'activités, celles de nature algorithmique". (6) A sua performance fica condicionada pela transmissão correcta de um algoritmo, ou seja, uma série encadeada e finita de regras ou instruções que traduzidas em linguagem de máquina constituem o que se designa por programa.

Todavia a competance e a utilização do computador não se limita a executar um número finito de instruções e a fornecer os resultados finais. "Grâce à l'interactivité, continue R. MOREAU, et à son complément indispensable la virtualité, la façon de poser et résoudre un problème a, en effet, complètement changé. Elle a changé, tout d'abord, dans la mise au point des algorithmes et de leur programmation. En effet, il est maintenant possible d'écrire des programmes en conversant en quelque sorte avec la machine, celle-ci signalant immédiatement certaines erreurs de logique".

É assim que o primeiro contributo do computador (todos os praticantes da informática ligada às ciências humanas são unânimes em apontá-lo) é de ordem metodológica. Impondo restrições, obriga o investigador de ciências humanas (de formação ordinariamente enciclopédica e de tendências universalizantes) a definir e analisar ou desdobrar as questões, à boa maneira platónica. (7)

Uma certa ascese científica se impõe. ALBERT KIENTZ aponta como regras de base para uma análise em computador : 1) ser objectiva; 2) ser sistemática; 3) fixar-se no conteúdo manifesto; 4) quantificar. (8)

Sublinhe-se nesta enumeração que ser objectivo significa a utilização de critérios interpretáveis univocamente por qualquer investigador e em qualquer momento de trabalho (da fase de entrada de dados à de obtenção de resultados). Restrição óbvia, sem a qual toda e qualquer tentativa de resultado será infrutífera, mas que pressupõe a definição de conceitos e a delimitação de problemas na sua globalidade sem nada omitir. A fixação no conteúdo manifesto constitui uma regra não só prudencial, que evite a subjectividade, mas contrapõe também o critério descritivo ao critério interpretativo; a comutabilidade do implícito para o explícito (e a codificação algorítmica é justamente a passagem de um ao outro) só pode conhecer a equivalência e nunca a aproximação.

Daqui que a análise deva ser gradativa. Como meio incluído para apreciar o alcance da análise. O mesmo autor propõe como etapas: 1) definir os objectivos da investigação ; 2) constituir um corpus; 3) decompor o corpus em unidades; 4) reagrupar as unidades em categorias; 5) tratar quantitativamente.

Definir os objectivos da investigação não quer dizer que o computador sirva apenas para apressar a resolução de um problema cuja solução já se encontrou de outro modo mais demorado. Além da aplicação em actos repetitivos ou de solução morosa, o computador deverá ter um lugar importante na verificação de hipóteses de trabalho, onde, como já ficou assinalado, a própria correcção formal é verificada. Só uma concepção demasiado estreita poderia pôr em causa um tipo de investigação desinteressada e anular os aspectos lúdicos e imprevisíveis que a ciência comporta. A heurística pode encontrar também na máquina um apoio importante.

3. Se o computador não faz ciência, ele é hoje um mecanismo importante que ajuda a fazer ciência, comandando a própria investigação.

Para descermos ao plano concreto do tratamento de textos, assinalaremos que a investigação neste domínio tem sabido aproveitar aspectos tão importantes como: a) capacidade de constituição de ficheiros volumosos; b) acessibilidade quer na constituição quer na exploração (ou actualização e substituição) de tais ficheiros; c) complexificação de ficheiros pela intercorrelação; d) rapidez na obtenção de resultados finais.

Sem pretendermos comentar cada um destes aspectos (outros poderiam ser apontados), salientemos que um elemento fundamental é a constituição inicial de um ficheiro. Dela dependem todas as explorações futuras e por isso é importante tanto a escolha do suporte como a definição das unidades de base (unidade de análise ou unidade de registo, zonas de trabalho, etc.) ou o simples tipo de referenciação. No que a este se refere, pensar, por ex., que 28 posições de referenciação num registo do CETEDOC é um excesso dispensável equivalente provavelmente a não ter uma noção exacta do que é uma página usual de texto com as respectivas unidades componentes, mas suprimir tal luxo terá certamente graves consequências na manipulação futura de qualquer registo, ainda que seja apenas para corrigir uma posição do ficheiro (que mais não seja a substituição de erro de caracter/grafema numa forma de texto). A escolha de suporte condiciona igualmente a manipulação da informação. Corrigir um ficheiro em cartões perfurados é mais fácil que fazê-lo quando o suporte é constituído por fita perfurada, e trabalhar ou simplesmente transportar os cartões do ficheiro de Plauto, por ex., ainda que fosse apenas relativo a finais de frase, é algo de imcomportável.

