TRATAMENTO DE TEXTOS EM COMPUTADOR

A LEMATIZAÇÃO

A exploração lexical foi a primeira a utilizar e a beneficiar da aplicação do computador ao tratamento de textos. Por um lado, a segmentação de unidades no código escrito facilitava a constituição de dados e a sua comparação; por outro, a investigação encontrava um meio adequado para manipular ficheiros volumosos sem ter que proceder a exclusões arbitrárias. Se razões pragmáticas levavam ao aproveitamento das possibilidades abertas pela automatização ou se o índice quantitativo de dados trabalhados e acessíveis e o grau de fiabilidade no tratamento garantiam um índice correlativo de rentabilidade de exploração, as exigências de ordem metodológica criavam igualmente um suplemento de rigor crítico e científico logo que se ultrapassasse o nível de levantamento das formas delimitadas pelas convenções da escrita para chegar às unidades de língua. Não apenas a forma de exploração questionava o grau de pertinência informativa das opções tomadas ao longo do tratamento, mas era possível ainda avaliar da aceitabilidade e carácter operatório de critérios de decisão utilizados bem como da consistência mantida na codificação de análise. Favorecendo e impondo explicitação de operações e gradatividade de complexificação como base de acesso a um aumento de informação, a nova metodologia assegurava igualmente a qualidade da investigação. Resolvendo limitações anteriores, abria simultaneamente novas vias para uma descrição e enumeração exaustiva das unidades de língua presentes num texto.

Vários são os problemas, de ordem teórica e prática, implícitos nessa operação, designada usualmente por lematização. Ao apontá-los aqui, não esconderemos o enquadramento particular em que tivemos de os solucionar. Não existe, aliás, norma comum aceite como tal, ou universalmente adequada. Julgamos, contudo, que, para além das coordenadas de situação, se reconhecerá uma metodologia cuja aplicação se não confina às fronteiras reduzidas da exemplificação apresentada.

1. *Um problema de lexicografia*

«Lematizar consiste em atribuir a cada uma das palavras de um texto a forma correspondente encontrada no dicionário de referência»,[1] ou, por outras palavras, «consiste, por um lado, em reagrupar as formas heterográficas dum mesmo vocábulo, por outro lado, em separar as formas homógrafas que pertencem a vocábulos diferentes».[2]

Definida assim esta operação, poderia parecer que ela não revestiria senão aspectos práticos, mormente de escolha do(s) dicionário(s) de referência e da respectiva consulta. Na realidade, a exploração automática dos textos, a quantificação dos resultados, a criação rápida de ficheiros, e a possibilidade de comparação de dados e análises, obrigando a critérios rigorosos e uniformes como ponto de partida, e revelando a inconsistência de procedimentos díspares, traz a claro a incoerência reinante no domínio dos instrumentos de trabalho tradicionais.[3]

Uma solução imediata seria a de ir procedendo a correcções sucessivas, à medida que as deficiências fossem sendo notadas. Restaria, no entanto, a questão de fundo: que critério tomar como norma ? A pura extensão de um tratamento a casos similares ou paralelos ?

A lematização é fundamentalmente um problema de lexicologia que deve assentar numa certa teoria de língua, ou pelo menos sobre uma percepção clara e operatória das unidades de uma língua. Só a partir daí se poderão formar critérios que normalmente os dicionaristas não explicitam ou que não é possível extrair da sua prática assistemática. Sem pôr em questão a utilidade dos instrumentos tradicionais (dentro das suas limitações não deixam de ser os respositórios mais abundantes dos dados de uma língua e continuam a fornecer a sua descrição mais operatória e acessível e podem constituir uma norma de referência) haverá que secundá-la não apenas com o acrescentamento de dados, mas com a ordenação crítica dos mesmos, tirando partido do contributo oferecido pelas novas técnicas.

