O SABRE E A MONTADA. OBSERVAÇÕES SOBRE A PRÁTICA DE UMA EDIÇÃO ACADÉMICA DIGITAL

Rita Marquilhas
CLUL, Universidade de Lisboa

DOI: 10.17075/cbfc.2020.007

1. INTRODUÇÃO

Deve-se a Louis T. Milic um apelo feito há mais de cinquenta anos, e desde então muito repetido, para que os académicos das ciências humanas saibam integrar bem os computadores no seu trabalho. O ano era o de 1966, o espaço era o da inauguração da revista *Computers and the Humanities*, e o autor estava a reunir aquele que viria a ser o *Century of Prose Corpus*, um corpus histórico representativo da prosa setecentista em inglês britânico (Milic 1990). Milic punha as coisas em termos de dar "um passo em frente" – usou o título *The next step* – e de assumir um controlo sobre a máquina semelhante ao do braço do esgrimista sobre a arma ou ao do corpo do cavaleiro sobre o animal (Milic 1966: 4):

> O nosso medo de que o estudo da literatura se venha a tornar mecânico se for processado por computador tem-nos impedido de tentar apreender as suas ricas e genuínas possibilidades. Se não tentarmos compreender o computador da mesma maneira como, enquanto académicos, fazemos com outros instrumentos, não só nos tornaremos incapazes de explorar adequadamente os seus recursos como correremos o risco de nos transformarmos em vítimas suas. O controlo nasce da compreensão, da fusão do utilizador com o seu instrumento, *como a do braço com o sabre ou a do cavaleiro com a montada* (traduzido do inglês; itálico nosso).

Este mesmo texto de Milic já foi usado por Willard McCarty para argumentar que foi pela década de 1960 que se inaugurou todo um género textual, feito de lamentações e atribuições de culpa, praticado pelos académicos na sua auto-crítica às humanidades digitais e à falta de imaginação que os seus projetos deixam transparecer (McCarty 2013: 36–37). Mas eu prefiro retomar o texto de Milic por uma outra razão, que tem a ver com a felicidade do estilo daquele autor. Com efeito, as imagens do esgrimista com o sabre e do cavaleiro com a montada são imediatamente sugestivas para todos quantos tenham empreendido uma ini-

ciativa em humanidades digitais. Em algum momento do seu trabalho estes são académicos que já se sentiram aquele mesmo atirador cujo sabre lhe prolonga o braço e o ajuda a atacar e a defender-se num assalto. Em outras alturas, ter-se-ão sentido o cavaleiro que corre distâncias e transpõe elevações socorrido da velocidade, do impulso e da elegância do seu cavalo. São dois conjuntos típicos, aos quais Louis T. Milic não hesitou em juntar um terceiro, formado pelo académico e pela sua máquina, juntos na delineação de um projeto digital (no caso de Milic, o processamento eletrónico de textos literários da história do inglês). Fosse o autor um cientista social e esta sua certeira descrição teria vindo a ser invocada no quadro da *teoria ator-rede*, aquela abordagem sociológica que leva longe a proposta de que formamos constantemente à nossa volta objetos híbridos, feitos de humano, não-humano e discurso. Será preciso encontrarmos uma maneira de representar a existência de tais híbridos, sob pena de nos escapar o correto diagnóstico da sociedade moderna, argumenta-se ainda nesta teoria (Latour 1991: 22).

Neste meu artigo, vou ficar certamente aquém do género do ensaio antropológico. Compus o texto por ocasião de um balanço que me foi pedido pelos colegas do Instituto da Lingua Galega aos primeiros cinco anos do projeto *P.S. Post Sriptum*, e vou portanto argumentar algo mais modesto: que a montagem de tal recurso eletrónico – que correspondeu à constituição de um *Arquivo Digital de Escrita Quotidiana em Portugal e Espanha na Época Moderna*, como informa o seu subtítulo – permitiu também observar de perto uma transição muito atual, experimentada pelas ciências humanas. No caso concreto, concentrar-me-ei no desafio que enfrenta a filologia ao ter de revelar capacidade para se manter fiável num mundo muito diferente daquele de onde veio, um mundo em que os arquivos e as bibliotecas eram ainda espaços físicos, a cultura era impressa e o público era escasso, composto de especialistas e estudantes. Agora, numa era tecnológica bem distinta, teremos de sair do "silo" da academia e "sermos vistos a viver e a trabalhar em público, com o público e para o público" (Pound / Liu 2013: §33).

2. ORIGEM E CARACTERIZAÇÃO DO *P.S. POST SCRIPTUM*

O *P.S. Post Scriptum*, uma iniciativa em humanidades digitais, arrancou em 2012 no Centro de Linguística da Universidade de Lisboa com o objetivo de, ao longo de cinco anos, mobilizar uma equipa que procedesse à campanha de recolha siste-

mática, edição digital e estudo histórico-linguístico de cartas particulares escritas em português e em espanhol ao longo da Idade Moderna desde o século XVI até ao início do século XIX[1].