Quer isto dizer que toda e qualquer opção tem consequências directas, mais ou menos imediatas, em tratamentos subsequentes. A frequência dos Centros de Investigação torna-se assim importante e quase indispensável, a fim de evitar erros cuja correcção posterior poderá equivaler à anulação de todo ou de parte do trabalho já efectuado, ou a fim de usufruir de uma prática convenientemente testada cujo desconhecimento acarretaria, pelo menos, uma perda de tempo e

gastos desnecessários.

Se dos aspectos enunciados nos limitamos ao primeiro, o da constituição do ficheiro, é porque daqui depende toda uma pedagogia de trabalho que é menos evidente nos restantes, alguns dos quais de natureza mais técnica que não pretendemos abordar.

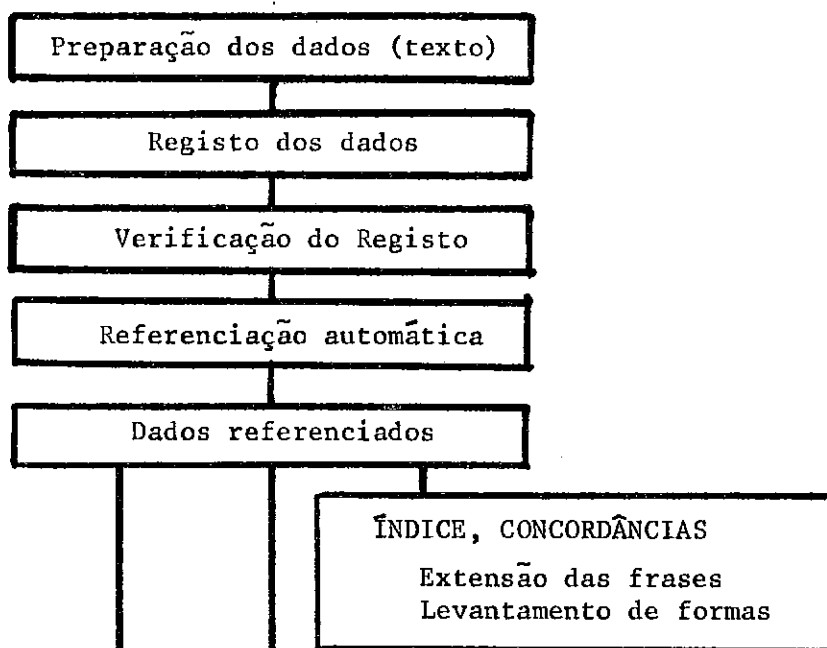
Complexo seria demorarmo-nos a descrever todas as fases de trabalho que envolve a preparação e análise de um texto por tratamento em computador. O organigrama seguido habitualmente no CETEDOC de Lovaina será mais intuitivo. (9)

Como se vê por tal organigrama (cf. página seguinte), duas fases distintas são tidas em consideração, segundo a intervenção ou não intervenção de um trabalho de análise sobre os dados do texto.

