A edição digital de fontes histórico-linguísticas

Rita Marquilhas

I Seminário Internacional de Humanidades Digitais no Brasil

Eu aceitei com muito gosto o convite dos organizadores para participar neste colóquio não só porque me revejo no termo Humanidades Digitais como sobretudo porque uso bastante e admiro o trabalho dos autores que costumam ser identificados com essa área de estudos.

Como estamos numa Mesa Redonda, por definição um contexto mais intimista do que o de uma conferência ou o de uma comunicação, vou manter ao longo da minha intervenção um tom mais pessoal, também para as pessoas ficarem com uma ideia de um perfil possível de investigadora cujo percurso se cruza com o das Humanidades Digitais.

A minha área de formação é a Filologia. Trabalhei durante duas décadas em edição e estudo de fontes textuais relevantes para o estudo do Português Clássico e fi-lo usando o que se pode chamar, ironicamente, de "baixa" tecnologia (ou "low" tech): a low tech das máquinas de escrever e de calcular, ou a low tech do editor de texto Wordstar, como milhões de outras pessoas; tentava imitar no texto editado no meu computador pessoal um modelo ideal que é o da empaginação tradicional dos trabalhos académicos dos séculos XIX e XX (aqui um exemplo de uma publicação de Carolina Michaëlis de Vasconcelos): como sabem do vosso treino de leitura de textos académicos, a argumentação do autor vai no corpo da página, a erudição, no espaço do rodapé; a transcrição mais alargada das fontes, nos anexos; a reprodução facsimilar, nas lâminas de figuras.

Quanto às fontes, depois de estudar as origens da nossa ortografia, passei a estudar a correspondência privada das pessoas anónimas, por vezes semi-analfabetas, que no século XVII foram perseguidas pela Inquisição Portuguesa, convencida eu de que a sua maneira ingénua de escrever me mostraria um pouco da língua oral de há 300-400 anos atrás, bem como os usos sociais da escrita na época.

Tive a sorte de descobrir uma coleção de fontes muito rica, demasiado rica para poder ser explorada isoladamente por um investigador. Demasiado rica, também, para poder ser reduzida, ao nível da edição, ao formato tradicional do livro impresso. Tratava-se de milhares de cartas informais, muitas delas familiares, ou de amizade, ou de amor, que os tribunais religiosos e civis da Idade Moderna (séculos XVI a XIX) confiscaram para as poderem usar como instrumento de prova para inculpar os seus acusados. Feitas para

terem uma vida efémera, ou para serem só guardadas em arquivos familiares, tiveram um destino muito diverso. Foram guardadas por uma das muitas autoridades judiciais do Antigo Regime, com grande protagonismo para a Inquisição Portuguesa. Junto a elas, arquivaram-se também os interrogatórios a réus e a testemunhas, portanto é possível saber, em muitos casos, o que aconteceu imediatamente **antes** e imediatamente **depois** de aquelas cartas terem sido escritas, quem as compôs concretamente, e quem as leu concretamente.

Era preciso pensar numa campanha de estudo desse material que estivesse à altura da sua complexidade, bem como do seu elevado número. Para contextualizar aqueles enunciados linguísticos enquanto práticas quotidianas, socialmente situadas, era necessário recorrer a historiadores da cultura. Para interpretar a gramática e o léxico que elas testemunhavam, eram necessários linguistas. E para garantir a integridade do texto editado eram necessários filólogos. Finalmente, para conseguir que os públicos interessados nestes materiais, que seriam públicos especializados, mas também públicos leigos, era preciso montar uma edição digital que conjugasse múltiplas camadas de informação, comentário e imagem num único hipertexto intuitivamente pesquisável.

Assim nasceu o projeto CARDS (Cartas Desconhecidas), que durou de 2007 a 2010, e que deu depois lugar ao projeto P.S. Post Scriptum, estendido a cartas também em espanhol: Arquivo Digital de Escrita Quotidiana em Portugal e Espanha na Época Moderna. Pelo meio, entre 2010 e 2012, ainda montámos o projeto FLY, dedicado a cartas do século XX da guerra, emigração, prisão e exílio.