Anteriormente, no âmbito de uma monografia (Marquilhas 2000) e de dois projetos financiados (*Por toda a parte... uma certa história da língua portuguesa* e *CARDS, Cartas Desconhecidas*[2]), já se tinham explorado fundos arquivísticos portugueses que consistiam em processos criminais da Idade Moderna e do século XIX contendo, como documento de prova, cartas originais escritas por gente comum. Tinha-se percebido que tais documentos existiam efetivamente em abundância, enviados por criados, crianças, cônjuges, amantes, ladrões, soldados, artesãos, padres, lutadores políticos, entre muitos outros tipos de agentes sociais. Os respetivos percursos tinham-se cruzado com a ação persecutória e punitiva dos múltiplos tribunais que, naqueles séculos, usaram a correspondência privada, apreendida como prova, para a instrução das suas causas (a justiça inquisitorial, a justiça episcopal, as justiças central e periférica da Coroa e a justiça das Ordens Militares, sobretudo). Como tais documentos se tinham guardado nos arquivos, acompanhados de autênticos inquéritos *sociológicos*, empreendidos pelos promotores ou juízes que interrogaram os réus e as testemunhas dos casos sob investigação, valia a pena recolher estes vestígios das comunidades tradicionais na faceta das suas relações interpessoais. Verificou-se que, apesar de escondidos (em cada 100 processos da Casa da Suplicação, por exemplo, só uma média de 17 continham cartas), os textos dessas cartas registavam, muito frequentemente, uma retórica quase oral, atravessada pela verbalização de emoções, pela explicitação de juízos do senso-comum, pela confissão de crenças não-ortodoxas ou pela narração de incidentes triviais. Incluíam, portanto, um discurso *popular*, ao qual tanto os historiadores da língua como os historiadores da cultura têm normalmente dificuldade em aceder se não quiserem ficar por amostras só lateralmente representativas como são as dos textos de teatro, satirizando tipos sociais e suas circunstâncias, ou as do discurso político, que ajuízam o grau dos contrastes sociais.

[1] O projeto foi financiado pelo European Reasearch Council por meio da atribuição de uma bolsa com o seguinte código: 7FP/ERC Advanced Grant 2011, Grant Agreement 295562.

[2] Financiados, respetivamente, pelo Instituto Camões em 2004-2005 (programa Lusitânia), e pela Fundação para a Ciência e a Tecnologia em 2007-2009 (PTDC/LIN/64472/2006).

Antes do *P.S. Post Scriptum* já se tinham, portanto, detetado 2.000 dessas cartas, das quais se tinha feito, porque a edição em papel era impensável, a transcrição paleográfica em XML-TEI e a contextualização baseada nos conteúdos dos processos criminais. Tinha sido possível, a partir das mesmas fontes, elaborar até alguns estudos de caso, tomando como fonte ora cartas isoladas ou pequenos conjuntos e explorando sobretudo o seu discurso (ex: Marquilhas 2009). Quanto ao corpus em XML-TEI que se tinha conseguido montar no âmbito do CARDS, sobre ele fizeram-se logo dois ensaios no plano do processamento da linguagem natural, o primeiro visando a segmentação automática do discurso formular e não formular (Hendrickx / Généreux / Marquilhas 2011) e o segundo a normalização automática das grafias variantes (Hendrickx / Marquilhas 2011).

Assim, em 2012, quando se iniciou o *P.S. Post Scriptum*, a noção era a de que com uma equipa mais ampla e mais heterogénea do que as que tinham trabalhado anteriormente, não só se reuniria uma amostra muito mais representativa daquelas fontes epistolográficas como se conseguiria conciliar o seu tratamento filológico – uma edição crítica digital – com a criação de um corpus linguisticamente anotado e a respetiva publicação *online* numa plataforma capaz de convidar públicos variados a estudarem ou simplesmente a usufruírem do recurso.