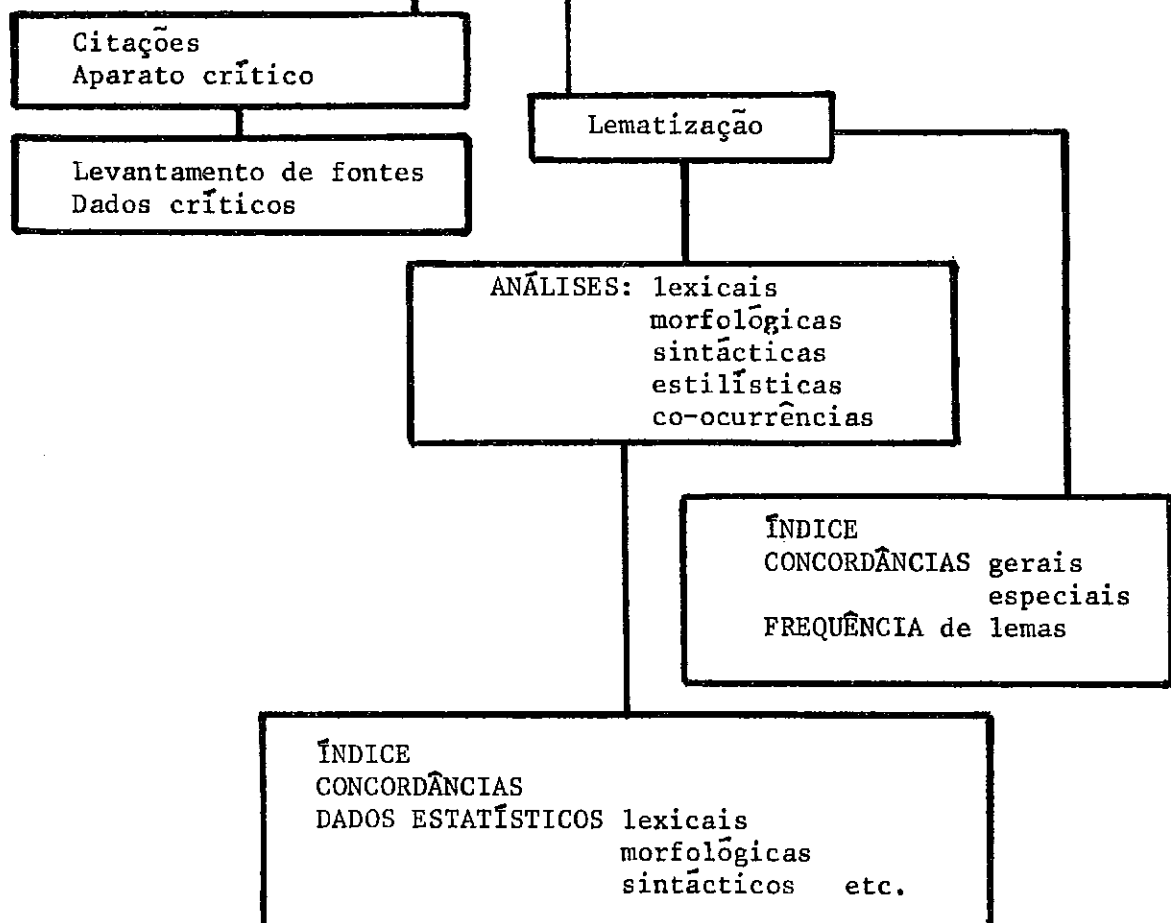
Escolhido um texto determinado (pressupõe-se que a qualidade de edição é um problema previamente e definitivamente resolvido) e introduzido de forma contínua em suporte adequado, com os códigos aconselháveis para a referência (normalmente, códigos correspondentes a fim de palavra, frase, linha, página, livro/capítulo, obra) esta é atribuída automaticamente (mediante programa) a cada uma das unidades de base de tratamento, a palavra/forma do texto. (Para o efeito, considera-se como palavra aquilo que no texto é compreendido entre dois brancos). Deste modo, em qualquer momento de tratamentos futuros, poderemos situar cada palavra no seu contexto exacto, pois que possuímos toda a descrição da sua posição na obra, no capítulo/livro, na página, linha e frase.

A partir daqui é possível obter já todo um conjunto de informações importantes tais como: índice alfabético (directo ou inverso) das formas, concordâncias de contexto (contexto frase, contexto optimizado de x posições, etc.), dados estatísticos relativos à frequência de formas, extensão de frases, relação entre extensão de frase e capítulo/livro/obra/autor, relação entre frequência de vogais e de consoantes, etc.. Para isso nada mais é necessário que explorar o código de representação da imagem do texto. De notar que quanto à pontuação ordinariamente se tem apenas em conta a pontuação forte, o ponto final; razões específicas poderão levar a man-

1.a FASE: TRATAMENTO DE DADOS BRUTOS (=NÃO ANALISADOS)



2.a FASE: TRATAMENTO
DE DADOS ANALISADOS



ter a restante pontuação, mas nesse caso e a fim de evitar inconvenientes futuros, como fosse a disparidade de tratamento de formas com pontuação colada e formas sem pontuação, haveria que prever uma operação suplementar de limpeza das formas e constituição de novo ficheiro.

