[R. & I. Hendrickx (2016). "Avanços nas humanidades digitais". In *Manual de Linguística Portuguesa*, ed. Ana Maria Martins & Ernestina Carrilho. MRL Series. De Gruyter. (Pre-print. Do not cite from this PDF version]

9. Avanços nas Humanidades Digitais

Rita Marquilhas e Iris Hendrickx

Sumário

Neste capítulo acompanham-se os avanços da filologia do português desde que o ambiente digital se começou a anunciar como o contexto mais apropriado para a circulação do conhecimento. Remonta-se às primeiras experiências de processamento mecânico de textos portugueses, quando se entreviam já duas grandes vantagens no auxílio informático para efeitos de estudo histórico da língua: prevenção de erro humano em transcrições e edições e prevenção de abandono de tarefas demasiado gigantescas para a capacidade humana. Acompanha-se uma fase ulterior, em que os académicos, a nível internacional, deixaram de instrumentalizar apenas o digital para passarem a harmonizar-se com ele, tentando compreender quantos conceitos e métodos é preciso revolucionar para que a filologia possa continuar a cumprir a responsabilidade de disciplina que se ocupa da peritagem dos textos e do seu diálogo com a história da cultura e a história da língua. Analisam-se aqueles modelos de edição académica que correspondem, por terem codificação explícita e consistente, ao imperativo da legibilidade por máquina, ao mesmo tempo que permitem, fruto da linguagem de marcação e da anotação rica que adotam, uma crescente manipulação das suas representações computacionais. E demonstra-se como a filologia do português ganhou um ritmo acelerado de experimentação a este nível.

Palavras-chave: filologia, texto, tecnofobia, codificação de caracteres, linguagem de programação, linguagem de marcação, XML, TEI, edição académica digital, corpus histórico, anotação em stand-off, anotação alinhada.

9.1 Introdução

A investigação em linguística, tal como acontece em muitas áreas de investigação científica, procede obrigatoriamente à validação das suas hipóteses junto de dados empíricos. Quando a investigação em causa envolve o estudo histórico de uma língua, falar em dados empíricos é o mesmo que falar em textos de épocas passadas, textos que se procura conhecer de uma forma que não pode envolver a mais pequena cedência a processos de mistificação. Os artefactos (*testemunhos*) que conservam as formas mais genuínas desses textos, não necessariamente os mais antigos, são geralmente manuscritos com o formato de documento ou de códice. Guardam-se em espaços físicos, como os arquivos ou as áreas reservadas das bibliotecas, e só são acessíveis a leitores acreditados. É preciso, portanto, publicá-los, o que corresponde a preparar uma edição rigorosa, comparada e completa, cujo conteúdo serve na composição de glossários, dicionários, gramáticas, antologias e estudos de linguística diacrónica.

A forma de publicação em causa é um tipo de edição tão enriquecida que contém, baseada no que vem escrito nas páginas iniciais das introduções, nos rodapés das notas, ou ainda nas páginas em espelho, uma proposta de leitura empiricamente defensável. Corresponde, em algumas tendências de edição inspiradas no método do alemão Karl Lachmann, àquilo que se julga simbolizar melhor, por aproximação, o conteúdo textual da fonte original (Timpanaro 1981). Em outras tendências, mais influenciadas pelo método do francês Joseph Bédier, corresponde ao conteúdo do testemunho menos modificado por uma história de transmissão textual (Bédier 1970). São edições que recebem o rótulo de *edição crítica*. Ao tipo de peritagem que a elas conduz chama-se *crítica textual*, ou *filologia*, termo que preferiremos aqui.

Desde o século XIX, tem-se experimentado e teorizado sobre a forma mais segura de se fazerem boas edições críticas, se as referidas Lachmanniana ou Bédieriana (Dionísio 2005). A discussão girou sempre em torno da oposição entre atitudes objetivas e subjetivas perante testemunhos históricos que foram herdados, lidos e transmitidos com a intervenção de múltiplos acasos, difíceis de reconstituir e causadores de incessante variação. Discutiu-se, inclusivamente, se a objetividade em filologia seria alguma vez possível. Uma coisa, no entanto, não foi preciso nunca discutir, até finais da década de 1980: o formato e o suporte de publicação para uma edição crítica. Seriam sempre, indiscutivelmente, o formato do livro, o suporte do papel. Esse tempo chegou ao fim, até porque se percebeu logo, ainda antes da explosão da internet, que a informática ia ser instrumental num novo tipo de edição crítica, que passava agora pelo "elogio da variante", título de um trabalho de Bernard Cerquiglini em que se reconhecia o determinismo da tecnologia impressa nos extremismos filológicos assumidas no passado (Cerquiglini 1989).

Neste capítulo, apresentaremos uma síntese das transformações resultantes da entrada da filologia portuguesa num mundo em que o ambiente mais natural para a circulação do conhecimento é precisamente o ambiente digital. Foi um processo que, como é inevitável em investigação científica, esteve sempre integrado em experiências internacionais, pelo que em certos aspetos seria artificial separá-lo de tal contexto. Por outro lado, foi um processo experimentado, com crescente solidariedade, por todas as áreas do conhecimento que interrogam textos e memórias coletivas, as quais são, principalmente, estas três: i) os estudos literários, porque veem nos textos a manifestação da capacidade humana de criação artística, ii) a história, porque se documenta nos textos por serem testemunho das sociedades e das culturas no tempo, e iii) a linguística histórica, porque toma os textos escritos como realizações antigas das línguas naturais, impossíveis de documentar por outra forma. Na sua faceta de *peritagem dos textos*, a filologia funciona hoje como disciplina auxiliar de qualquer destes estudos. Com efeito, cada filólogo é normalmente, também, ora um historiador da língua, ora um estudioso da literatura e da cultura. O próprio termo *filologia*, antes da

autonomização disciplinar oitocentista que levou à distinção entre linguística, história da literatura e crítica textual, serviu para cobrir todas elas.[1]

Contudo, a partilha do método filológico, no passado, serviu sobretudo de pretexto para se definirem as diferenças entre as modalidades de edição que as três áreas diferentemente exigiam. Ora a adoção de métodos digitais e a criação de recursos levou a que as "paredes" entre aquelas áreas, numa dinâmica exemplar de interdisciplinaridade, começassem a cair (McCarty 2005, 118). Buscou-se, inclusivamente, um termo suficientemente genérico que a todas englobasse, na sua faceta de adesão às práticas computacionais e à exploração da publicação em linha. Foi busca que parece ter terminado quando, em 2001, se cunhou a expressão *Humanidades Digitais* (Kirschenbaum 2010, 2-3).

Em termos de problemas, o mais desafiante que as Humanidades Digitais têm encontrado é o de perceberem que certas metodologias e conceitos que tomavam como adquiridos se revelam, afinal, incomodamente dependentes de simplificações criadas pelas tradicionais culturas do manuscrito e do impresso. É portanto necessário embarcar numa aventura crítica que se espera venha ajudar a entender melhor o lugar do texto, e da comunicação escrita em geral, dentro das lógicas da sociedade tradicional, da sociedade moderna e da nova sociedade da informação.

9.2 Os começos da filologia digital

9.2.1 Nota sobre filologia portuguesa tradicional

A tarefa de dedicar cuidados filológicos a textos portugueses foi inicialmente cumprida pelos fundadores Gonçalves Viana, Epifânio da Silva Dias, Adolfo Coelho e Jules Cornu e prosseguida por Leite de Vasconcellos, Carolina Michaëlis de Vasconcelos, Júlio Moreira, José Maria Rodrigues, José Joaquim Nunes, Manuel Said Ali, David Lopes, Álvaro da Silveira, Cláudio Basto, João da Silva Correia e Manuel Rodrigues Lapa, todos nascidos ainda no século XIX. Consolidaram-na Joseph Piel, Paiva Boléo, Serafim da Silva Neto, Paul Teyssier, Celso Cunha e Lindley Cintra, para registar apenas os que nasceram no primeiro quartel do século XX (para a maioria das suas biografias e publicações, cf. Prista/Albino 1996).