Para concretizarmos o estudo interdisciplinar e a publicação dos milhares de documentos históricos com que lidamos, temos tido sempre que compatibilizar ferramentas digitais com as metodologias tradicionalmente seguidas pela crítica textual, pela linguística histórica e pela história cultural. Temos procurado automatizar o que é automatizável ao longo de vários processos, e aqui o que fazemos sobretudo é a chamada "Computação para as Humanidades" (ou Computing Humanities): a) para a marcação textual da transcrição conservadora dos manuscritos, baseamo-nos na grande aceitação do protocolo TEI, que em linguagem XML tem servido as edições críticas em suporte digital por um ambiente académico cada vez mais vasto; as etiquetas pensadas para a edição de fontes primárias têm uma minúcia que é totalmente compatível com o espírito conservador do filólogo, que quer assegurar-se de que a sua edição tem um formato o mais próximo possível do texto que toma como base da edição, seja ele original ou não (nós usamos quase só originais porque queremos ter a certeza de que os

enunciados históricos que utilizamos têm a maior naturalidade possível); usamos, como interface, o editor Oxygen, que permite algumas operações semi-automáticas e que, não sendo gratuito, tem um preço razoável; para mais, foi criado num país pobre, a Roménia, país que os mais afortunados podem sentir uma certa obrigação moral de ajudar.

b) normalização da variação gráfica presente nas transcrições conservadoras; depois de termos experimentado a modernização automática com alguns bons resultados, apoiados no trabalho de colegas da Universidade de Lancaster, em Inglaterra, e da Unesp (Dicionário Histórico do Português do Brasil), usamos, desde há um ano, a preciosa ferramenta eDictor, que não vou descrever longamente porque o público deste Encontro teve já oportunidade de perceber a que corresponde; vou apenas precisar que a adaptámos de maneira a utilizá-la com algumas finalidades que não estavam na sua conceção original por parte dos colegas de Campinas. Assim, não usamos a janela Transcrição, porque já tínhamos demasiado material transcrito de origem no formato TEI; usamos sim um script de conversão, escrito por nós, que nos permite passar imediatamente à janela Edição e aí modernizar o texto, marcá-lo textualmente e atribuir-lhe alguma anotação meta-linguística. Por exemplo, para a indexação em termos de palavras-chave, usamos o elemento Tipo de palavra (originalmente pensado para anotação lexical e não metalinguística) e o elemento Propriedades de Secção (pensado para marcação textual e não metalinguística). Para a anotação textual em si, por seu lado, usamos Propriedades de Sentença e Propriedades de Parágrafo. Um novo script de conversão permite reabrir o XML do eDictor como se de TEI se tratasse. Ou seja, em vez de criarmos nós aplicações novas, concentrámo-nos em aprender a programar para podermos compatibilizar as aplicações já disponíveis.

c) anotação de categorias linguísticas nos ficheiros normalizados; usamos o anotador eDictor, ou seja, o sistema do Tycho Brahe de Campinas e do CORDIAL do CLUL, para o português e o anotador freeLing para o espanhol;

d) análise do corpus — suas variantes, seu léxico, suas categorias — em termos estatísticos; usamos as ferramentas de análise lexical automática WordSmith Tools e AntConc;

e) cruzamento entre os dados linguísticos e os dados extralinguísticos recolhidos pela equipa de historiadores do projeto; esta busca fica assegurada pela carregamento dos ricos ficheiros em XML-TEI, que comportam dados textuais e metadados históricos e linguísticos, num website pesquisável pelo utilizador.

Traçado este breve panorama, queria passar agora a algumas reflexões, ou observações críticas. Começo por listar as vantagens desta Computação para as Humanidades, ressalvando que estou a falar sobretudo de Filologia e de História Cultural.

A vantagem da introdução de tecnologia informática em disciplinas que dependem de fontes textuais é óbvia e já foi suficientemente sublinhada neste colóquio. Se estamos a falar do trabalho desenvolvido individualmente, a adoção da tecnologia avançada permite acelerar o ritmo (isto parece claro para toda a gente; só não está ainda explicada é a razão de tanta produtividade por parte dos filólogos do século XIX e do início do XX, se não tinham todo este apoio tecnológico; pegando no que eles fizeram, dá ideia de que nunca conseguiríamos chegar a tanta produção, mesmo que tivéssemos várias vidas); o que aumenta, seguramente, é o número de fontes a que um investigador pode aceder, no caso de estarem digitalizadas; também embaratece a investigação, por permitir viagens virtuais a horas proibidas a arquivos e bibliotecas longínquas; elimina-se ainda o risco do erro em operações sobre o texto, e favorecem-se novas descobertas, apoiadas em operações automáticas.

Sem toda esta mecanização do estudo dos textos, não nos seria possível, por exemplo, descobrir que os padrões de sequência de duas palavras do discurso das cartas familiares do século XX era este que encontrámos nas duas mil cartas do FLY.