Saltando para 2017, conseguiu-se nesse ano que a edição digital das cartas do *P.S. Post Scriptum* correspondesse, efetivamente, a um *corpus* bilingue de envergadura considerável, contendo dois milhões de palavras (um milhão de palavras em cartas portuguesas e um milhão também em cartas espanholas). O *corpus* é filologicamente consultável por estar totalmente sobreposto a uma edição académica digital, um desiderato incontornável uma vez que os documentos a estudar eram manuscritos originais que tinham sido precisamente escolhidos na sua qualidade de fontes para a história da língua portuguesa em termos do seu uso quotidiano: todos os traços de origem, dos paleográficos aos textuais e aos materiais, tinham de ficar rigorosamente registados sob pena de se perder a informatividade linguística e cultural das fontes. Simultaneamente, a edição é do interesse dos historiadores por conter ligações entre os textos das cartas e pequenas composições, com o resumo da documentação judicial que as cartas integram dentro dos arquivos onde se conservaram até hoje. Uma lista de 245 palavras-chave permite, por outro lado, a consulta das cartas agrupadas por coerência temática. E a ancoragem de coordenadas geográficas ao ficheiro de toda e qualquer carta cujo local de

envio tenha sido identificado dá origem, por sua vez, ao mapeamento virtual e interativo do conjunto dos documentos.

Em terceiro lugar, o *corpus* é linguisticamente pesquisável por estar lematizado e anotado em termos de Parts-of-Speech (POS) e de secções textuais (abertura, arenga, narração, peroração e fecho), além de marcado com palavras-chave, sobretudo de natureza fonológica. Parcialmente, ainda, em 20% das partes textuais narrativas, i.e., não formulares, as cartas estão anotadas sintaticamente.

Trata-se, finalmente, de um *corpus* e de uma edição de livre acesso, estando todos os materiais alojados num servidor do Centro de Linguística da Universidade de Lisboa, mantido pela Reitoria da mesma universidade. Tem este endereço: http://ps.clul.ul.pt.

Em termos cronológicos, os documentos cobrem um período amplo, indo de 1504 a 1833. Em termos geográficos, abrangem a Península Ibérica, naturalmente, mas também os espaços dos antigos impérios português e espanhol. E como estão maioritariamente contextualizados em termos dos respetivos enquadramentos situacional e social, permitem que o público tenha uma apreensão invulgar das vidas de 5.000 indivíduos diferentes, que viveram em algum tempo da Idade Moderna e que, enquanto remetentes ou destinatários de cartas, se envolveram numa interação que foi julgada na sua época como doméstica, próxima, íntima ou proibida. A distribuição média é de duas cartas por remetente, estatística que converte este *corpus* no mais variado alguma vez reunido para a história do português e do espanhol. Além disso, fruto do tipo textual das fontes, elas são também extremamente valiosas por causa da *baixeza* dos registos linguísticos que contêm: i) 23% do *corpus* contém texto escrito por mulheres e ii) 60% do *corpus* contém texto de remetentes que não pertenciam a uma élite: não eram do clero nem da nobreza.

Os tribunais da época interessaram-se na altura por estes textos por causa da sua dimensão de prova de culpa. Nós, leitores do nosso tempo, olhamos para eles de outro modo, obviamente. Através das palavras destas testemunhas involuntárias do quotidiano privado e do registo linguístico informal em espanhol e português da Idade Moderna – palavras de raiva, amor, obscenidade, vingança ou superficialidade – podemos ir à origem, se formos linguistas, de muitas propriedades dos dialetos atuais e acompanhar o seu desenvolvimento em combinação com fatores extralinguísticos. Se formos historiadores, podemos observar

um lado escondido da herança das instituições da Idade Moderna ao lermos o discurso daqueles que foram perseguidos por lhes faltarem as virtudes apropriadas para o seu tempo, i.e., pela sua falta de moralidade, de catolicismo, de honra, de respeito pelas leis do Rei ou de lealdade para com o estado real.

Resumindo, no âmbito do *P.S. Post Scriptum* foi possível i) montar um corpus robusto a partir destas fontes dispersas e escondidas debaixo do peso da burocracia judicial; ii) desenvolver um método apropriado (em termos de marcação do texto e anotação das estruturas) que fizesse justiça à riqueza do discurso contido nas mesmas fontes, e iii) oferecer ao público, tanto o académico como o leigo, um recurso *online* que funcionasse ao mesmo tempo como uma gramática, um glossário, uma edição académica digital, uma base de dados, um mapa, e, não menos importante, um longo "romance" polifónico com personagens e histórias inesperadas de há 200, 300, 400 e 500 anos.

3. O *SABRE* PERANTE O DESAFIO DA TECNOLOGIA EXISTENTE E A *MONTADA* PERANTE OS OBSTÁCULOS DA EXPERIMENTAÇÃO

Exposições como a da secção anterior permitirão identificar uma iniciativa em humanidades digitais. No entanto, não a explicam para quem a queira replicar nem a descrevem para quem a queira julgar. Infinitamente mais útil será identificar traços do comportamento da equipa, que, para garantia da pertinência dos seus achados no contexto tecnológico atual, teve de assumir-se ora "atiradora" ora "hípica", para retomar as imagens de Milic referidas na introdução. Isto equivale a dizer, em linguagem mais direta, que é útil identificar tanto os *perigos* como as *dificuldades* que uma equipa enfrentou ao pretender usar a tecnologia para fins académicos.