O propósito da ladainha de nomes, praticamente exaustiva, é o de demonstrar que, ao longo de mais de cem anos, esteve toda uma pequena elite "livresca" encarregada da preparação de materiais eminentemente didáticos (antologias, gramáticas, dicionários, glossários, histórias da língua portuguesa), baseados em edições críticas que os mesmos, ou pares muito próximos, elaboravam. A sua matéria empírica era constituída por manuscritos e impressos antigos, que tinham cuidadosamente escrutinado dentro de exemplares de acesso muito restrito. O sistema pressupunha uma escala hierárquica: no topo, as instituições (e alguns colecionadores privados), guardiões físicos dos artefactos

[1] Foi August Schleicher o primeiro a sugerir, no século XIX, que a *linguística* era uma ciência natural e a *filologia* (ainda no sentido de estudo dos textos herdados, sobretudo os literários) um tipo de história (Timpanaro 2005, [1]1963).

que continham a memória textual da língua nacional na forma mais genuína, e na quantidade mais numerosa, que se conseguiu conservar. As credenciais para aceder a esse nível de topo estavam reservadas a académicos com credibilidade suficiente para lhes ser confiado o manuseio das relíquias. O seu papel era o de converter a credibilidade que as instituições lhes reconheciam, bem como o privilégio de acesso a objetos valiosos, em publicações com estatuto de autoridade. A autoridade não era, contudo, automática: discutia-se em correspondência com os pares, transbordava para artigos de revistas científicas, e, uma vez aprovada, acabava por atingir os escalões de base, o do ensino, sobretudo o secundário e o universitário. De permeio, atuavam as casas editoras (as imprensas nacionais, portuguesa e brasileira, as imprensas de Universidade, de biblioteca ou de fundação), nada alheias ao sistema porque se tratava normalmente de publicações subsidiadas. Eram, precisamente, edições dispendiosas, sem grande correspondência entre o investimento que envolviam e o público que as consumia. Mas eram tidas como necessárias por se entender, desde que a ideologia nacionalista de Oitocentos tinha triunfado, que cabia ao Estado velar pela repetida celebração da herança cultural da Nação, cristalizada nos seus diferentes monumentos, incluindo os textos históricos e os textos literários. Este panorama ainda se mantinha quando a informática e a cultura se começaram a cruzar, mas tornou-se claramente desajustado num mundo de onde já parece ter desaparecido o problema da conservação da memória textual, para dar lugar ao da sua sobre-representação.

9.2.2 O começo da filologia portuguesa digital

O início da filologia digital remonta à apropriação da tecnologia informática no sentido de se conseguir o processamento de textos de grandes proporções, considerados representativos de uma cultura através do léxico da sua língua. Chamou-se aos produtos desse trabalho, sintomaticamente, *Tesouros*. Veja-se como o caso se passou em Portugal, entre as décadas de 1960-1980.

As primeiras experiências resultaram da colaboração entre a linguista Maria Helena Mira Mateus e o Centro de Cálculo Científico da Fundação Calouste Gulbenkian (Mateus 1968, 227). A preocupação era lexicográfica e filológica. Pretendia-se transpor para o caso português o que se praticava em França com a construção do *Trésor de la Langue Française*,[2] i.e., "elaborar um dicionário histórico que desse, em períodos sucessivos, um quadro tão completo quanto possível do vocabulário da língua desde a origem aos nossos dias" (Mateus 1974, 3).

O *Tesouro* português começou a ser construído com o *Glossário da Vida e Feitos de Júlio César, tradução portuguesa quatrocentista de Li Fet des Romains* e o resultado

[2]Em 1960, três anos depois de o projeto ter sido pensado, o CNRS decidiu criar em Nancy, e com direito a edifício novo, o *Centre pour un Trésor de de Langue Française*. O *Trésor* foi publicado primeiro em papel, em 16 volumes saídos entre 1971 e 1994. Em 2002 saiu o *Trésor de la Langue Française Informatisé*, nas versões CD-ROM e em linha (http://atilf.atilf.fr/) (Mateus 1974, Del Mancino/Pierrel 2009).

foi a publicação da obra em papel, primeiro em sucessivos números do *Boletim de Filologia*, entre 1974 e 1992, cobrindo o léxico das letras *A* a *S*, depois em edição integral, numa publicação encadernada (Mateus, 2010).

Na verdade, a intervenção de matemáticos, informáticos, programadores (e máquinas) do Centro de Cálculo Científico consistiu, estritamente, num auxílio externo, destinado a tornar mais eficiente um processo que podia ter também realização manual (Mateus 1974, 5):

> É evidente que a elaboração de um Tesouro da Língua se não restringe à etapa (informática) indicada. Mas é evidente, também, que esta etapa, vital em relação à essência da própria obra, se não poderá realizar com eficiência sem esse importante contributo.

À época, porque ainda não havia microcomputadores, o termo "computador" era usado em itálico, havia necessidade de o descrever recorrendo à perífrase "complexo mecanográfico que inclui o computador como unidade principal", e não se hesitava em encará-lo enquanto "instrumento", se bem que a noção da sua progressiva indispensabilidade para a investigação já fosse óbvia: "[tem-se] tornado indispensável na coadjuvação do esforço do homem para o progresso científico". Podia-lhe ser entregue, no caso do estudo do português, "a tarefa de evidenciar, alfabetizar e organizar todos os vocábulos, não de um único texto, mas de milhares de obras consideradas representativas (da) língua" (Mateus 1974, 5)

O *Tesouro* não se fez na altura: só o *Glossário da Vida e Feitos de Júlio César* ficou pronto. Em termos de descendência, o recurso em linha que atualmente mais se aproxima dos objetivos do *Tesouro* pertence à área da lexicografia. Trata-se do *Corpus Lexicográfico do Português*, da Universidade de Aveiro e do Centro de Linguística da Universidade de Lisboa, da responsabilidade de Telmo Verdelho e João Paulo Silvestre. Foi lançado em 2003 e disponibiliza, em modelo crescentemente pesquisável dentro do sistema DICIweb, a dicionarística portuguesa dos séculos XVI a XIX (http://clp.dlc.ua.pt/DICIweb/).

O trabalho que se seguiu à experiência de Maria Helena Mateus decorreu no início da década de 1980, empreendido pelo lusitanista britânico Stephen Parkinson, que lançou o AOPT, *Achive of Old Portuguese Texts* (Parkinson, 1983). Parkinson contava com a colaboração do *Computing Centre* da Universidade de Aberdeen e com a máquina Honeywell H66, que já usava a fita magnética em substituição dos cartões perfurados. Já estava também disponível um programa de concordâncias, o *Oxford Concordance Program* (Hockey/Marriott 1980) e um editor para conversão automática de texto.

Aparecia claramente a Parkinson a grande vantagem que já acima referimos, ao mencionar Cerquiglini, e que se tornou depois banal para todos os filólogos que praticam codificação de texto para publicação eletrónica: deixou de haver razões para a aplicação de normas de edição irreversíveis, sempre de difícil coerência interna. Eram normas que se tinham usado no passado porque era preciso assegurar que as edições mantivessem alguma legibilidade do ponto de vista do leitor comum. No caso português, um exemplo que se tornou clássico foi decidido por um grupo de trabalho formado em 1970 no Centro de Linguística da Universidade de Lisboa e dirigido por

Luís Filipe Lindley Cintra. Inclui regras de conversão deste tipo, relativo à transcrição das *vogais* nos manuscritos medievais (Castro/Castro/Cepeda/Madureira 1973, 418):

> [vogais] *não etimológicas*: são sempre transcritas como vogais simples, mesmo que, para alguns textos, se possa suspeitar que, na intenção do escriba, a geminação representava abertura ou tonicidade da vogal. (Em casos destes, o facto será devidamente registado na Introdução.)
> Ex: *taes* por *taaes*, *ó* ou *oh* por *oo*, *doe* por *dooe*, *ceo* por *ceeo*, *som* por *soom* (<SUM).

A tecnologia veio permitir que se passasse a proceder de outro modo, como foi experimentado e relatado por Parkinson. Perante um manuscrito antigo e linguisticamente relevante, passou a ser possível progredir por etapas. Primeiro, procedia-se à recolha de dados crus, para usar um anacronismo, já que a expressão da altura era "transcrição diplomática pré-editada". Para o problema das abreviaturas, por exemplo, havia a solução dos símbolos compostos, "sequências de símbolos que o computador interpretará como símbolo único". A sua resolução era indicada à parte, numa lista de regras de expansão que podia mudar consoante o resultado editado que o filólogo pretendesse.

Na notícia que Parkinson deu deste trabalho, em artigo do *Boletim de Filologia*, o autor reconhecia a inspiração retirada de trabalhos pioneiros, ensaiados para o alemão medieval (Murdoch, 1971) e expunha a virtude da fidelidade e da reversibilidade nas edições assistidas por computador (Parkinson 1983, 242-243):

> [É] possível dar instruções para a conversão de quaisquer símbolos ou sequências de símbolos em quaisquer outros símbolos ou sequências, inclusive os "control characters". Neste poder transformacional reside uma solução para o problema de transcrição de textos medievais. O editor de textos medievais teve sempre que escolher uma forma de transcrição, bem cônscio de que esta decisão afastava um ou outro sector do público do documento. Numa edição modernizadora, aliás, os efeitos da decisão eram permanentes e irreversíveis porque não seria possível deduzir as formas manuscritas sem o apoio de aparatos críticos. O poder transformacional do computador abre caminho para uma prática editorial reversível. (...) O computador fará uma tal tradução de transcrição diplomática para edição com abreviaturas desenvolvidas com muito mais fidelidade que um amanuense.