2. *A delimitação da unidade lexical*

O objectivo da lematização é enumerar as unidades da língua, sem ter em conta as suas variantes, quer automáticas quer livres quer gramaticais. Procura-se determinar a unidade de base ou unidade menos marcada, susceptível de ser integrada nas situações variáveis resultantes do discurso, da sequência fonética ou de uma escolha particular (directamente intencional ou não). Intenta-se delimitar o vocábulo como unidade do léxico[4] ou como unidade codificada[5] existente na língua, independentemente da sua utilização concreta nos actos da fala. Enquanto codificada ou lexicalizada, o utente é livre de a utilizar ou não, mas não pode modificá-la no seu corpo de significante, sem comprometer a comunicação. Enquanto forma livre mínima, tal como a define Bloomfield, ela constitui uma unidade a nível superior ao do morfema, não pode ser decomposta em outras formas livres significantes mais pequenas, e apresenta um grau de coesão interna, manifestada pelo acento e pela sua natureza não discreta relativamente a outros elementos exteriores. Situa-se, pois, como resultado da primeira condição apontada, num segundo nível de complexidade organizativa, o da determinação de um morfema por outro (mesmo que este seja o morfema zero, já que nesta situação fica criada uma oposição com outros morfemas de caracterização). Determina-se também, em virtude da segunda condição (a coesão interna) como um todo que ultrapassa convenções gráficas ou condicionalismos diacrónicos.

É sabido como, no decurso da história da língua, unidades se fundem para formar, em dado momento, um todo fonético. A fusão é tanto mais estreita quanto os utentes esquecem os elementos constitutivos anteriores, e a nova unidade entra no sistema de comutações da língua. Ela só é possível, todavia, porque esses elementos apareciam associados na fala, e, de tal associação, resultou uma unidade indissociável, codificada na língua e mantida por uma coesão interna, com fronteiras materializadas por um acento único. O ponto de chegada pode ser o amálgama, após fases de justaposição, ou de aglutinação, com graus intermédios (sujeitos muitas vezes aos hábitos da escrita), mas o funcionamento conjunto dos diferentes elementos determina a formação de uma unidade da língua, que poderá ser analisada ou decomposta a nível de determinação de constituintes, mas não segmentada a nível de utilização (fala). O desmembramento de "respublica", "cachorro-quente", "caminho de ferro", "mãe d'água", é tão impraticável como será o de "declaração", "incómodo", ou "Hotel Dieu", "tempestas", "scutagium", "hastiludium". Esta coesão interna pode, por vezes, ficar obscurecida ou comprometida por fenómenos de disjunção, cujo grau extremo será formado pela tmese. Na língua latina, conhecemos casos audaciosos como os de Énio: *cere saxo interemit brum*; ***Massili.... tenos***. Serão casos limite de dissociação de uma forma codificada na língua, mas a partir daí podemos concluir que não é absolutamente necessário, a nível de utilização, ter o corpo fonético (ou gráfico) de manter-se inalterado para se poder reconhecer uma unidade de língua. Não é raro encontrarmos em latim *ante... quam*, ***post... quam***, ou, ao inverso, formas como *plus... quam*, aglutinadas em *plusquam* e equivalendo a *ultra*. Se tal dissociação joga com o carácter discreto do significante, parece bem claro que, para perceber como unidades de língua os elementos assim dissociados, há que referi-los ao significante na sua globalidade. (Haveria que recordar outros exemplos, tais como *non... solum ... sed etiam*, ou simplesmente as formas compostas do verbo). O grau de coesão interna pode, pois, ser variável, mas ele torna-se indispensável como critério para delimitar o significante na sua articulação própria com o conjunto da

língua e com a sequência linear do discurso.

3. *Variantes lexicais*

A unidade lexical, como elemento de língua / sistema, apresenta na realização concreta do discurso um número mais ou menos finito de variantes, cujos tipos principais podem ser reduzidos a três:

a) variantes complementares alomórficas, automáticas (porque codificadas na língua) e obrigatórias: — tipo fonético e combinatório — *ab illo / a te; do* (*de* + *o*), *bel homme / beau garçon*;

b) variantes livres, não obrigatórias e não automáticas, opcionais, de acordo com o utente e a sua intenção de uso: — tipo fonético intensivo — *rever / re-ver; impleo / in-pleo*;

c) variantes complementares sintagmáticas, de distribuição na sequência frásica: — tipo morfo-sintáctico. Pertencem a este grupo as variantes que resultam da integração discursiva; assim, as marcas de oposição masc. / fem. / neutro (marca de género), de oposição sing. / pl. / dual (marca de número), ou todas as outras que afectam o modo, tempo, voz, pessoa, etc..