A segunda fase de tratamento assenta sobre um trabalho de análise minuciosa e sistemática (já que a ela é submetida cada uma das formas do texto) e mais ou menos aprofundada conforme os objectivos pretendidos e a grelha estabelecida para isso. Vários tipos de análise são possíveis e alguns deles são hoje praticados, desde a pesquisa de fontes, análise textual por comparação de manuscritos, análises gramaticais, estilísticas, etc..

Como metodologia de trabalho, convirá apontar que uma das grandes vantagens do computador está em que nem tudo precisa de ser feito de todas as vezes pelo investigador. Todos os actos repetitivos, de certa frequência pelo menos, podem ser confiados à máquina. Com a vantagem de não ficarem sujeitos à fadiga ou a factores externos causadores de desatenção e divergência de análise. Considere-se, por ex., o problema dos invariáveis de uma língua. Uma vez estabelecida uma análise de aplicação constante, pode esta ser-lhes atribuída automaticamente, sem que o investigador tenha de intervir em cada ocorrência.

Um outro exemplo tomado da lematização. Advirta-se que se considera como lema a forma de base ou de referência, i.é, aquela que dá entrada no dicionário de uma língua: singular para o substantivo, masculino do singular para o adjectivo, 1.ª pessoa sing. pres. ind. / infinitivo (alternativa conforme se trata de línguas clássicas ou de línguas modernas) para o verbo, etc.. Suponhamos que no primeiro levantamento de formas obtivemos uma frequência de 50 para determinada forma. Bastará que façamos a correspondência entre esta forma e o lema respectivo e que este seja atribuído (mediante programa) sempre que a forma ocorra para que o trabalho de lematização fique, no caso, realizado. Suponhamos ainda que decidíamos que todas as formas terminadas em -s precedido de vogal a, e, o teriam como lema a mesma forma com supressão

de -s. Numa língua como o português, teríamos sem dúvida ganhado uma percentagem elevada de lematizações na classe dos substantivos. É evidente que enquanto não tivermos definido os tipos de automatismos da língua nunca será possível dispensar a intervenção humana. De resto, as ambiguidades são extremamente abundantes em qualquer língua e só o recurso ao plano sintagmático do discurso permite desfazê-las — restrição insuperável para a máquina, como demonstraram as experiências da tradução automática. A intervenção humana, no entanto, pode ser suavizada, já que nunca poderá ser dispensada, pela exploração dos automatismos da língua e pelo reconhecimento efectivo das suas formas pelo computador. O trabalho a realizar neste sector variará, em grande parte, segundo os objectivos visados, a natureza da língua em estudo, a metodologia praticada ou o grau de rentabilidade procurado. Para referirmos apenas duas experiências formadas sobre o tratamento de textos latinos, salientaremos que nem sempre os métodos mais sofisticados são os mais rentáveis. O Centro da Universidade de Liège, L.A.S.L.A., utiliza uma metodologia complexa, com a decomposição da palavra em elementos, desinências e radicais, a cada um dos quais é atribuído um código de análise. (Desinências e radicais, aliás, são apenas designações operatórias e não coincidentes necessariamente com as noções e realidades linguísticas dos mesmos) (10). No plano prático, todavia, o trabalho do investigador é multiplicado já que para uma forma terá de verificar todas as hipóteses de análise consentidas pela ambiguidade ou pluralidade de forma e de contexto.

O CETEDOC da Universidade Católica de Lovaina optou por uma modalidade mais simples. Todo e qualquer texto tratado passa a constituir um ficheiro de base confrontável e aplicável a textos subsequentemente analisados. Para sermos concretos, apontaremos o que nós mesmos aí pudemos realizar nos meses do verão passado. Começámos por tratar a Vida de S. Frutuoso de Braga, para o que tomámos a edição crítica estabelecida pelo Prof. Manuel C. Díaz y Díaz. Uma vez referenciado o texto (de harmonia com os critérios apontados acima), recebeu ele uma lematização automática, por aplicação

do leemário do CETEDOC (leemário, i. é, índice de lemas a que estão associadas as formas dos textos já trabalhados e existentes em biblioteca). Sempre que uma forma encontrava equivalente recebia o lema correspondente. Não tínhamos então mais a fazer seguidamente que rever as ambiguidades (por ex., distinção entre cum preposição e conjunção, etc.) e completar as lacunas. Uma vez analisado o texto (não recorremos aqui à análise automática do Centro, pois que preferimos seguir uma grelha particular) ficou ele (ainda que com as restrições julgadas oportunas) a constituir um ficheiro de base aplicável a tratamentos subsequentes.