Os exemplos aqui evocados, dos trabalhos desenvolvidos por Mateus e por Parkinson antes da generalização dos computadores pessoais, demonstram claramente as duas grandes oportunidades que a filologia entreviu no recurso ao auxílio informático. Eram as oportunidades de prevenir a hipótese de erro humano em tarefas de escrita (ao longo de transcrições e edições) e de prevenir a hipótese de abandono de trabalhos demasiado pesados para a capacidade humana, ainda que teoricamente necessários do ponto de vista da investigação filológica. Um *robot*, na sua aceção etimológica de "escravo",[3] poderia fazê-los.

[3] O termo robotics pertence ao campo semântico do trabalho porque começou no empréstimo ao próprio inglês do termo checo robotnik 'escravo'.

O destino da experiência de Stephen Parkinson já foi, ao contrário do que acontecera com o trabalho de Maria Helena Mateus, o de um armazenamento digital. Os documentos tabeliónicos medievais que editou passaram a integrar o primeiro corpus histórico de textos portugueses, o CIPM ou *Corpus Informatizado do Português Medieval*, constituído a partir de 1994 no Centro de Linguística da Universidade Nova de Lisboa, precisamente em torno do resultado do trabalho de Parkinson e com a sua colaboração (Xavier/Brocardo/Vicente 1995). Hoje são pesquisáveis em linha quer no site do próprio CIPM (http://cipm.fcsh.unl.pt), quer no *Corpus do Português*, já com anotação morfossintática, (http://www.corpusdoportugues.org/).

Entretanto, na transição entre as décadas de 1980 e 1990, as condições mudaram. Surgiram dois desenvolvimentos que levariam ao abandono deste tipo de experiências individuais de codificação dirigidas para uma só língua. Aconteceu que a codificação de texto e a codificação de caracteres se tornou numa experiência de crescente dimensão coletiva, da qual muitas diferentes filologias puderam começar a beneficiar: criaram-se o Unicode, de iniciativa empresarial, e o TEI, de iniciativa académica. Antes de abordarmos tais codificações, convém contudo analisar as consequências desta primeira fase da informatização nas Humanidades.

9.2.3. A tecnofobia e a crítica digital

Referimos na secção anterior que a conversão de materiais filológicos num formato mecanicamente legível, bem como a correspondente constituição de bases de dados, resolveu dois problemas clássicos inerentes à atividade da edição crítica de textos: o dos erros humanos e o das limitações humanas. Teoricamente, uma máquina bem programada não se engana, uma máquina bem mantida não se cansa.

Acontece que estas mesmas duas virtudes da mecanização (o automatismo das operações e a viabilização dos empreendimentos gigantescos) é também o nó de um problema cultural e social, inevitavelmente criado quando há abandono de atividades manuais mediante a adoção da tecnologia. É o problema da "proletarização", muito utilizado pelo ceticismo antitecnológico, o que não quer dizer que não seja um problema real.

A posição mais cética em relação ao trabalho em Humanidades Digitais aponta-lhe o seguinte: é trabalho que parece levar apenas a poupar horas e recursos, mas não leva a nada de novo ao nível da construção do conhecimento. O próprio Roberto Busa, reconhecido por muitos como o pai das Humanidades Digitais, discutiu tal perversão (Busa 2004).

Sociologicamente analisado, o perigo parece ser o de se estar a proletarizar uma atividade cultural (Stiegler 1998, Robertson 2015). Proletarização tem aqui vários dos sentidos que Marx atribuiu ao termo na sua crítica à ligação entre capitalismo e industrialização nas sociedades ocidentais. O paralelismo é o seguinte: tal como o produtor tinha sido afastado da terra, seu antigo meio de produção, e obrigado a entrar como operário não especializado na engrenagem industrial, também o académico se

afastou do conhecimento com a chegada da digitalização às Humanidades, sendo arrastado para um trabalho mecânico que ele, académico, só pode desenvolver a níveis pouco especializados já que não tem uma preparação de engenheiro. Ao mesmo tempo, o engenheiro informático que se ocupa da infraestrutura da edição académica (ou do corpus histórico, ou do museu virtual) vai lidar com temas de cultura ao nível da modelização dos programas, só que não tem preparação humanística para tanto. O resultado combinado será, argumenta-se em setores mais conservadores da esfera pública, uma perda para a língua, a história, a literatura, a cultura, enfim, para o conhecimento.

Esta é também a linha da tradicional crítica à técnica, que tem origem num tipo de ressentimento com a mesma antiguidade da própria filosofia. Nas palavras de Bernard Stiegler, "desde a sua origem e até agora, a filosofia tem reprimido a técnica como objeto de pensamento. A técnica é o não pensamento" (Stiegler 1998, [1]1994, ix). Contudo, como aponta o mesmo autor, precisamos muito pragmaticamente de uma rápida reação a tal repulsa, por maior tradição que ela tenha: "A mudança de perspetiva e de atitude torna-se necessária, obrigando a uma capacidade de reação tão urgente quanto inevitável" (Stiegler 1998, [1]1994, x).

A reação mais urgente, que é precisamente a que estão a assumir as Humanidades Digitais, é a de compreender a máquina, investindo na busca de uma solução para a falta de harmonia entre cultura e técnica. Tal falta de conciliação nasce sempre que se olha para as máquinas enquanto meros substitutos do indivíduo-artesão, "portador de ferramentas". Supera-se vendo nelas "indivíduos técnicos", que é preciso conhecer mediante a construção de uma "mecanologia" que detete até que ponto a máquina vai ganhando "capacidade de regulação" (Stiegler 1998, [1]1994, 69, retomando ideias de Gilbert Simondon).

Na primeira fase de informatização da filologia, o computador era simplesmente instrumentalizado. Pedia-se-lhe para, repetindo o que escreveram Maria Helena Mateus e Stephen Parkinson, "evidenciar, alfabetizar e organizar vocábulos", ou então para fazer a conversão de símbolos "com muito mais fidelidade que um amanuense". Na segunda fase, estão a ser os próprios académicos a harmonizar-se com o digital, a envolver-se nele de forma a compreenderem quantos conceitos é preciso revolucionar para que a filologia possa continuar a cumprir a responsabilidade de disciplina que se ocupa da peritagem dos textos e do seu diálogo com a cultura e a língua.

Ao mesmo tempo, este envolvimento dos académicos com a tecnologia dentro do movimento das Humanidades Digitais é um processo que equivale a resistência. Resistência a duas forças: a da vetusta tradição académica e a da ameaçadora agressividade empresarial. Por um lado, as Humanidades Digitais caracterizam-se por uma cultura que, reconhecidamente, envolve "colaboração, abertura, relações não hierárquicas e agilidade" (Kirschenbaum 2010, 5), por contágio das práticas entretanto naturalizadas por programadores informáticos e por frequentadores da blogosfera. Por outro lado, um maior protagonismo daquele movimento significará que o controlo da

edição digital não fica tacitamente entregue às empresas, enfrentando a crítica textual uma nova missão, de se converter também em *crítica digital*. Se os filólogos tivessem olhado mais cedo para a constituição de bibliotecas e livrarias em linha, argumenta Jerome McGann, a *Google Books* prestaria hoje um serviço muito diferente, documentalmente mais responsável (MacGann 2014, cap. 7).

9.3 O conceito de texto na esfera digital

Como foi referido anteriormente, uma das grandes vantagens que a revolução digital trouxe a qualquer área do conhecimento que lide de perto com textos, cultura e memória foi a de ter oferecido instrumentos valiosos no auxílio à reflexão sobre "o que é um texto?". Na formulação de Maria Clara Paixão de Sousa, que é, na área da filologia portuguesa, e com João Dionísio (Dionísio 2005), quem mais se tem dedicado a refletir sobre o tema, "a difusão digital exige, [pelo] menos, transformações profundas nas nossas perspetivas conceituais sobre o texto" (Sousa 2013a, 20, cf. também Sousa 2013b).