4. *A forma fundamental: o lema*

A lematização não pode deixar de ter em conta a natureza específica de cada tipo de variantes. Para o primeiro tipo, a entrada deverá ser constituída pela forma considerada fundamental. Dever-se-ia, no entanto, assinalar logo de seguida a(s) variante(s) derivada(s) da distribuição (o que supõe uma certa descrição da forma do significante segundo os contextos imediatos de associação fonética). Para o tipo de variantes livres, não haverá dúvida em tomar a forma não marcada. No terceiro tipo, a abstracção de integração discursiva levará a optar pela forma de base (não marcada) para lema. Não será fácil todavia um acordo unânime sobre o que há-de considerar-se como tal. A prática corrente considera que o singular e o masculino (adjectivo) constituem a forma de base para o nome (substantivo e adjectivo). Para o verbo, a opção varia segundo se trata de línguas modernas ou línguas clássicas. Enquanto naquelas se prefere o infinitivo como forma de entrada, nestas opta-se pela 1.a pessoa do presente do indicativo. Haverá razões para manter tal prática, dada a diferença existente, por ex., entre o sistema do verbo latino construído sobre os dois eixos

do "infectum" e do "perfectum" e uma vez que nele a forma de base e a menos marcada é certamente o "infectum", e, neste, a 1.a pes. do sing. do pres. do indicativo. Aliás, o estatuto do infinitivo, de natureza mista (forma nominal-verbal) e vária (marca de sufixos supostamente nominais para o "infectum", acrescida da característica específica do "perfectum", verbo auxiliar para a passiva no "perfectum" e futuro) não permite considerá-lo como forma de base, ou mesmo de referência. Não constitui porém surpresa total que uma das últimas e mais bem elaboradas realizações de lexicografia latino-medieval, preparada pela Academia Britânica,[6] tenha optado pelo infinitivo em lugar do "infectum". A opção justifica-se, certamente, dentro de uma percepção de latim medieval em que a antiga oposição significativa de "infectum / perfectum" acaba por ser anulada em proveito da aproximação das línguas românicas. Atente-se mais no aspecto tradicional ou mais na inovação operada, a opção será sem dúvida diferente. O objectivo perseguido, a metodologia adoptada (determinação dos estratos vocabulares — clássico ou não clássico — no texto; comparação do vocabulário utilizado por mais que um autor, ou estudo de fontes) ou apenas a operacionalidade de um grupo de trabalho (por recurso a uma obra de referência) será igualmente determinante.

Um problema de escolha põe-se (e a uma dimensão que abrange a quase totalidade dos dicionários conhecidos) de um modo particular para as formas supletivas, sobretudo para o verbo e para o grau de adjectivo e advérbio.

Não haverá hesitação quanto ao substantivo em reter duas formas, embora uma não represente mais que a variante de género: *veado / corça; cerf / biche; bos / vacca*. Estamos, na realidade, não ao nível de morfema, mas de formas livres e dentro de uma categoria funcional que não é estritamente gramatical, mas também semântica e que remete para o referente.

Para o verbo, o supletivismo não é mais que um condicionamento linguístico: aspecto de "perfectum" oposto ao de "infectum". A organização do discurso não depende do supletivismo de raízes (não era o caso dos substantivos, onde a forma supletiva segundo o género obriga à reescrita do grupo ou da frase).

Para os adjectivos e advérbios, no que respeita ao grau (comparativo e superlativo), poder-se-á encarar a questão sob diversos ângulos. Há evidentemente um ponto de vista prático que está relacionado com o tipo de informação obtido directamente a partir de um índice do texto, onde se registem as formas supletivas utilizadas. Mas um léxico não é um índice, e a elaboração deste pode não representar mais que uma fase preparatória daquele. Subsistem, porém, problemas teóricos, como o de saber se o grau de comparação traduz uma situação interna do discurso ou se tem antes um conteúdo semântico, se é uma variante intensiva ou se veicula uma referência exterior. Se há situação semântica, dever-se-iam

manter todas as variantes, como no caso do substantivo; se não há mais que situação linguística, haveria a reter apenas a forma de base.

Que o grau afecte a organização linguística, não pode deduzir-se imediatamente do morfema de grau, pois que tal marca pode apresentar apenas valor intensivo (*meliores, optimi homines*), mas a maior parte das vezes oferece um correlativo de comparação na sequência da frase. Estamos assim perante uma marca mista, intensiva e comummente com valor distribucional.

Nas formas de morfema preso (e destacável relativamente à forma de base, designada por positivo) reter-se-á apenas, como nos outros casos, a forma não marcada. Para as formas supletivas (*melius, optimus*, etc.), uma vez que não estamos dentro de uma situação que se possa definir simplesmente de organização linguística interna, será legítimo proceder como nos casos dos substantivos e admitir as formas supletivas para o lema. Apelar para um critério de analogia com o verbo, pelas razões atrás assinaladas, não pderá aqui considerar-se pertinente. Assim preferiríamos seguir a opção dos dicionários que mantêm (no todo ou em parte) as entradas das formas supletivas de grau, mas não as de verbo.