Para que se faça uma ideia da rentabilidade deste método (prático e que não exige as complexificações de outros), bastará apontar que, na análise de textos filosóficos medievais, se obteve uma percentagem de 70% de palavras tratadas automaticamente (11).

4. Quer-nos parecer que, com o que acabamos de referir, é fácil deduzir ou simplesmente reconhecer as perspectivas abertas ao tratamento de textos com o recurso ao computador. Pense-se apenas no que ele pode representar na constituição de dicionários de autores ou de língua (12). Em lugar de acumular verbetes, com atestações tantas vezes incompletas e segundo critérios muito provavelmente divergentes pela intervenção de redactores diferentes, é bem mais objectivo partir da constituição de listas de concordâncias vocabulares obtidas a partir do registo do texto em computador, onde não apenas o nome do autor, da sua obra, mas igualmente o da data de publicação ou de edição, a página, etc. poderão ficar assinalados. Não só o seu manejo é mais rápido, mas estaremos seguros de possuir uma imagem fiel e completa de todas as atestações. A análise subsequente fica assim facilitada, ao mesmo tempo que fica anulado o risco de alteração do texto por manipulações sucessivas do ficheiro e por intervenções diferentes. A própria selecção de atestações não necessitará de recorrer à cópia manual se houver o cuidado de prever um código conveniente e se se estabelecer um programa particular para o efeito.

Não temos prática do que possa significar o tratamento de texto para o estabelecimento de uma edição crítica. Todavia não será difícil de admitir a priori que um índice global ou uma concordância do mesmo tipo podem valer mais que todas as conjecturas eruditas na reconstituição de um original (13).

Todo e qualquer aspecto de descrição de uma língua, diacrónico ou sincrónico, formal ou semântico, poderá obter uma resposta, mais ou menos larga, a partir das possibilidades oferecidas pelo computador. O problema está em saber interrogá-lo, sem dúvida. Mas, ainda que somente nas modalidades enunciadas, os resultados são já volumosos e podem considerar-se significativos.

Apenas podemos pronunciar-nos sobre a aplicação que fizemos sobre 13 textos latino-medievais portugueses (intencionalmente diversificados no tempo, ainda que não tanto diversificados quanto ao género literário como pensáramos inicialmente) e um glossário de verbos latino-português medieval do fundo de Alcobaça. Com este conjunto pensamos chamar a atenção para um domínio limiarmente ignorado, apesar do interesse que o latim medieval tem despertado desde há 50 anos pelo menos nos meios universitários europeus e norte-americanos. Julgámos que recorrendo à informática poderíamos, de algum modo, recuperar o atraso que nos separa da erudição europeia quanto ao estudo deste sector ou mesmo quanto à constituição de um léxico nacional de latim medieval.

Não pretendemos, no entanto, supor que será este o domínio onde seja mais urgente uma aplicação semelhante; não nos pertence definir hierarquias, mas também não cremos que haja problemas de concorrência. Apesar das restrições económicas em curso, talvez não seja descabido nem utópico pensar no tratamento dos textos dos nossos autores com a ajuda da informática... Que segredos nos revelariam Fernão Lopes, Camões, Mendes Pinto, Eça, Aquilino e tantos outros ? (14)

Para terminar, acrescentaremos que alguns dos programas que constituímos em Lisboa (15) poderão ser aproveitados para esse fim ou para outros, no todo ou em parte, com ou sem alteração.

NOTAS

1) Portugal continua, por enquanto, incluído no número de tais raridades. Contamos apresentar proximamente os primeiros resultados de uma investigação por nós começada sobre textos latino-medievais portugueses. Os programas de base poderão ser aproveitados para outro tipo de textos. Aliás, no número dos ficheiros tratados, no prosseguimento do nosso trabalho no Centre de Traitement Electronique des Documents (CETEDOC) da Universidade Católica de Lovaina incluímos um glossário medieval de verbos latino-português do fundo de Alcobaça.