O termo *texto* está carregado de conotações antigas, associadas às tecnologias da escrita e à noção de autoridade intelectual, bem como de conotações modernas, ligadas à história recente da defesa do direito de autor. Sintomático da sua polissemia é o facto de as mesmas palavras, por exemplo, *obra*, *discurso*, *mensagem*, *enunciado*, serem explicadas muitas vezes como equivalentes de *texto*, e igual número de vezes como distinguindo-se dele.

O uso comummente dicionarizado é o de "palavras fixadas pela escrita", que acaba por ser também recolhido por alguma bibliografia didática de filologia (Roncaglia 1975, 23, Blecua 1983, 17). Mas tanto na mesma área da filologia, como em teoria literária e em linguística do texto, o termo é visivelmente incómodo, levando certos autores a conduzir ensaios inteiros só para o definirem e outros a optarem por o retirar da terminologia mais precisa.

Roland Barthes, num conhecido e influente ensaio intitulado *De l'oeuvre au texte*, alongou-se na busca das melhores metáforas para explicar a complexa distinção entre texto literário – um "campo metodológico", um "espaço social", um "plural irredutível" – e obra literária – "um fragmento em substância [que] ocupa uma porção do espaço dos livros (numa biblioteca, por exemplo)" (Barthes 1994, [1]1971).

Paul Ricoeur, aquele cujo esforço de definição é porventura dos mais citados, entregou-se ao paradoxo: explicou texto como aquilo que é um discurso não oral – "discurso fixado pela escrita" –, precisando, no entanto, que se trataria de um discurso que vivia "no ar", "fora do mundo" ou "sem mundo" (Ricoeur 1986, 154 e 158).

Esta contradição entre a materialidade inevitável de tudo quanto seja "fixado" e a imaterialidade daquilo que vive "fora do mundo" diz bem da dificuldade em explicar um termo que remonta a Quintiliano, autor que escolheu dar um sentido figurado à designação latina de "tecido" (*textus*), referindo-se às palavras que os humanos

compõem enlaçando-as num tecido ("verba qua compositione in textu iungatur" (*apud* Roncaglia 1975, 23).

Em linguística do texto, uma disciplina que, tal como a crítica textual, recusa *a priori* a equivalência entre texto e texto literário, a questão da materialidade é posta em termos de *canal* ou *modo*. Tal facto decorre de se conceber, aqui, que os textos tanto podem pertencer à oralidade como à escrita. São tomados como instância de comunicação: situacionalmente, podem adotar um modo que pode ser transiente (tipicamente, o falado), ou permanente (tipicamente, o escrito). Numa das abordagens mais difundidas, fala-se menos de texto e mais de *textualidade*, uma propriedade das expressões linguísticas que se agregam de forma a apresentarem coesão, coerência, intencionalidade, aceitabilidade, informatividade, situacionalidade e intertextualidade (Beaugrande/Dressler 2005, [1]1972, 46). Outra abordagem formula a existência de um campo abstrato que os textos ocuparão, campo esse que é cruzado por uma oposição de meio, ou modo, (oralidade *versus* escrita) e por uma escala espacial (da extrema proximidade à extrema distância comunicativa). Pretende-se assim dar conta do fenómeno da variação textual, com a conversa face a face num dos polos, e a obra literária canónica, no outro. Em qualquer ponto do campo pode surgir um género textual diferente, com outras tantas idiossincrasias (as *tradições discursivas*) que se manifestam em todos os níveis de descrição da língua (Koch/Oesterreicher 2007, [1]1999).

O papel da materialidade na constituição dos textos é, por conseguinte, um fator que reúne pouca unanimidade, suscitando diferentes atitudes para o considerar ou desconsiderar intelectualmente dentro da reflexão teórica sobre o texto. Mas quando a informação passou a circular digitalmente, uma coisa se tornou bem clara. Percebeu-se que nem se pode dizer que os textos são desprovidos de matéria específica, nem se pode considerar que eles são constituídos por matéria específica. A componente da materialidade está instalada nos textos, sim, mas ocupa um nível abstrato, "subespecificado" no sentido de latente e imprevisível, para usar termos da Fonologia teórica (cf. Steriade).[4] Tornou-se mais claro como a atividade textual envolve a nível abstrato, para quem fala e escreve, uma conceção da materialidade dos seus textos que só deixa de ser polivalente quando há convergência de mais planos, i. e., quando o plano físico e o interpretativo se vêm juntar ao concetual. Referindo-se ao que se aprende sobre texto na análise dos textos literários já nascidos-digitais, N. Katherine Hayles resume (traduzido de Hayles 2004, 72):

> Nesta perspetiva de materialidade, ela não é uma coleção inerte de propriedades físicas; é antes uma qualidade dinâmica que emerge da articulação entre o texto enquanto artefacto físico, o seu conteúdo concetual e as atividades de interpretação de leitores e escritores[…]. A materialidade não pode ser especificada à partida.

O caráter subespecificado, emergente e dinâmico da materialidade dos textos estava intuído nos trabalhos de Donald F.McKenzie, publicados nos anos 1970 e 1980, quando

[4] O autor explica também os diferentes sentidos que o termo *subespecificação* (de traços fonológicos) pode ter.

pugnou por uma filologia do livro impresso enquanto *bibliografia histórica*, ou *sociologia do texto*. Referindo-se à edição setecentista das peças reunidas do dramaturgo inglês William Congreve (*Works*, 1710), McKenzie proclamava (traduzido de McKenzie 2002, [1]1977, 200):

> O livro em si [Works 1710] é um meio de expressão. Aos olhos, as suas páginas oferecem uma agregação de significados, tanto verbais como tipográficos, a serem traduzidos para o ouvido [na encenação da peça]; mas temos de aprender que a forma que ele toma na nossa mão também nos fala do passado. A explicação plena desses significados, em toda a sua riqueza contextual, é a principal função textual da bibliografia histórica.

Concluía McKenzie que "qualquer ênfase na estrutura integrada de um texto é portanto bem vinda"; para o conjunto integrado da edição textual participavam, além de todos os "detalhes" e "defeitos", as "versões diferentes", as "formas não verbais", os "comportamentos de leitura" e as "decisões históricas feitas por autores, desenhadores e artesãos" (McKenzie 2002, [1]1977, 223).

As conceções de McKenzie, que pertenciam sobretudo à defesa de uma modalidade mais culturalmente empenhada da prática da filologia, revelam-se penetrantes porque no caso dos textos que o filólogo analisava, saídos da baixa tecnologia da imprensa manual, o caráter subespecificado da materialidade não estava tão obviamente presente. Os artefactos manuscritos e impressos são, com efeito, muito sólidos na sua aparência física, e facilmente se cria a ilusão de que suscitam representações idênticas em qualquer pessoa, momento ou lugar da história. No mundo da comunicação digital, a subespecificação da materialidade, pelo contrário, só não se torna óbvia para os infoexcluídos profundos. Qualquer variação de software ou de hardware, qualquer caminho diferente que se tenha seguido na navegação, qualquer histórico que tenha ficado na memória dos seus computadores pessoais vai afetar a forma como um texto se materializa para leitores e escritores, que de qualquer forma também variam entre si na agilidade com que se adaptam a novidades e *upgrades*.

A lição para os filólogos da edição digital tem sido dupla: não podem sonhar com uma edição crítica estável, até porque ela é impossível; mas podem lidar de frente com tal impossibilidade. Primeiro, harmonizando convenções entre si, para que o grau de subespecificação material das suas edições (que são, elas próprias, textos) não seja tão elevado quanto a tecnologia o permite. Depois, convertendo as edições em abordagens integradas, onde se codifique o que no texto é físico, o que é concetual e o que é social. Um seguidor da *sociologia dos textos* de McKenzie, Jerome McGann, foi precisamente o pioneiro deste tipo de experiência, quando lançou a edição académica digital das obras plásticas e literárias de Dante Gabriel Rossetti (http://www.rossettiarchive.org).

Nas secções que se seguem procuraremos demonstrar a viabilidade técnica destas edições. O argumento é o de, como expôs Wilard McCarty, ser preciso construir modelos de edição que correspondam ao imperativo da legibilidade computacional, pelo que têm de ser completamente explícitos e consistentes. Por outro lado, tem de se

aceitar que serão edições sempre manipuláveis, por ser isso mesmo que acontece nas representações computacionais (McCarty 2004).

9.4. Questões de codificação e de marcação

9.4.1 Codificação e linguagens de programação

Em termos de linguagem, os computadores apenas conseguem lidar com código binário: trata-se de unidades básicas de computação de informação, ou bits (*binary digits*), que são ligadas e desligadas. Representam-se numa série de 0s e 1s, que correspondem frequentemente em termos físicos aos sinais elétricos de alta e de baixa voltagem que dão entrada no computador. Todos os dados que aí se contenham, quer se trate de programas quer de conteúdo, são representados desta maneira, em sequências de cadeias binárias.