5. *Lema e classe gramatical*

As variantes anteriormente apontadas eram compreendidas sempre integradas na unidade de classe gramatical. Só uma segmentação formal sem referência a esta poderá aceitar uma organização de entradas como é proposta por Roberto Busa, s. j., para o *Index Thomisticus*: «Adverbs are attached to nouns or verbs whenever possible as a type of 'adverbial case', e. g. *suaviter* was lemmatized as an 'adverbial case' of *suavis, e*, and *diligenter* was lemmatized as an 'adverbial case' of the participle of the verb *diligo, ere*. They were not treated as separate lemmas as Forcellini does, for example».[7]

Aquele autor não aduz as razões da sua decisão. Não parece todavia que, quer de um ponto de vista linguístico teórico, quer de um ponto de vista de exploração automática dos textos (que é também o seu), quer ainda do de informação imediata a partir de levantamentos linguísticos (índices, concordâncias, dados quantitativos, etc.), haja qualquer vantagem em desprezar a natureza específica dos morfemas de classe e dos respectivos contextos de distribuição. Será mais fácil aceitar o critério de neutralização de lemas para o caso do adjectivo funcionando como substantivo, tanto mais que a fronteira paradigmática entre um e outro é relativamente fluida e se poderá deixar para uma fase de análise a determinação da natureza do adjectivo como forma presa ou como forma livre (substantivada). A atribuição de índices numéricos (à semelhança do que faz o Forcellini) dependerá do grau

de informação que se pretenda obter a partir da consulta a uma lista lematizada construída sobre um texto.

Tal uniformização não parece, porém, que se haja de manter no que respeita ao particípio verbal, cuja natureza mista obrigará a reparti-lo pelas formas de base (verbo ou adjectivo), segundo a função desempenhada.

6. *Problemas de ambiguidade*

Situações como estas, cuja definição ultrapassa o âmbito da forma, impõem uma consulta ao contexto. O problema da ambiguidade surge, todavia, numa amplitude maior, a nível de exploração lexical, dada muitas vezes a neutralização de morfemas verificada na forma da base.

Para o solucionar, recorre-se a índices numéricos colados à forma, através dos quais se passam a distinguir os lemas. Nem sempre, porém, a sua atribuição obedece a critérios uniformes. Forcellini, por ex., manteve certas constantes. Observa uma ordem de derivação (*subditus 1* = adj.; *subditus 2* = subst.); segue a ordem das declinações e conjugações (*subditus 2* = subst. 2.a decl.; *subditus 3* = subst. 4.a decl.; *dico 1* = inf. *dicare; dico 2* = = inf. *dicere; volo 1* = inf. *volare; volo 2* = inf. *velle*). Todavia é oscilante quanto à ordenação de classes: *aspergo, -onis* = 1; *aspergo, -is* = 2; mas *capio, -is* = 1; *capio, -onis* = 2.

Não seria difícil substituir, nestes casos, o índice numérico pelo morfema de caracterização. Ganhar-se-ia em informação e evitar-se-ia o recurso contínuo a listas de homógrafas, cujo acrescentamento se vai operando à medida que se vão reconhecendo novas homógrafas, e onde por vezes só o puro convencionalismo ditou a atribuição de índices (cf. os últimos casos citados). Aliás, o funcionamento do nome como adjectivo ou substantivo, ou a distinção entre homógrafos substantivos, ficaria resolvido definitivamente desta maneira. Apenas certas desvantagens práticas, numa fase de lematização manual poderá aconselhar o contrário.

Note-se que, na distinção de lemas, a prática consignada por um uso generalizado tem em conta apenas um critério formal, e, juntamente com este, um critério etimológico. Excluem-se, em princípio, os critérios semânticos e os critérios sintagmáticos. Esta limitação há-de considerar-se tanto mais restritiva quanto o que se pretende é a descrição de uma língua (de um estado ou da utilização). Manter, por ex., a mesma entrada para "populare" com sentidos tão antagónicos como o de "devastar", "despovoar" (ao uso clássico) e "povoar" (do uso medieval) significa não se dar conta ou não assinalar alterações evidentes ou pelo menos deixar confundidos os planos sincrónico e diacrónico da língua, e não forne-

cer os dados mínimos de descrição linguística que a lexicografia deve servir.[8]

Aceite-se, no entanto, numa fase intermédia de organização de ficheiros, a solução tradicional, já que a alteração iria privar-nos imediatamente da maior parte das obras em referência; reserve-se a sua revisão para uma fase ulterior em que a exploração de ficheiros largamente compreensivos e em sistema de concordância permitirá uma descrição global.