Tal uso é constantemente ameaçado, em primeiro lugar, pelos perigos de se adotarem sem problematização sistemas que venham de fora, seja do mundo da tecnologia comercial seja do das experiências anteriores dentro das próprias humanidades digitais. É preciso reagir-lhes com uma atitude defensiva mas ao mesmo tempo dinâmica. Por outras palavras, é preciso saber esgrimir, parar e responder, usando a máquina como uma arma que prolonga o braço humano e lhe permite ir mais longe do que parecia fisicamente possível. É assim que se vão imaginando sistemas tão eletrónicos como os outros, mas mais naturalmente, e não forçadamente, humanísticos, além de cada vez mais eficientes do que os anteriores.

Por outro lado, há também aquelas dificuldades comuns a todas as iniciativas de investigação que criam simultaneamente tecnologia e que progridem à custa da execução de testes, cada vez mais numerosos e cada vez mais desafiantes. Esta faceta da experimentação exige que o experimentador falhe, bloqueie, retroceda e recomece constantemente, entendendo agora a máquina tal como o cavaleiro tem de entender a sua montada para que os saltos num campo de obstáculos possam ser dados em conjunto, estando o verdadeiro impulso na capacidade para combinar as faculdades de ambos. É assim mesmo que se consegue, depois de muito treino, levar por exemplo um computador a etiquetar bem textos tão desafiantes como podem ser os dos manuscritos históricos, a descobrir-lhes regularidades e exceções, bem como ligações escondidas a variáveis externas. É também assim que se abre caminho para o desenho de interfaces bem integradas no contexto a que pertencem os públicos atuais, que é um contexto inexoravelmente tomado pela circulação da informação em rede.

Identificado o objetivo desta secção, passarei à narração crítica do processo de montagem do *P.S. Post Scriptum*, incluindo a identificação das falhas cometidas ao longo de uma política de sucessivas escolhas em nome da obtenção do desenho, nunca antes tentado, de um local da *web* para onde pudessem convergir os interesses da linguística histórica e da história da cultura escrita na exploração do conteúdo de textos muito particulares: milhares de cartas de gente vulgar, apreendidas pelos tribunais de Portugal e Espanha ao longo da Idade Moderna.

A abordagem crítica impõe-se também por me parecer útil contrariar certo registo *comercial* que costuma demasiadas vezes acompanhar os relatos de como nasce o software digital. Com efeito, e como foi assinalado pelo historiador da tecnologia Michael S. Mahoney, a história do que já fizemos ao "metermos o mundo dentro de computadores" sai bastante dificultada por ser uma história que quase só se pode basear em relatos "de sucesso" (Mahoney 2005: 128). Ora há uma óbvia separação (feita de falhas, desaires, desastres até) entre o que se imagina e o que se consegue alcançar ao longo dos processos de envolvimento dos humanos com a tecnologia que criam. As iniciativas em humanidades digitais passam forçosamente pela experimentação de modelos, ferramentas e plataformas. Este sentido da experimentação vem-lhes do que têm de comum com as ciências da computação, as quais foram introduzindo modificações nas linguagens e nos objetos digitais não tanto por terem descoberto muitos "novos

princípios" mas mais porque exploraram as "possibilidades da prática" (Mahoney 2005: 131). Quer isto dizer que é, sobretudo ao nível da razão prática e não da teórica – i.e., na reflexão envolvida nas deliberações humanas que levam à ação e não tanto na reflexão sobre factos e sua explicação (Wallace 2003/2018: §1) – que as comunidades da computação têm surgido com inovações.

No nascimento das humanidades digitais, houve, com efeito, e desde cerca de 1980, várias deliberações criadas pelo problema inicial de se ter pedido emprestada a computação dos outros, já que não se conseguia desenvolver uma computação própria. Para os seus próprios fins, eruditos ou educacionais, os estudiosos das Letras começaram a utilizar produtos que tinham sido desenhados para clientes empresariais, passando-lhes despercebido o facto de aquelas ferramentas trazerem incorporada uma sua história muito própria (Mahoney 2005: 133). O que tem movido as humanidades digitais desde esse início tem sido a verificação de que é preciso obter "a confiança e a fluência" necessárias (Berry 2011: 26) para se produzirem artefactos desenhados de raiz em torno das questões importantes para o estudo das Letras. O processo envolve experiência atrás de experiência e falha atrás de falha. As decisões de desenho, porque, na expressão de Jeffrey Schnapp, é de "desenho do conhecimento" que se trata, passam por aceitar sempre a premissa de que "nem os métodos que produzem conhecimento humanístico nem as formas e géneros em que esse conhecimento vai ser moldado são coisas dadas" (Schnapp 2014: 5).