Em termos de estrutura interna, o núcleo do computador é constituído pela sua unidade de processamento central, ou CPU (acrónimo de "Central Processing Unit"), aquela parte que executa o código-máquina. Código-máquina, por sua vez, é a linguagem de programação mais primitiva de todas, uma linguagem em que se codificam as instruções para o computador; não é legível por humanos, portanto a maior parte dos programas de computador (o chamado *software*) está escrita em linguagens de programação de nível superior. Exigem uma ulterior compilação, ou interpretação, em código-máquina para poderem então ser executadas. Pode dizer-se, portanto, que uma linguagem de programação é nada mais nada menos do que uma linguagem formal na qual se podem escrever instruções para um computador executar.

São linguagens que existem aos milhares, variando entre as que se destinam a tarefas específicas, as que servem todos os propósitos e as que são independentes do componente físico, ou "ferro", o *hardware*. Compreensivelmente, interessam aos programadores de Humanidades Digitais só aquelas linguagens que permitem manipular conteúdos textuais, as quais variam em termos de sistemas operativos. Em computadores do tipo Unix, i.e., computadores que têm como sistema operativo o Linux, o BSD, ou mesmo o Mac OS X, há uma parte das suas ferramentas básicas que é muito apropriada à manipulação de texto. Com uns poucos e simples comandos Unix, podem-se extrair, por exemplo, listas de frequências de todas as diferentes ocorrências (tokens) de uma palavra, ou de todos os diferentes tipos de palavras numa coleção de textos.

Já para computadores com outros sistemas operativos, ou para tarefas mais complexas, usa-se muito uma linguagem de programação para manipulação de texto chamada Perl (Wall/Christiansen/Orwant 2004), que foi desenvolvida em finais da década de 1980. Em Perl podem-se empregar expressões regulares muito sofisticadas, sendo que uma expressão regular é uma sequência padronizada de caracteres que permite encontrar e substituir (*find and replace*) cadeias de sinais e de palavras. A máquina de expressões regulares do Perl está hoje reimplementada em linguagens de programação mais

recentes, como é o caso das linguagens Python e Java. Na Tabela 2 da secção 9.4.2, abaixo, mostraremos um exemplo simples de um programa escrito em Perl.

Como já referimos, há uma discrepância grande entre as cadeias binárias que um computador consegue entender e os textos com a escrita própria de uma língua natural. Para que estes últimos possam ser trabalhados por uma máquina, os caracteres legíveis por humanos têm de ser transformados em cadeias binárias e vice-versa. Tal passo de conversão recebe o nome de codificação de caracteres e implica a tradução de bits para caracteres. Um dos mais antigos sistemas de codificação de caracteres foi o que se usou na década de 1960 com os primeiros computadores: chama-se ASCII, *American Standard Code for Information Interchange*. O código ASCII tem espaço para 27 bits, o que lhe permite codificar até 128 caracteres diferentes, incluindo um conjunto de caracteres de controlo, tais como os que codificam "tabulação" e "nova linha". É o suficiente para codificar o alfabeto que se emprega na escrita do inglês, mas é recurso demasiado escasso quando se pretende abranger todos os sistemas de escrita do mundo.

O facto gerou problemas à medida que, na era da computação, se foi assistindo à criação de múltiplos sistemas de codificação alternativos, por parte de produtores em países de língua não inglesa. Viu-se um exemplo de tal iniciativa quando, acima, se referiu a criação, por parte do lusitanista Stephen Parkinson, de um sistema de combinações de símbolos destinadas à codificação de abreviaturas em textos medievais portugueses. Com experiências desse tipo, a troca de programas e de ficheiros entre diferentes computadores foi-se tornando crescentemente difícil, o que conduziu, na década de 1980, à definição de um novo padrão unificador, o Unicode. É um sistema descritivo para codificação de caracteres que cobre, atualmente, mais de 100.000 caracteres diferentes, desde o dos hieróglifos do Antigo Egito até aos mais variados ícones, incluindo os *emoticons* nossos contemporâneos. Note-se que o Unicode é um padrão descritivo, pelo que o valor final de cada carácter, i.e., o seu corpo, fonte e estilo, ainda é atribuído pelo programa que o interpreta, seja ele o sistema de programas do navegador da internet (*web browser*) ou o de um editor de texto.

A implementação de Unicode que hoje mais se usa é a codificação em UTF-8, que é compatível com ASCII. Para o inglês, codifica os caracteres em códigos ASCII de 7 bits, mas recorre-se a mais bits para o caso de outros caracteres. Na perspetiva das Humanidades Digitais, o padrão Unicode tem uma série de inconvenientes, como foi já salientado, por exemplo, por Fiormonte (2012). Há que ser-se cuidadoso e crítico em virtude de se tratar de um padrão industrial orientado para as necessidades do Ocidente, sobretudo do Ocidente moderno, o que o torna potencialmente desajustado para muitos sistemas de escrita. São limitações de que estão bem cientes muitos investigadores da área das Humanidades Digitais, que prosseguem, aliás, com a criação original de tipos para computador quando as suas edições o exigem. É o caso, em Portugal, do *Notator Mono*, um tipo medieval para computador criado para permitir edições diplomáticas com "uma representação tipográfica rica, complexa e fidedigna dos sistemas de escrita usados em Portugal entre os séculos IX e XIV na produção de documentação notarial" (Emiliano 2005, 139). A título de exemplificação, reproduz-se abaixo na Tabela 1 o

caso da representação em *Notator Mono* das letras traçadas por traços verticais e oblíquos previstas para todas as combinações possíveis entre letras minúsculas e sinais de -*UM*, uma terminação muito frequente em palavras latinas.

CAR#	DESIGNAÇÃO	VALOR
#156	M minúsculo traçado	-m (um)
#159	N minúsculo traçado	-n (um)
#168	R minúsculo redondo traçado	-r (um)
#171	R minúsculo traçado	-r (um) [carácter da letra visigótica]
#178	T minúsculo visigótico traçado	-t (um) [carácter da letra visigótica]
#179	T minúsculo traçado	-t (um)

Tabela 2 Representação no tipo *Notator Mono* das abreviaturas de terminações latinas em *-mum*, *-num*, *-rum* e *-tum* (reproduzido de Emiliano 2005, 152)

A codificação de caracteres mantém-se um obstáculo problemático em muitos projetos de Humanidades Digitais. Surge em dois contextos. Por um lado, quando se quer retomar documentos digitais já antigos, codificados em formato pré-Unicode, como é o caso de muitas edições preparadas em computador nas décadas de 1960, 1970, 1980 e mesmo ainda em 1990, cuja codificação não é corretamente reconhecida pelos programas atuais, além de que não se converte facilmente para outros formatos. Por outro lado, quando se coligem documentos de proveniências diferentes, é sempre possível que eles tragam codificações variadas. É essencial, por conseguinte, criar uma versão estandardizada em que tal variação desapareça, mas nem sempre o responsável pela coleção consegue controlar o formato exato que os seus documentos digitais receberam na origem. No momento de recorrer a dados textuais fornecidos por voluntários, por exemplo, pode acontecer que alguns dados tenham sido obtidos por cópia a partir de uma qualquer aplicação, seguida de colagem num editor de texto, sendo que o documento pode ser enviado já no formato desse programa de edição, sem consciência, muitas vezes, da codificação que se pode ter perdido ou corrompido no momento da cópia.

A propriedade da subespecificação, dinamismo e fluidez material dos documentos e programas feitos em computador é portanto motivo para que na prática das Humanidades Digitais haja um alerta constante em relação a problemas de codificação e um investimento crescente em técnicas para a prevenção dos mesmos.

9.4.2 Marcação

Os documentos de texto digitais têm, para além do seu conteúdo codificado concreto, uma camada adicional de informação sobre o texto em si e respetivo contexto, i.e., sobre a fonte arquivística, o género, o autor, a data, as revisões, etc., bem como sobre o formato geral que a visualização deve ostentar quando publicada, em pormenores como os das fronteiras de parágrafos, das cores, do corpo da letra, etc.