Como depõe Ch. Muller, "le depouillement lexical ne saurait tout enregistrer, et ne peut retenir que des distinctions très nettes. Le reste appartient à une autre phase de recherche. Aussi est-il sage, à ce stade, d'adopter une norme pratique, sans grandes prétentions scientifiques, ce qui conduit à la rattacher autant que possible à un ouvrage de référence. C'est aussi ce qui pousse à adopter, pour la délimitation du mot, une norme aussi analytique que possible (les regroupements étant remis à plus tard), et pour la délimitation du vocable, une norme très synthétique (les distinctions peuvent attendre).[9]

7. *Recurso a obras de referência*

Apesar de todas as objecções, por mais fundadas que elas sejam sob o ponto de vista linguístico, o critério mais operatório é o de seguir «a norma tradicional que é a dos dicionários; ou pelo menos de se afastar o mínimo possível, num número de casos bem definidos e fáceis de enumerar».[10]

Para o latim, na falta ainda de um léxico tão compreensivo como promete ser o *Thesaurus Linguae Latinae*, aquele que oferece melhores condições de trabalho é, sem dúvida, o Forcellini.[11] Para o latim medieval, encontrará ele os complementos necessários nos trabalhos especiais gerais ou regionais e particulares.[12]

8. *Automatização e lematização*

Pela rapidez dos processos (criando novos ficheiros a partir de um ficheiro fundamental, em curto espaço de tempo) e pelas exigências metodológicas (rigor e uniformidade de critérios), a informática põe à disposição do investigador um instrumento de trabalho altamente rentável e operatório. Pela fusão rápida e completa de ficheiros, permitirá, por ex., ir corrigindo progressivamente a disparidade de classificação encontrada nas obras de referência. Ajudará igualmente a ultrapassar o critério formal-etimológico e adoptar critérios formais paradigmáticos e sintagmáticos de ordenação de homógrafas, etc.. Para tratramento destas ainda, não terá pequena importância a possibilidade de estabelecer automaticamente

inter-relações do lema (vocábulo) com a análise (palavra inserida no texto e respectiva função) e obter índices de classificação uniformes e sistemáticos correspondentes às várias classes gramaticais. Além disso, permite reagrupar os diferentes elementos do vocábulo dispersos na sequência discursiva e tratados analiticamente numa fase inicial, recorrendo à recriação do respectivo contexto.[13]

Mas não é apenas na correcção ou alteração de critérios que a informática pode prestar um concurso valioso. O acrescentamento progressivo do ficheiro de base fica facilitado também, pois a automatização pode suprir todas as análises repetitivas, ou pelo menos aquelas que não ofereçam ambiguidade, quer directamente, quando a oposição de formas seja explícita, quer por recurso ao contexto, quando este se possa considerar suficiente para determinar a forma.

A lematização pode processar-se assim em três fases:

a) manual: ao lado de cada forma, recebida em lista de computador segundo uma ordem sequencial do texto, ou segundo uma ordenação alfabética ou ainda em ordenação de frequência decrescente, escreve-se o lema (vocábulo lexicalizado) respectivo que seguidamente é introduzido em suporte adequado na zona previamente determinada.

b) semi-automática: uma vez constituído um primeiro ficheiro lematizado, por selecção sobre uma primeira lista de formas ordenadas por ordem de frequência decrescente, ou como resultado de tratamento de textos anteriores, os lemas podem ser atribuídos automaticamente a todas as formas idênticas, por sistema de comparação. Para as homógrafas, escolher-se-á o índice convencional com probabilidade de maior frequência, e proceder-se-á posteriormente à verificação mediante recurso a concordâncias. Este método apresenta como vantagem não só o facto da simplicidade de processos, mas ainda o da rentabilidade sem recurso a uma programação pesada e também o aproveitamento progressivo do trabalho anterior. Nesta perspectiva, cada novo texto entra a fazer parte de um dicionário de base. Graças a ele, o computador, pela comparação entre a lista-memória e as formas do texto a analisar, atribui a cada forma encontrada o lema adequado que detecta em tal lista-memória.