No desenrolar do *P.S. Post Scriptum*, e perante a frequente inadequação do que lhe era *dado*, foi também sendo necessário ir moldando novos métodos, formas e géneros.

Em primeiro lugar, foi preciso ter presente que a lógica na sociedade atual, no que diz respeito à oferta de conhecimento em rede, passou a ser uma lógica de *abundância* de recursos em vez de uma lógica da escassez. O facto tem uma importância cada vez mais evidente, a ponto de, a Willard McCarty, lhe lembrar o "proverbial ladrão na noite", na perspetiva das ciências humanas e da sua missão:

> Para as humanidades, o proverbial ladrão na noite parece-me ser a acumulação de recursos primários e secundários, todas essas *juke-boxes* do conhecimento a abarrotar de textos, imagens e sons. O que é que significa ter isto como mobiliário da nossa investigação? (McCarty 2013: 46).

No desenrolar do *P.S. Post Scriptum,* essa mesma pergunta, sobre o que é que se pode fazer de racional, e racionalmente histórico-filológico, com tanto material de arquivo, fez-se efetivamente muitas vezes. Logo de início estava claro que o desiderato tinha passado a ser, nesta como em qualquer outra modalidade de edição crítica ou de história cultural, o de consultar *toda* a informação, e não só a mais representativa, mesmo que tal não parecesse imediatamente realizável. Portanto, e perante a constatação óbvia de que durante os cinco anos do financiamento inicial do *P.S. Post Scriptum* os investigadores mobilizados para os arquivos não iriam conseguir ver *todos* os fundos judiciais portugueses e espanhóis da Idade Moderna que contivessem cartas privadas como instrumento de prova, eles, mesmo assim, precisavam de prevenir um futuro em que todos esses documentos seriam tão facilmente alcançáveis como o são já, por exemplo, os conteúdos de um *Google Books*. Com os olhos nessa possibilidade, começou a trabalhar-se no conceito do que seria a resposta ideal na perspetiva das humanidades digitais, chegando-se à conclusão de que, para não se vir só a ter uma *juke-box* de conhecimento quando os arquivos vierem a ter os textos dos seus manuscritos todos eles processáveis, é preciso ter sistemas que não saibam só, quais OCR, reconhecer caracteres manuscritos. É preciso ter um *pipe-line* o mais longo possível, que comece por unir o conteúdo das prateleiras dos arquivos (as imagens dos seus manuscritos) ao dos ficheiros das salas de índices, mas que não fique pela transformação desses conteúdos em texto processável por computador. A ligação tem de continuar, garantindo sempre a reversibilidade, da função de reconhecimento de caracteres para a da correção de texto, sua lematização e sua anotação, gramatical, semântica-lexical e discursiva. Assim se foi formando a ideia do sistema ideal que o projeto poderia ajudar a construir, e que acabou por vir a concretizar-se na forma do *TEITOK*, de que será questão a seguir e que está também descrito noutro capítulo desta mesma publicação (cf. Janssen / Vaamonde).

Em segundo lugar, e ainda a propósito da necessidade de recusar muitas vezes o que é *dado* em iniciativas de humanidades digitais, cito abaixo os autores de *digital_humanities*, que assim sintetizam a sua prática:

> [Nas humanidades digitais] favorecemos o processo em detrimento do produto; favorecemos a elaboração de versões e extensões em detrimento das edições definitivas e dos

> silos de conhecimento. A capacidade das humanidades digitais de perguntar, desenhar e modelar novas perguntas para a sua investigação abre novas possibilidades para aqueles que querem correr riscos. É demasiado frequente que o experimental esteja ausente do discurso cultural estabelecido ou que nele seja rapidamente apagado, uma vez que o diálogo já está tão consolidado que as novas abordagens são, na melhor das hipóteses, tão-só um incremento das antigas. Mas não teremos nunca uma experiência se ela não puder falhar (Burdick *et al.* 2012: 22; traduzido do inglês original).

Inspirada pela referência ao centramento nos "processos" (e não nos objetos) e à elaboração de "versões" e "extensões" (e não de edições definitivas e de silos de conhecimento), referir-me-ei a partir de agora aos tipos de erro que julgo terem sido cometidos durante a montagem do *P.S. Post Scriptum* no que diz respeito à sua edição digital, tanto ao nível da idealização, por centramento involuntário no objeto e não no processo, como ao nível da concretização, por desenho de artefactos que se viriam a revelar mais inflexíveis do que extensíveis.