O procedimento mais usual que se adota para indicar esta informação adicional é o da marcação de texto (*textual markup*), o que implica usar uma linguagem de marcação. É, tal como as de programação, uma linguagem formal, mas tem propósitos específicos. Enquanto uma linguagem de programação se usa para formular instruções para computador, uma linguagem de marcação usa-se para anotar um documento com informação extra, dando continuidade à tradição da inclusão de símbolos descritivos e de notas de aparato na publicação em papel de edições diplomáticas, críticas e genéticas (Roncaglia 1975, 75–78, Blecua 1983, 147–152, Castro 2001). Com efeito, é na camada da marcação de uma edição crítica eletrónica que se inclui a indicação de lacunas, rasuras ou acrescentos no original do texto que se edita, conjeturas de editor ou existência de variantes dentro de uma tradição textual.

Marcação significa emprego de etiquetas para distinção efetiva, a níveis diferentes, entre a camada de texto e a camada de informação de natureza editorial. As etiquetas podem revestir diferentes formas em função do programa que se estiver a usar, podendo ser simplesmente *negrito*, _sublinhado_, /cursivo/ ou <etiqueta>. Uma linguagem de marcação fornece uma descrição, e depois é sempre necessário dispor de programas específicos e de ferramentas capazes de a interpretar e de lhe dar um destino útil.

Uma das linguagens de marcação mais conhecidas é o HTML (*HyperText Markup Language*), usada na descrição das páginas da internet, padrão que é mantido pela organização *World Wide Web Consortium* (W3C) (http://www.w3.org/). Nas etiquetas de marcação dessas páginas, incluem-se indicações para o navegador disponibilizar a visualização do conteúdo por meio de chamadas de atenção para o início e o fim da informação adicional. Veja-se um exemplo em (1) e (2), onde se demonstra o uso de e para assinalar as partes de um texto que devem ser visualizadas a negrito quando se consulta a página da internet.

> (1) Dentro desta frase em particular, esta secção deve ser a negrito para efeitos de disponibilização na internet.
>
> (2) Dentro desta frase em particular, **esta secção deve ser a negrito** para efeitos de disponibilização na internet.

Na visualização oferecida em (2), a qualidade do negrito é percebida analogicamente, como sempre aconteceu na tradição impressa, i.e., os caracteres são eles próprios diferentes, neste caso mais carregados em termos de preto. Pelo contrário, em (1), a qualidade do negrito indica-se e visualiza-se digitalmente, constituindo uma informação com existência discreta em relação aos caracteres em que o texto vem escrito.

Uma outra linguagem de marcação, esta extremamente relevante para o universo das Humanidades Digitais, é o XML (*eXtensible Markup Language*), uma linguagem descritiva de aplicação generalizada, descendente de um padrão anterior, o do SGML (Goldfarb1999). Ao contrário do HTML, o XML não fornece um conjunto fixo de etiquetas, limitando-se a definir um formato. São depois os utilizadores quem escolhe que nomes dar e que significado associar a cada etiqueta de marcação. Por exemplo, uma indicação de negrito em XML tanto pode ser <b>, como <n>, como <negrito>, como <meunegrito>, desde que se estabeleça no lugar apropriado o que significa a etiqueta e como ela precisa de ser interpretada pelos programas que reconhecem tal formato. É por isso que cada documento XML tem de ser acompanhado por definições externas que especificam o significado e a estrutura das etiquetas ali particularmente usadas, definições essas que recebem o nome de *XML schema*, ou então o nome do modelo seu antecessor, DTD (*Document Type Definition*).

As etiquetas XML organizam-se dentro de uma estrutura em árvore. Por exemplo, querendo transpor o conteúdo de um livro para linguagem XML, as etiquetas de parágrafo ficam em ramos inferiores, filhos dos nós de secção, por sua vez filhos dos de capítulo; todos juntos, organizam-se num todo ramificado que representa o conteúdo do livro.

Dado que a linguagem XML é, como o nome indica, muito maleável ou "extensível", torna-se potencialmente infinita a variedade dos seus elementos, indicados com parênteses angulares <xxx>, e seus atributos, indicados entre aspas dentro da categoria dos elementos <xxx y="zzz">. Assim, não tardou a surgir, manifestada por académicos, bibliotecas e arquivos, a necessidade de se dispor de um formato estandardizado para o uso desta linguagem de marcação na descrição e edição eletrónicas de conteúdos textuais. Foi por isto que se impôs na comunidade das Humanidades Digitais o padrão TEI (*Text Encoding Initiative*), desenvolvido desde finais da década de 1980. O objetivo foi o de criar e manter um padrão independente de marcação para a codificação de dados em Humanidades Digitais. A versão atual do TEI segue definições XML e criou um Manual muito detalhado que permite a adoção de um mesmo padrão em resultados digitais tão diferentes como são a edição crítica de uma obra literária, a transcrição de diálogo num arquivo de registo oral, a codificação detalhada de metainformação sobre a relação entre objetos digitais e seus originais físicos, ou a codificação sobre a proveniência e a anotação de corpora linguísticos. Em relação a estes, foram publicadas em 1996 as recomendações para se passar a dispor de corpora linguísticos estandardizados em TEI (Ide/Priest-Dorman/Véronis 1996), as quais têm tardado, contudo, a ser aplicadas a corpora portugueses.

Cada documento XML-TEI divide-se em duas partes, a do conteúdo, etiquetada como <text>, e a do cabeçalho, etiquetada como <TeiHeader>. Esta última descreve a metainformação relativa ao documento de que se trate, a qual pode ser usada por motores de busca desenhados para pesquisar este tipo de XML. Tudo quanto se registe em termos de metadados – relativos à história, aos suportes e às edições do texto em causa e à edição eletrónica que se estiver a elaborar – fica disponível para pesquisas

avançadas, articuláveis com as que incidam sobre os conteúdos textuais em si, anotados com marcação que pode ser filológica, linguística, geográfica, cronológica ou outra.

Na Tabela 2 incluímos a demonstração de como algumas etiquetas TEI podem servir a codificação, dentro do elemento<text>, das abreviaturas e leituras difíceis de um manuscrito (cf. também as Tabelas 4 e 5):

1	<expan> <abbr>Ill</abbr> <ex>ustrissi</ex> <abbr>mo</abbr> </expan>
2	<choice> <expan>Illustrissimo</expan> <abbr>Illmo</abbr> </choice>
3	<supplied resp=”CA” reason=”damage”>pois</supplied>
4	Tinha <unclear>vinurazer</unclear> hum bocado de contrabando

Tabela 2 Codificação de abreviaturas, conjeturas e *loci desperati* segundo o protocolo TEI P5 (TEI Consortium 2015)

Linha 1: codificação da abreviatura “Illmo” e indicação do seu desenvolvimento em “Illustrissimo”.

Linha 2: codificação alternativa do desenvolvimento da mesma abreviatura.

Linha 3: codificação da conjetura “pois” feita pelo editor “CA”, devidamente identificado no cabeçalho do documento XML-TEI; esta é uma conjetura motivada por danificação do manuscrito

Linha 4: codificação de um *locus desperatus* em “vinurazer”, uma palavra cujos caracteres se leem bem no manuscrito, mas que desafiam o entendimento do editor.

A reprodução facsimilada do manuscrito de onde provêm as formas acima codificadas, uma carta particular escrita provavelmente em 1827 por um criado que assim pedia ajuda ao seu antigo patrão, pode observar-se no sítio em linha do projeto P.S. *Post Scriptum*.[5] Na mesma localização escolhe-se a visualização de uma edição diplomática, crítica ou modernizada, com ou sem indicação, conforme desejado, de lemas e de anotação morfossintática.

O significado variado das etiquetas, mesmo sem se sair do padrão TEI, ajuda não só a transformar o documento em causa na base para vários formatos simultâneos de saída,

[5] Localização: http://ps.clul.ul.pt/pt/index.php?action=edit&cid=CARDS0002.

que vão da edição mais ou menos diplomática à mais ou menos modernizada, mas podem ser também, entre muitos outros, um glossário, uma lista de abreviaturas desenvolvidas, uma amostra de treino para operações automáticas ou um corpus linguisticamente anotado.

Ao mesmo tempo, torna-se possível com este tipo de recurso correr testes automáticos à qualidade do trabalho já executado e desencadear campanhas de correção, quer manual quer automática, dos procedimentos adotados ao longo da edição digital. Para tanto, são de grande ajuda os programas escritos em Perl, acima referidos. Ilustra-se um caso na Tabela 3, que contém um programa em Perl destinado a extrair e listar em separado o conteúdo textual dos parágrafos do corpo de um texto codificado em XML-TEI.