2) O português não faz excepção já que tem sido objecto de estudos particulares em universidades estrangeiras. Gil Vicente, por ex., foi objecto de tratamento electrónico por parte de W.W. MOSELEY em Fort Collins, Colorado. Não seria demais pensar no tratamento lexicográfico (índices, concordâncias, frequências...) de outros autores nossos. Acrescentaremos que não pretendemos referir-nos aqui ao estabelecimento do português fundamental, cujo empreendimento se vem arrastando há anos e de cujos processos de trabalho não estamos suficientemente informados para emitir uma opinião. De resto, o corpus de tratamento não é estabelecido na base de textos completos, como aqui consideramos.

3) Pense-se apenas na diferença existente (seja de concepção seja de exploração) entre a constituição manual de um ficheiro para um dicionário (de autor ou de língua) e a sua organização em ficheiro electrónico.

4) Um centro como o CETEDOC da U.C. de Lovaina funciona como unidade de investigação (particularmente interessado desde a fundação em textos latino-medievais sob a orientação do seu Director, Prof. Paul Tombeur) e como unidade pedagógica e científica de apoio técnico e informático a investigadores.

Seja-nos permitido exprimir aqui todo o nosso reconhecimento mais sincero a toda a equipa do CETEDOC com quem nos foi dado trabalhar no maior espírito de colaboração interu-

niversitária, durante o ano lectivo transacto, ao mesmo tempo que frequentávamos o Institut d'Etudes Médiévales da Universidade Católica de Lovaina.

5) Para uma informação, ainda que não seja completa, dos centros de investigação de tratamento de textos por computador, poderá consultar-se Statistique et analyse linguistique (colloque de Strasbourg, Avril 1964), Paris, 1966, ou também JOSSE DE KOCK, Introducción a la lingüística automática en las lenguas románicas, Madrid, 1974. Sobre o CETEDOC de Lovaina aparecerá proximamente em Euphrosyne uma nota subordinada ao título "A informática ao serviço da filologia latina", onde se dá conta dos trabalhos publicados por este Centro.

6) R. MOREAU, "Informatique et résolution de problèmes", IBM/Informations, nº 76, págs. 5 ss.

7) É ainda R. MOREAU que relembra: "pour Platon, tout ce qui ne peut s'explicitier sous forme d'un enchaînement de règles et d'instructions n'est pas connaissance mais croyance".

8) ALBERT KIENZ, Pour analyser les media — l'analyse du contenu, Paris, 1971.

9) Cfr. para referência e desenvolvimento destes aspectos PAUL TOMBEUR — ANDRÉ STAINIER, "Les méthodes et les travaux du Centre de Traitement Electronique des Documents", Bulletin de Philosophie Médiévale, X-XII (1968-70), pp. 141 ss.

10) A. BODSON — E. EVRARD, "Le programme d'analyse automatique du latin", Revue, 1966, nº 2, pp. 17 ss.

11) É evidente que uma análise com as variáveis de uma análise estilística não oferece grau apreciável de automatização.

12) O Trésor de la langue française, convirá recordá-lo, está a ser estabelecido em ficheiro electrónico, sob o patro-

cínio do C.N.R.S.

13) Na colação de manuscritos, o computador pode prestar relevantes serviços. São conhecidos os trabalhos de Dom J. FROGER (particularmente La critique des textes et son automatisation, Paris, 1968). Outros, como G.P. ZARRI, G. PHILLIPARD (este em tese de doutoramento apresentada recentemente à Universidade Católica de Lovaina e realizada com o concurso do CETEDOC) têm recorrido ao computador para resolver problemas de crítica textual e constituição de stemmata codicum.

14) Derivando para outro domínio, porque não impulsionar a constituição de um ficheiro nacional que nos desse a conhecer os fundos das nossas bibliotecas ou que nos revelasse os instrumentos de trabalho (coleções, revistas, etc.) existentes no país? No domínio da investigação (e oxalá as ciências humanas, para não dizer as letras simplesmente, não venham a ser preteridas em favor de qualquer projecto míope) continuar a viver de ficheiros privados é pelo menos manter aferrolhada uma riqueza que, na expressão de Chesterton, só produz quando espalhada. E, além disso, não somos tão ricos que cada qual se possa dar ao luxo de ter o seu solar...

15) E que, de resto, pagamos a firma comercial inteiramente à nossa custa, sem que até hoje tenhamos obtido resposta definitiva das entidades a quem solicitamos apoio (o decoro obriga-nos a omitir referências).

AIRES AUGUSTO NASCIMENTO