c) automática: a decomposição da palavra nos seus elementos fundamentais e a atribuição de uma forma de base poderá ser em certa medida automatizável, mediante um certo conjunto de regras.[14] Resta, no entanto, saber até que ponto é operatório e rentável um sistema desta natureza, quando em fase semi-automática se chegam a obter 70% de lematizações sobre o total de um texto.[15] Aliás, quando na exploração de textos como os de latim medieval as variantes gráficas se multiplicam, a complexidade será extrema. Fora disso, o recurso ao contexto (ainda quando acessível à máquina, o que não pode ser feito sem

largas restrições) não é suficiente para resolver as ambiguidades, e pode acontecer que o investigador tenha de perder mais tempo (com uma sobrecarga psicológica mais elevada e com todas as consequências daí advenientes) perante uma consola a verificar hipóteses de solução propostas pela máquina, que a atribuir uma análise directa. A possibilidade teórica (com todas as restrições que supõe a criatividade linguística, a arbitrariedade do signo linguístico e a variabilidade da frase no discurso) nem sempre constituirá a melhor solução do ponto de vista prático.

Tendo em vista uma rentabilidade prática poderão adoptar-se ainda várias modalidades de trabalho.

Uma vez obtida uma ordenação sequencial ou alfabética do texto, procede-se à transcrição em computador do conteúdo da zona-forma para a zona-lema. Desta maneira, e ressalvados sempre os casos de ambiguidade, obtêm-se as lematizações dos invariáveis e também as de todos os variáveis cuja forma presente no texto seja idêntica à forma de base ou lexicalizada. Seguidamente proceder-se-á à eliminação de desinências. Poderá optar-se por um tratamento manual ou também recorrer à automatização, enquanto se considere suficiente o índice de não-ambiguidade. Esta segunda alternativa será viável em certos casos como o do plural em português: poderá recuperar-se um número relativamente elevado de lematizações para substantivos, adjectivos e pronomes desde que se estabeleça uma instrução de apagamento do -s final de palavra. Perderíamos, sem dúvida, alguns lemas, mas ganharíamos uma quantidade maior. Depois de uma operação como esta, haveria ainda que proceder à substituição de elementos residuais quer de desinências quer de formas supletivas e à indexação para eliminação de homografias e ambiguidades. Progressivamente, no entanto, fomos reduzindo os casos a tratar singularmente.

Uma outra hipótese de trabalho consistirá partir de uma lista de frequências de formas do texto e criar para os tipos mais frequentes o lema correspondente que seguidamente poderá ser atribuído e escrito automaticamente na zona respectiva.

Esta modalidade é inteiramente compatível com a anterior. Os casos pendentes poderão ser decididos posteriormente em fase de análise ou mediante concordância.

9. *Soluções a adoptar*

As soluções técnicas terão que ser escolhidas em função dos problemas concretos que podem envolver tanto os objectivos do investigador como o volume de dados a tratar ou as disponibilidades técnicas de acesso ao computador ou ainda a existência de programação adequada. Convirá recordar, no entanto, que as modalidades mais lentas raramente

são as mais económicas e as menos sujeitas a erro.

As soluções teóricas terão que obedecer também a um critério de funcionalidade. Se uma descrição de língua deve obedecer a uma percepção das unidades de base, um modelo particular de análise poderá não ser o mais adequado para um trabalho de grupo ou para servir a um público de formação heterogénea. A obliteração dos critérios mais frequentes na lexicografia implicaria a anulação ou pelo menos assinalável redução do índice informativo e descritivo que uma lematização traz consigo.

Dada a contingência e variação dos condicionamentos técnicos não nos alongaremos quanto a soluções neste domínio. E, porque o domínio da lexicografia é demasiado vasto, apontaremos apenas alguns aspectos decorrentes de uma experiência sobre textos do latim medieval, assente numa prática testada ao longo de alguns anos no CETEDOC de Lovaina. Aceita este a Forcellini como obra de referência principal, procedendo, no entanto, a algumas alterações:

1) modificam-se grafias de acordo com as melhores edições críticas e de forma a obter o maior grau de uniformidade: *cottidie, cum / quum, solacium, paenitentia*, etc..