1	`#!/usr/bin/perl`	1	Declaração de que se trata de um programa Perl.
2	`use strict;`	2	Instrução dada à interpretação do Perl para que seja estrita...
3	`use warnings;`	3	...e para que emita avisos sempre que surjam erros no código.
4	`use XML: :LibXML;`	4	Instrução de utilização de uma biblioteca Perl, uma biblioteca preexistente, formada especificamente para lidar com o formato XML.
5	`my $filename = "CARDS0001.xml";`	5	Especificação do ficheiro de entrada, aquele onde se deseja que os comandos venham a ser executados, no caso, o ficheiro CARDS0001.xml.
6	`my $xmlparser = XML: :LibXML->new();`	6	Instrução de criação de um analisador de sintaxe (um parser) capaz de lidar com XML.
7	`my $doctree = $xmlparser->parse_file($filename);`	7	Instrução para o parser analisar a estrutura do ficheiro de entrada: o resultado é a distribuição por uma estrutura em árvore de toda a informação do ficheiro de entrada.

8	`foreach my $par ($doctree->findnodes)('/TEI.2/text/body/p')){`
	`print $par->to_literal, "\n";`
	`}`
		8	Instrução para que no documento com a estrutura em árvore ($doctree) sejam encontrados os nós com a etiqueta 'p', hierarquicamente inferiores aos nós etiquetados com 'body', 'text' e 'TEI.2', sucessivamente, sendo que para 'p' encontrado, o respetivo conteúdo textual deve ser impresso e fechado, no final, com um código significando 'nova linha'(\n)

Tabela 3 Programa Perl exemplificado à esquerda e respetiva explicação à direita

No exemplo, usa-se a função 'findnodes' para chegar à informação contida dentro de certos elementos, os elementos 'p' (no caso, com o valor de 'parágrafo'). Para tanto, indica-se concretamente o caminho a seguir até se encontrar a localização de 'p', que,

dentro da estrutura do documento, é filho de 'body', por sua vez filho de 'text', e este filho de ´TEI.2'. Para cada 'p' encontrado, extrai-se para impressão em ficheiro separado o seu conteúdo textual, fechando-se o resultado com uma instrução para mudança de linha (barra n).

Neste exemplo, o conjunto de instruções em Perl é relativamente curto e simples, mas a mesma linguagem pode servir para inúmeras manipulações, impostas pelas necessidades de investigação ou tão-só pela imaginação. Pode-se conciliar todo o género de variáveis extratextuais previamente registadas ao nível de metadados (ex: género, cronologia, classificações sociológicas, coordenadas geográficas) com a extração de partes ou da totalidade do texto. Podem-se procurar padrões indetetáveis a olho nu, determinados pelo contexto em que as palavras ou os caracteres ocorrem. Podem-se esconder para efeitos de busca todas as formas conjeturadas de uma edição crítica, ou todas as leituras provenientes de um testemunho menos fiável.

9.5 A edição académica digital

Esta última secção destina-se a referir recursos para o estudo histórico do português que estejam a ser construídos no âmbito das Humanidades Digitais. O seu formato emblemático é o da edição académica digital (*scholarly digital edition*), um "recurso de informação que oferece uma representação crítica de documentos ou textos (normalmente) históricos", na definição de Patrick Sahle (Sahle 2014). Seguindo o mesmo autor, responsável por uma pormenorizada elaboração de critérios para avaliação e descrição de edições académicas digitais e pelo respetivo catálogo (http://www.digitale-edition.de/), não cabem nesta classificação as meras publicações em formato digital por não serem "sistemas de informação que seguem uma metodologia determinada pelo paradigma digital", como é o caso da metodologia que envolve cuidado na codificação de caracteres e nas linguagens de programação e de marcação que acabámos de ver. As simples publicações em formato digital seguem, por seu lado, metodologias do paradigma impresso. Não se podem considerar edições académicas digitais, por conseguinte, nem as edições impressas digitalizadas, ainda que críticas, nem a maioria dos projetos de digitalização empreendidos por bibliotecas e arquivos, mesmo que acompanhados de descrição, transcrição e indexação. Igualmente excluídos estão os casos das edições digitais que não envolvam representação crítica, i.e., que não respeitam critérios filológicos mínimos, despreocupadas que estão com a responsabilidade de oferecer leituras não mistificadas dos textos históricos publicados.

Em sentido estrito, os corpora anotados hoje disponíveis para o estudo histórico do português (CIPM[6], *Corpus do Português*,[7] *Tycho Brahe*,[8] P.S. *Post Scriptum*[9], WOChWEL[10] e *Colonia*[11]) também não foram montados para funcionarem como

[6] Localização: http://cipm.fcsh.unl.pt.
[7] Localização: http://www.corpusdoportugues.org/.
[8] Localização: http://www.tycho.iel.unicamp.br/~tycho/corpus/index.html.
[9] Localização: http:/clul.ul.pt.
[10] Localização: http://alfclul.clul.ul.pt/wochwel/index.html.
[11] Localização: http://corporavm.uni-koeln.de/colonia/.

edições académicas digitais, desprovidos que estavam todos eles, no início, de linguagem de marcação textual de natureza filológica e da respetiva ancoragem à anotação gramatical, fosse ela POS, morfossintática ou sintática. Cumpriam a prática consagrada da linguística de corpus de separar enquanto objetos fisicamente autónomos os documentos que contêm o texto-fonte dos documentos cujos tokens são acompanhados de anotação linguística (*stand-off annotation*, cf. McEnery/Wilson 2001, 38).

A anotação em *stand-off* tem inegáveis vantagens, reconhecidas mesmo no âmbito das edições académicas digitais (cf., por exemplo, Schmidt 2010). Permite i) o cruzamento de ramos ao longo de níveis diferentes de anotação, o que não podeacontecer em XML por desencadear malformações estruturais, ii) a ausência de interferência entre os vários níveis de anotação adotados, iii) diferentes versões de uma mesma anotação, iv) o acrescento de níveis ulteriores de marcação, sem alteração do texto-fonte, v) o trabalho simultâneo de várias pessoas sobre os mesmos dados, primários e inalterados e vi) a salvaguarda em relação a problemas gerados pela legislação que protege o direito de autor, uma vez que os documentos primários estão arquivados separadamente.

No entanto, algumas destas vantagens convertem-se em desvantagens quando se passa da teoria à prática. Ao anotar informação a níveis diferentes, a ocorrência de erros é muito comum, como observam Grover et al., no caso, para a anotação de entidades biomédicas: surgem constantemente incompatibilidades entre as dependências das fronteiras de palavra e as dependências das etiquetas para as diferentes entidades identificadas (Grover/Matthews/Tobin 2006). Nas edições académicas digitais, por seu lado, acontece constantemente surgirem novas interpretações de natureza filológica a propósito das fontes primárias, o que conduz a necessidades de modificação da anotação de uma mesma ou de várias palavras, já entretanto registadas em níveis diferentes. Pode torna-se por conseguinte muito pouco prático o procedimento do *stand-off*, já que os níveis de anotação se distinguem aí, precisamente, pela independência recíproca e pela não-interferência. A mesma mudança tem de ser repetida, nesta modalidade, tantas vezes quantos os níveis que houver. Há a possibilidade técnica, que está inclusivamente a ser experimentada (cf. por exemplo Druskat 2014), de criar dependências entre os vários níveis de uma anotação em *stand-off*. No entanto, quanto mais complexas e adaptadas, no sentido de não-estandardizadas, forem as camadas de uma anotação, mais difícil e demorada se torna a escrita de programas capazes de lidar simultaneamente com todas elas. Noutro sentido, e uma vez que é muito comum haver investigadores diferentes a trabalhar em simultâneo sobre a mesma coleção de textos, torna-se essencial que as anotações e revisões de cada um sejam sistematicamente visíveis para os outros, o que não acontece quando se trabalha em planos independentes. Também por causa disto, a anotação alinhada (*embedded* ou *inline*), alternativa à anotação em *stand-off* por fazer coincidir num mesmo documento todos os níveis de anotação, pode tornar-se mais desejável (cf.).

No sentido de mudar do sistema em *stand-off* para um sistema alinhado, o projeto *Tycho Brahe* da Universidade de Campinas, coordenado por Charlotte Galves, construiu a

ferramenta *eDictor* (http://edictor.net/), a combinação de um editor de XML e de um etiquetador morfossintático que gera automaticamente edições navegáveis, i.e., em HTML, que podem ser ora diplomáticas, ora semidiplomáticas, ora modernizadas, bem como versões com anotação morfossintática, tanto em texto simples como em XML, sendo que se anuncia também um futuro módulo para anotação sintática. Na retaguarda de todos estes diferentes formatos de saída estão documentos em XML que integram, alinhadas em nós irmãos, tanto a informação textual, como a filológica e a morfossintática (Cf. Tabela 4). Qualquer mudança ou revisão dos processos de registo dessas informações só tem de ser introduzida uma vez, o que permite não só o trabalho em equipa em torno de uma mesma edição como a prevenção de desalinhamentos e incompatibilidades entre os formatos de saída, que podem ser gerados de novo na sequência dos processos de correção.