A nível sintático-semântico, observa-se saliência do predicado epistémico de polaridade negativa 'não sei', que serve também a expressão da disforia; pragmaticamente, sobressaem ainda outras proposições apoiadas em 'não' ('não te', 'não me', 'não é'), centrais tanto nos atos diretivos como em múltiplas estratégias de delicadeza; textualmente, percebe-se quão significativo é o protótipo da argumentação, dada a saliência dos operadores 'mas (não)' e 'por isso'; lexicalmente, revelam-se significativas as expressões fáticas e idiomáticas do coloquialismo familiar, 'meu querido' e 'graças a Deus', dada a constante presença das mesmas nas partes formulaicas, logo fixas, das cartas.

Se estamos a falar de trabalho desenvolvido em equipa, a tecnologia em linha permite ainda, para além das vantagens acima enunciadas, a harmonização do trabalho do grupo, centralizando a informação, impedindo a repetição de tarefas, acelerando ainda mais o ritmo produtivo e enriquecendo os resultados em consequência da interação que assim se favorece entre diferentes sujeitos pensantes.

Pondo de lado as vantagens metodológicas, lembraria agora as vantagens sociais. A área das Humanidades, muito apoiada na memória e na erudição, sempre favoreceu os

estudiosos que já beneficiavam de ter crescido em meios onde os temas culturais estavam omnipresentes. Uma biblioteca de família, uma educação em instituição de élite, uma vida em ambiente urbano... eram meio caminho andado para se poder vir a ser filósofo, historiador, literato... Esta área de estudos também sempre favoreceu a idade, pelo tempo que demora um indivíduo a acumular na memória os saberes dos clássicos, acrescentados depois com os dos modernos. Os académicos trabalhando em Humanidades foram assim, sem surpresa, as primeiras vítimas da revolução tecnológica, da liberalização da economia, da globalização da cultura. Nas Universidades, os cursos perderam prestígio, perderam atração, perderam alunos. Tais académicos continuam a trabalhar, mas cada vez mais isolados, lamentando com amargura a suposta ignorância dos jovens que à sua volta veem triunfar. Eram (são) a versão moderna do sábio isolado na sua torre de marfim.

O mundo das Humanidades Digitais está a alterar este panorama. A era digital, se é verdade que acelerou vertiginosamente a circulação da informação e foi responsável pela transformação da cultura no que se passou a chamar um produto industrial, também é verdade que democratizou o conhecimento em todos os ramos do saber, incluindo os das Humanidades. Não é preciso pertencer originalmente a uma classe de élite, ou ser-se já muito adulto, para se chegar a sábio (só que hoje não se lhe chama sábio, chama-se 'nerd').

Ao mesmo tempo, numa equipa de Computação para as Humanidades, há um lugar privilegiado que só pode ser ocupado por uma ou várias pessoas jovens, com conhecimentos de engenharia informática. O seu domínio da tecnologia é imprescindível ao grupo e a sua capacidade de programar operações automáticas com os textos ou sobre os textos garante-lhes o respeito dos seus companheiros de letras, que tudo dariam para conseguirem criar, eles também, um pouco daquela magia. Esse acaba por ser mesmo o passo seguinte neste tipo de equipas: os filólogos, historiadores ou linguistas de origem, por exemplo, aprendem um mínimo de programação, o que lhes permite imaginar pedidos mais certeiros em termos de novas ferramentas de trabalho. Os informáticos, por seu turno, são forçados a conhecer cada vez melhor as coleções de dados sobre as quais estão a programar, deixando progressivamente de as verem como dados e olhando-as cada vez mais como criações do espírito humano.

Mas não quero terminar sem matizar um pouco o tom ingenuamente eufórico com que vos falei da Computação para as Humanidades. O lado mais escuro deste empreendimento é o que encontramos também em todos os setores da atividade em que

se empreguem processos mecânicos. É que a falha humana, a este nível, tem consequências muito mais devastadoras do que as de falhas em produções artesanais. Algumas estratégias de prevenção do erro são as que passam pelas rotinas de verificação pelos próprios e pelo controlo de qualidade pelos pares. Uma estratégia paralela é a da
insistência na inclusão de todas as gerações de académicos, mesmo dos que à partida têm tendência para ser info-excluídos. Não se podem, por isso também, cortar todas as amarras com as formas tradicionais de investigação e de publicação de resultados: não devemos pensar usar o digital para reinventar as Humanidades. É mil vezes preferível usá-lo para as fazer renascer.