2) reforçam-se os lemas com sublemas, por recurso a índices numéricos, quando se verifica mudança de categoria gramatical (adjectivo substantivado) ou do próprio F. figuram com indicação de emprego "absolute" (*ex. bonum*).

3) aumentam-se as entradas com formas marcadas por asterisco, quando elas estão atestadas apenas depois de 550 em textos de fora da Itália ou depois de 904 em textos procedentes de Itália.[16] Para o seu estabelecimento, utilizam-se os vários léxicos latinos já publicados ou em vias de publicação.[17] O acrescentamento realiza-se também de acordo com os novos registos do texto em análise, sempre que a forma em questão não represente uma variante gráfica, mas uma unidade de língua.

Em qualquer hipótese, de alteração ou de inovação estabelecem-se critérios a seguir:

1) atende-se à forma e não à semântica. Assim não é marcado com asterisco *oculosus*, com o sentido de prudente, pois F. apresenta-o, se bem que com outra acepção: *lapis oculosus* = pedra preciosa. Não o será igualmente *populare* na acepção de povoar.

2) a etimologia resolve os casos de homografia: *iterare*, de *iterum, iter* ou *ita*.

3) a cadeia de derivação estabelece também, no caso das homógrafas, a ordem de índices numéricos a atribuir. O advérbio terá um índice mais baixo que a preposição; o adjectivo mais baixo que o substantivo homógrafo.

4) fenómenos de aglutinação geram lemas novos: *intantum, nullomodo*, etc., à semelhança de *huiusmodi*.

5) palavras de língua vulgar são tratadas como medievais e consequentemente é-lhes atribuído também asterisco.

6) os nomes próprios recebem um código específico na lematização. Tanto quanto possível, utilizam-se as formas do *Onomasticon* do F., ou do *Orbis Latinus*.[17]

Combinam-se pois critérios vários (formal, etimológico, paradigmático), procurando-se ajustar tanto quanto possível os instrumentos lexicais com uma nova metodologia de exploração de textos. Pode conjecturar-se que o resultado final seja uma revisão desses mesmos instrumentos de trabalho e uma reformulação dos critérios adoptados.

O sistema de exploração é o de forma semi-automática, tomando-se como ponto de comparação os textos anteriormente constituídos em memória, por redução das formas repetidas e selecção das formas mais rentáveis em casos de ambiguidade possível.

*

É evidente que qualquer opção é susceptível de ser contestada e de ver contraposta uma outra assente em princípios teóricos diferentes ou simplesmente em factores de alcance prático ou em objectivos imediatos. No entanto, para além de discussões a qualquer nível (e tanto mais possíveis quanto não existe uma norma aceite), uma solução prática não poderá escamotear factores como os da rentabilidade de exploração, operatoridade na constituição de dados e acessibilidade de informação por parte dos utentes futuros dos dados constituídos. Qualquer destas fases, seja de criação seja de interpretação, exige critérios, não só uniformes, objectivos, não ambíguos e ajustáveis através de regras de decisão para os casos imprevistos, mas igualmente simples para serem operatórios. Quanto mais sobrecarregada ou complexa for a codificação de entrada menor será o índice de automatização e de obtenção de resultados ou de simples leitura, embora o índice informativo possa ser mais elevado. De igual modo, quanto maior for a transformação da forma para o lema maior intervenção exigirá na constituição dos dados e menor será a informação imediata sobre a realidade do texto. Daí que haja grande vantagem em manter a transformação a um nível de integração gramatical (por ex., no caso das formas supletivas do verbo) ou em não reduzir as oposições de morfemas de classe.

A lematização, tendo como objectivo delimitar as unidades lexicais de um texto, sem abdicar de um modelo de língua, não pretenderá antecipar ou suprir a fase de análise linguística propriamente dita. Não atenderá assim, por ex., às possibilidades de equivalência

funcional, e ater-se-á preferentemente ao nível da forma mínima livre e não ao nível do sintagma.

N O T A S

1 Instruções do CETEDOC sobre lematização.

2 Cfr. por ex. as críticas feitas por L. Delatte e E. Evrard a Gaffiot e a Forcellini, em *Sénèque, Consolation à Polibe; Index verborum, relevés statistiques*, Liège, 1962, pp. VI e VII.

3 Charles Muller, *Initiation à la statistique linguistique*, Paris, 1968, p. 148.

4 Ch. Muller, *op. cit.*, p. 133.

5 Josette Rey-Debove, "Lexique et dictionnaire; l'inventaire du monde", em Bernard Pottier, *Comprendre la linguistique*, Verviers, 1975.