Um dos primeiros problemas criados no arranque do *P.S. Post Scriptum* prendeu-se com a seleção dos modelos que na altura pareciam inspiradores e exemplares, montados em instituições estrangeiras e consistindo em edições académicas digitais já prontas, ou ainda em curso, e em *corpora* anotados já disponíveis. A idealização para a publicação *online* passava por tentar atingir três resultados ao mesmo tempo: primeiro, a mesma qualidade filológica e gráfica da edição de cartas manuscritas disponibilizada pelos *Mark Twain Papers Online* (MTPO)[3] da California Digital Library; depois, a racionalidade de método que transparecia do Manual de edição DALF, *Digital Archive of Letters in Flanders* (Vanhoutte / Van den Branden 2003), entretanto concretizada na edição *Van Nu en Straks. De Brieven*, da Academia Real da Língua e da Literatura Holandesa[4], bem como da do Manual para a edição digital do epistolário de Giacomo Puccini (Pierazzo 2007), ambos aplicando o sistema de marcação de texto XML-TEI. Já a idealização para a montagem do *corpus* passava por gerar um objeto compatível com o *Corpus Histórico Português Tycho Brahe*[5], da Universidade de Campinas.

[3] Endereço do MPTO: http://www.marktwainproject.org/homepage.html [consultado: 12/2018].

[4] Endereço da edição digital *Van Nu en Straks. De Brieven*: http://vnsbrieven.org/index.htm [consultado: 12/2018].

[5] Endereço do *Tycho Brahe*: http://www.tycho.iel.unicamp.br/corpus/ [consultado: 12/2018].

É claro que o problema a que me estou a referir não reside naquelas edições, naqueles Manuais ou naquele *corpus*, que se mantêm até hoje exemplares. O problema, na perspetiva do lançamento inicial do *P.S. Post Scriptum*, esteve em se fixar um *padrão duplo*, o que acarretou durante muito tempo a crença em que o resultado mais desejável que se poderia obter seria constituído por dois objetos distintos (edição digital das cartas por um lado e *corpus* anotado com o texto das cartas por outro). Durante dois anos, e ao mesmo tempo que iam visitando quase cinquenta arquivos diferentes, os membros da equipa organizaram pacientemente uma coleção de textos epistolográficos transcritos e anotados em XML-TEI, com o detalhe mais próximo possível dos ideais MTPO, DALF e Puccini, e prepararam os materiais para uma futura anotação morfossintática e sintática do corpo das transcrições. A ideia era a de lançar mão das mesmas ferramentas com que o *Tycho Brahe* trabalhava, *Edictor* para a anotação POS[6] e *CorpusSearch 2* para a anotação sintática[7], tirando igualmente partido dos *corpora* de treino que o *Tycho Brahe* tinha já montado.

A prática revelou que a opção era pouco imaginativa, e mesmo desastrosa, do ponto de vista da articulação de tarefas, desencadeando uma infindável repetição de campanhas de correção que pareciam ir criando cada vez mais erros em vez de os eliminarem. Uma vez anotado sintaticamente um ficheiro, tinha de se ir rever o seu "irmão" com anotação POS e o "irmão" deste com anotação filológica. Por outro lado, uma vez detetado um erro de leitura do manuscrito, tinha de se prolongar a correção pelos ficheiros com a anotação POS e a anotação sintática. Os artefactos (i.e. os ficheiros) que iam sendo gerados tinham assim um grave problema de inflexibilidade, para não falar da sua falta de extensibilidade, já que ia sendo adiada a questão, cada vez mais premente, de como tornar pesquisável a relação entre as variáveis contidas nos três tipos de artefacto, afinal o grande objetivo do *P.S. Post Scriptum*.

A chegada à equipa do Centro de Linguística da Universidade de Lisboa, no verão de 2014, do linguista computacional Maarten Janssen revelou-se providencial na superação do obstáculo, uma vez que lhe surgiu a ideia de, em vez de pensar que um dia se poderia fazer convergir para os ficheiros XML-TEI (que continham a edição filológica das cartas) a informação dos ficheiros com ano-

[6] Endereço do *Edictor*: https://edictor.net/[consultado: 12/2018].

[7] Endereço do *CorpusSearch 2*: http://corpussearch.sourceforge.net/ [consultado: 12/2018].

tação POS e sintática, se poderia tentar o caminho inverso. Aos poucos, foi-lhe surgindo a ideia do *TEITOK* (cf. Janssen 2016a e Janssen / Vaamonde nesta publicação), um sistema baseado num etiquetador automático que Janssen já tinha desenvolvido anteriormente, o *Neotag* (Janssen 2016b), e que foi enriquecendo com cada vez mais funcionalidades. Assim, um problema mais difícil de resolver: – *Como extrair texto, sem perda de informação, de uma edição académica digital para um formato aceitável que dê entrada num etiquetador automático?* – foi substituído por outro mais simples: – *Como refinar uma plataforma, que já etiqueta morfossintaticamente texto, com funcionalidades que lhe permitam ancorar a cada palavra cada vez mais informação, da filológica* à *extratextual?* (v. também Marquilhas / Hendrickx 2016: 273–274).