6 R. E. Latham, *Dictionary of medieval latin from british sources*, Londres, 1975 (1.o fasc.).

7 Roberto Busa, s, j., "The quantities of the latin vocabulary documented in the Index Thomisticus", *Revue*, n.o 3, 1976, pp. 1-45.

8 Cfr. J. Dubois, "Recherches lexicographiques; esquisse d'un dictionnaire structural", *Études de linguistique appliquée*, n.o 1 (1962), pp. 43 ss.

9 Charles Muller, *op. cit.*, p. 151. Importará recordar que para Ch. Muller "le vocable est une unité de lexique, le mot une unité de texte; on a lu un mot dans le texte, mais c'est un vocable que l'on trouvera dans le dictionnaire" (p. 133).

10 *Id., ib.*, p. 144.

11 A. Forcellini, F. Corradini, L. Perin, *Lexicon Totius Latinitatis*, tomos I-IV, Pádua, 1965 (rep. anast. e aum. de 1926).

12 Englobar-se-ão aí os léxicos organizados nas diversas nações, tanto mais que o *Novum Glossarium Mediae Latinitatis* está longe de poder servir de instrumento de trabalho adequado. Cfr. Manuel C. Díaz y Díaz, "Ruta crítica por la lexicografía latina medieval", *Helmántica*, Setembro-Dezembro, 1960, n.o 36, pp. 497-518; M. Hélin, "Le nouveau glossaire du Latin médiéval", *Revue Belge de Philologie et d'Histoire*, XXXVII (1959), n.o 1, pp. 104-111; J. F. Niermeyer, "En marge du nouveau Du Cange", *Le Moyen Âge*, LXIII (1957), pp. 329-360; Yves Lefrèvre, "Les dictionnaires du latin médiéval et l'Union Académique Internationale", *Comptes Rendus de l'Académie des Inscriptions et Belles Lettres*, Jul.-Out., 1975, pp. 402-414.

Não poderá esquecer-se a importância que representa no despojamento de textos latino-medievais o recurso aos léxicos de autores cristãos. A. Blaise, *Dictionnaire latin-français des auteurs chrétiens*, Paris, 1954, 2.a ed. cor., 1967, é um subsídio indispensável para datação do registo vocabular.

13 A divisão do trabalho em duas fases apresenta não só a vantagem de manter a ordem sequencial do texto sem alterações comprometedoras para análises ulteriores, como ainda simplifica a automatização (total ou parcial) na fase inicial. De um ponto de vista teórico, não parece, aliás, que haja objecção em seguir a ordem discreta do significante e só depois refazer as unidades de língua.

14 Para o latim clássico, o L. A. S. L. A. de Liège utiliza esta exploração automática. Cfr. A. Bodson e E. Evrard, "Le programme d'analyse automatique du latin", *Revue*, n.o 2, 1966, pp. 17-46.

15 Resultados mesmo superiores puderam ser obtidos no CETEDOC na análise de textos filosóficos medievais. Este mesmo Centro tem em constituição um léxico descritivo do vocabulário latino para aplicação em tratamento de textos.

[16] Note-se que 604 é a data de morte de Gregório Magno. Repare-se igualmente que estes critérios não são coincidentes com os do *Thesaurus Linguae Latinae*. Ficarão de pé, sem dúvida, questões importantes de periodização. S.to Isidoro é um homem da Idade Média ou representa ainda a antiguidade ? A regra de S. Bento pertence a um período ou a outro ? Não será de somenos importância, para quem faz do latim clássico o termo de comparação do latim medieval, determinar quais os factores de cisão entre um período e outro. A intervenção do cristianismo, como mentalidade e como valorização de processos linguísticos, as invasões germânicas com a quebra da influência da escola tradicional (pelo menos em extensão, já que os reis bárbaros não dispensam nas suas chancelarias os oficiais anteriores), a expansão do monaquismo e a nova organização escolar, o recuo da cultura clássica e o aparecimento de novas línguas, a recuperação da cultura tradicional pela escola carolíngea ? A dificuldade de optar por um critério ou por outro passa pela pertinência de transferir um critério social ou histórico como índice operatório para o domínio linguístico.

[17] Graese, Benedict, Plechl, *Orbis latinus; Lexikon lateinischer geographischer Namen des Mittelalters und der Neuzeit*, Brawnschweig, 1972.

AIRES AUGUSTO NASCIMENTO