A questão da convergência com a anotação sintática demorou mais tempo e até hoje ainda está só parcialmente resolvida. Já é possível pesquisar em *TEITOK* o texto sintaticamente anotado, é possível cruzá-lo com variáveis extratextuais, é possível corrigir a anotação na mesma plataforma, mas ainda é necessário guardar a informação sintática num ficheiro separado que a plataforma faz convergir no momento de cada consulta. Há também algumas vantagens neste formato da chamada *stand-off annotation* (McEnery / Wilson 1996/2001: 38), porque permitem alguma legibilidade humana dos ficheiros que ainda contêm a informação separada, pelo que não é claro que o caminho ideal do *TEITOK* seja o do aninhamento total, num mesmo ficheiro, da informação textual, filológica, histórica, morfossintática e, a que falta aninhar nos ficheiros do *P.S. Post Scriptum*, sintática.

Todo este foi um processo cujas fases iniciais, de experimentação de ferramentas alheias e de eleição de modelos já prontos, nunca se revelaram, em contrapartida, exercícios estéreis. Na plataforma *TEITOK* conseguiu-se, por exemplo, pôr a funcionar com rapidez um normalizador automático que converte uma leitura paleográfica de manuscrito numa sua leitura modernizadora porque entretanto já havia um conjunto de 250 mil palavras portuguesas e 400 mil palavras espanholas normalizadas à mão com recurso à ferramenta *Edictor*. O mesmo para a anotação POS do português, que foi inicialmente feita em *Edictor* e depois usada como corpus de treino para o *TEITOK* aplicar o seu próprio automatismo. Além de que uma equipa treinada na execução manual de uma série de tarefas que impliquem atenção simultânea a sistemas de etiquetas e aos elementos de um texto

corrido está depois muito bem preparada para passar a fases em que a atribuição de anotação é feita automaticamente e já só fica a faltar uma revisão manual do resultado. A fluência e naturalidade com que, mais tarde, se passou a imaginar que possibilidades oferecer aos visitantes nas janelas de buscas e de *downloads* beneficiou também muito deste treino manual anterior.

5. CONCLUSÃO

As ciências humanas, que se constituíram desde o século XIX em modalidades disciplinadas, mono-autorais, pausadas e elitistas estão a transformar-se, no enquadramento das humanidades digitais, em modalidades interdisciplinares, colaborativas, dinâmicas, culturalmente apelativas, socialmente integradas e tecnicamente atualizadas. Isto não tem de significar, de forma alguma, um abandono do pensamento típico das humanidades, que se mantém comprometido, como sempre esteve, "com questões de valor e de interpretação, com os domínios da retórica e da lógica, com juízos subjetivos combinados com a atenção a verdades verificáveis". Só que nas humanidades digitais tal pensamento foi desafiado a tornar-se "mais explícito nas suas premissas", de maneira a poder ser entendido em "ambientes computacionais" (Burdick *et al.* 2012: 4). Foi minha intenção, neste artigo, adotar precisamente essa argumentação, ao mesmo tempo que apresentava o caso concreto da iniciativa *P.S. Post Scriptum*, caracteristicamente de humanidades digitais por apresentar as três variantes que em geral identificam, segundo N. Katherine Hayles e Jessica Pressman, esses mesmos trabalhos:

> a digitalização de informação histórica por meio da realidade virtual e da realidade aumentada, a análise de texto contido em *corpora* demasiado vastos para poderem ser lidos por inteiro e a reflexão teórica sobre a natureza, os efeitos e as especificidades dos diferentes meios [digitais] (Hayles / Pressman 2013: §11; traduzido do inglês original).

Quis aqui demonstrar que o desafio das humanidades digitais é o de os seus estudiosos construírem um ambiente laboratorial dentro do qual, ao longo de sucessivas experiências, se alcancem equilíbrios entre a elaboração de questões das ciências humanas e a sua explicitação numa linguagem que a máquina saiba entender e os pares consigam replicar.

Referências bibliográficas

Berry, David M. (2011): *The Philosophy of Software. Code and Mediation in the Digital Age*. Basingstoke / Nova Iorque: Palgrave Macmillan.

Burdick, Anne / Johanna Drucker / Peter Lunenfeld / Todd Presner / Jeffrey Schnapp (2012): *Digital_Humanities*. Cambridge MA / Londres: The MIT Press.

Hayles, N. Katherine / Jessica Pressman (2013): «Making, Critique: A Media Framework», em N. Katherine Hayles / Jessica Pressman (eds.), *Comparative Textual Media. Transforming the Humanities in the Postprint Era*. Minneapolis / Londres: University of Minnesota Press, vii-xxxiii. <https://doi.org/10.5749/minnesota/9780816680030.001.0001>.

Hendrickx, Iris / Michel Généreux / Rita Marquilhas (2011): «Automatic Pragmatic Text Segmentation of Historical Letters», em Caroline Sporleder / Antal van den Bosch / Kalliopi Zervanou (eds.), *Language Technology for Cultural Heritage. Selected Papers from the LaTech Workshop Series*. Berlim / Heidelberg: Springer-Verlag, 135-153. <https://doi.org/10.1007/978-3-642-20227-8_8>.

Hendrickx, Iris / Rita Marquilhas (2011): «From old texts to modern spellings: an experiment in automatic normalisation», *Journal for Language Technology and Computational Linguistics(JLCL)* 26 (2): 65-76.

Janssen, Maarten (2016a): «TEITOK: Text-Faithful Annotated Corpora», *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016). Portorož, Slovenia*. European Language Resources Association, 4037-43. <http://www.lrec-conf.org/proceedings/lrec2016/pdf/651_Paper.pdf>.

Janssen, Maarten (2016b): «POS tagging and less resources languages individuated features in CorpusWiki», em Zygmunt Vetulani / Hans Uszkoreit / Marek Kubis (eds.), *Human Language Technology. Challenges for Computer Science and Linguistics. 6th Language and Technology Conference, LTC 2013, Poznań, Poland, December 7-9, 2013. Revised Selected Papers*. Cham: Springer, 411-419. <https://doi.org/10.1007/978-3-319-43808-5_31>.

Latour, Bruno (1991): *Nous n'avons jamais été modernes. Essai d'anthropologie symétrique*. Paris: La Découverte.

Mahoney, Michael S. (2005): «The histories of computing(s)», *Interdisciplinary Science Reviews* 30 (2), 119-35. <https://doi.org/10.1179/030801805X25927>.

Marquilhas, Rita (2000): *A Faculdade das Letras. Leitura e Escrita em Portugal no Século XVII*. Lisboa: IN-CM.

Marquilhas, Rita (2009): «'Eu ainda sou vivo'. Sobre a edição e análise linguística de cartas de gente vulgar», *Estudos de Lingüística Galega* 1: 47-65.

Marquilhas, Rita / Iris Hendrickx (2016): «Avanços nas humanidades digitais», em Ana Maria Martins / Ernestina Carrilho (eds.), *Manual de Linguística Portuguesa*, De Gruyter: 252-77. <https://doi.org/10.1515/9783110368840-012>.

McCarty, Willard (2013): «The future of digital humanities is a matter of words», em John Hartley / Jean Burgess / Axel Bruns (eds.), *A Companion to New Media Dynamics*. Chichester, West Sussex, UK / Malden, MA: Wiley Blackwell, 33-52. <https://doi.org/10.1002/9781118321607.ch2>.

Milic, Louis T. (1966): «The next step», *Computers and the Humanities* 1 (1), 3-6. <https://doi.org/10.1007/BF00188010>.

Milic, Louis T. (1990): «The Century of Prose Corpus», *Literary and Linguistic Computing* 5 (3): 203-208. <https://doi.org/10.1093/llc/5.3.203>.

McEnery, Tony / Andrew Wilson (1996/2001). *Corpus Linguistics*. Edimburgo: Edinburgh University Press.

Pierazzo, Elena (2007): *Progetto Epistolario: Manuale di codifica*. Manual inédito. Lucca: Centro Studi Giacomo Puccini.

Schnapp, Jeffrey T. (2014): *Knowledge Design: Incubating New Knowledge Forms / Genres / Spaces in the Laboratory of the Digital Humanities*. Herrenhausen Lectures. Hannover: Volkswagen Stiftung. <http://jeffreyschnapp.com/wp-content/uploads/2011/06/HH_lectures_Schnapp_01.pdf>.

Pound, Scott / Alan Liu (2013): «The Amoderns: Reengaging the Humanities. A Feature Interview with Alan Liu», *Amodern* 2 (October: Network Archaeology), s.p. <http://amodern.net/article/the-amoderns-reengaging-the-humanities/>.

Vanhoutte, Edward / Ron Van den Branden (eds.) (2003): *DALF guidelines for the description and encoding of modern correspondence material Version 1.0*. Gent: CTB-KANTL. <http://www.kantl.be/ctb/project/dalf/dalfdoc/>.

Wallace, R. Jay (2003/2018): «Practical Reason», em Edward N. Zalta (ed.), *The Stanford Encyclopedia of Philosophy* (Spring 2018). <https://plato.stanford.edu/archives/spr2018/entries/practical-reason/>.