1	<w id="20"> <o>Illmo</o> <e t="exp">Illustrissimo</e> <e t="norm">Ilustríssimo</e> <m v="ADJ-S"/> </w>
2	<w id="48"> <o>pois</o> <comment author="CA" date="09/24/15" title="supplied">damage</comment> <m v="C"/> </w>
3	<w id="93"> <o>vinurazer</o> <comment author="CA" date="09/24/15" title="unclear">unclear</comment> <m v="VB"/> </w>

Tabela 4 Estrutura de um XML editado em eDictor, correspondente à anotação alinhada do mesmo exemplo da Tabela 2

A ferramenta *eDictor* foi anunciada em 2010 (Faria/Kepler/Sousa 2010, Sousa 2013b) e tem estado a ser usada no sentido de criar corpora históricos anotados a partir de edições académicas digitais e vice-versa (o *Corpus Anotado do Português Tycho Brahe* da Universidade de Campinas, os projetos do Grupo de Pesquisas Humanidades Digitais da Universidade de São Paulo, os do Laboratório de História do Português Brasileiro da Universidade Federal do Rio de Janeiro, o *Corpus Eletrônico de Documentos Históricos do Sertão*, CE e o corpus WOChWEL do Centro de Linguística da Universidade de Lisboa (CLUL), entre os que têm já materiais disponíveis).

No projeto P.S. *Post Scriptum* do CLUL está a ser usada desde 2014 uma tática alternativa à do recurso à ferramenta *eDictor*. Dado que o *eDictor* não aceita ainda ficheiros originalmente elaborados em XML-TEI (só aceita as suas próprias definições de XML), dado que também ainda não é utilizável em linha, que não tem lematizador e que não oferece maleabilidade para a inclusão de anotadores automáticos diferentes dos do *TychoBrahe*,[12] o P.S. *Post Scriptum* passou a utilizar o sistema TEITOK,[13] desenvolvido no CLUL por Maarten Janssen, no âmbito daquele e de outros projetos da mesma instituição. Tal como o *eDictor*, o TEITOK cria um mesmo suporte em XML para o corpus linguisticamente anotado e para a edição académica digital, mas em TEITOK esse suporte edita-se num ambiente em linha. Trata-se de um sistema baseado na web para visualizar, criar e editar textos com marcação filológica rica acompanhada de anotação linguística. O sistema contém uma interface gráfica em que o documento anotado pode ser visualizado em formatos diferentes, dependendo dos interesses do visitante. Já para os administradores do mesmo sistema, o TEITOK permite que na mesma interface se edite, transforme e anote o XML-TEI subjacente. Pode-se assim modernizar automaticamente a ortografia dos textos transcritos sem perder a marcação de origem, pode-se lematizar e anotar morfossintaticamente os mesmos textos com recurso a anotadores automáticos definidos pelo utilizador e consultar todo o resultado em função das variáveis extratextuais registadas ao nível de metadados. Mais uma vez, na retaguarda de toda esta maleabilidade está uma anotação alinhada e não uma anotação em *stand-off*. Estão ficheiros em linguagem XML que alinham em torno da leitura crua das palavras da fonte primária, sucessivos atributos com as informações paleográficas, filológicas, lexicais e gramaticais que o investigador queira registar (Cf. Tabela 5).

1	`<tok id=”w-20” form=”Illmo” pos=”ADJ-S” fform=”Illustrissimo” nform=”Ilustríssimo” lemma=”ilustre”>Illmo</tok>`
2	`<supplied resp=”CA” reason=”damage”> <tok id=”w-48” form=”pois” pos=”C” lemma=”pois”>pois</tok> </supplied>`
3	`<unclear> <tok id=”w-93” pos=”VB” lemma=”vinurazer”>vinurazer</tok> </unclear>`

Tabela 5 Estrutura de um XML editado em TEITOK, correspondente à anotação alinhada do mesmo exemplo das Tabelas 2 e 4

O sistema tem também a vantagem de poder ser usado na edição e anotação de textos de qualquer língua (já está a anotar espanhol, por exemplo, ainda dentro do P.S. Post Scriptum), mas apresenta a desvantagem de exigir a presença de programadores experientes nas equipas que o usam, técnicos que consigam modificá-lo no sentido da inclusão de corpora de treino e de etiquetadores automáticos adicionais.

[12] Localização: http://www.tycho.iel.unicamp.br/~tycho/corpus/manual/tags.html.
[13] Localização: http://alfclul.clul.ul.pt/teitok/site/index.php.

Tanto o *eDictor* como o TEITOK continuam a ser constantemente apurados e expandidos em termos de sofisticação. No horizonte mais próximo, por exemplo, está o desafio do alinhamento da anotação sintática. Mas ambos demonstram para já, aliados a exemplos referidos nas secções anteriores, como a filologia do português sempre se manteve ativa ao nível da experimentação em Humanidades Digitais.

9.6 Observações finais

Vimos neste capítulo como as Humanidades Digitais se caracterizam por uma principal virtude: são abertas. Abertas no sentido de serem acessíveis, discutíveis, colaborativas, democráticas. Caracterizam-se também por uma imposição: são velozes. Os procedimentos mecânicos têm uma velocidade intrínseca que é à partida incompatível com o ritmo pausado da interpretação, ou exegese, método incontornável na construção de conhecimento em Humanidades. Entrou também, por conseguinte, na agenda das Humanidades Digitais a reflexão sobre novos métodos, adaptados à tecnologia, ou transformadores da mesma. Também se revisitam velhos temas, convertidos em tópicos atuais pela sociedade de informação: oposição entre técnica e conhecimento, diferença entre língua natural e língua artificial, relação entre texto e gramática.

Referências

Barthes, Roland (1994, [1]1971), *De l'ouvre au texte*, in Éric Marty (org.), *Œuvres complètes*, vol. 2, 1211–1217.

Beaugrande, Robert-Alain/Dressler, Wolfgang Ulrich (2005, [1]1972), *Introducción a la lingüística del texto*, Barcelona, Ed. Ariel.

Bédier, Joseph (1970), *La tradition manuscrite du Lai de l'ombre: réflexions sur l'art d'éditer les anciens textes*, Paris, Librairie Honoré Champion.

Blecua, Alberto (1983), *Manual de crítica textual*, Madrid, Editorial Castalia.

Busa, Roberto A. (2004), *Foreword: Perspectives on the Digital Humanities*, in: *A Companion to Digital Humanities*, Susan Schreibman/Raymond George Siemens/ John Unsworth (edd.), *A Companion to Digital Humanities*, Malden MA, Blackwell Publishing Ltd, xvi–xxi.

Castro, Ivo (2001), *Metodologia do Aparato Genético*, in: Manuel Simões/Ivo Castro/João David Pinto Correia (edd.), *Memória dos Afectos: Homenagem a Giuseppe Tavani*, Lisboa, Colibri, 69–81.

Castro, Ivo/Castro, Maria Helena Lopes de/Cepeda, Isabel Vilares/Madureira, Virgílio (1973), *Normas de Transcrição para Textos Medievais Portugueses*, Boletim de Filologia 23, 417–425.

Cerquiglini, Bernard (1989), *Éloge de la variante. Histoire critique de la philologie*, Paris, Seuil, 1989.

Del Mancino, William/Pierrel, Jean-Marie (2009*), Du trésor de la langue française à l'ATILF et au CNRTL*, La revue pour l'histoire du CNRS [en ligne] 24, s.f. http://histoire-cnrs.revues.org/9133.

Druskat, Stephan (2014), *An Open-Source Software Platform for Multi-Level Corpus Annotation*, in: Druskat, Stephan/Bierkandt, Lennart/Gast, Volker/Rzymski, Christoph/Zipser, Florian (edd.), *Proceedings of the 12th Konferenz zur Verarbeitung natürlicher Sprache (KONVENS 2014)*, Hildesheim, 228–234.

Emiliano, António (2005), *Tipo Medieval para Computador: uma Ferramenta Informática para Filólogos, Historiadores e Paleógrafos*, Signo. Revista de Historia de la Cultura Escrita 15, 139–176.

Ertuna, Irmak (2009), *Stiegler and Marx for a Question Concerning Technology*, Transformations 17. url: http://www.transformationsjournal.org/journal/issue_17/article_07.shtml

Faria, Pablo P. F./Kepler, Fábio N./Sousa, Maria Clara Paixão de (2010), *An Integrated Tool for Annotating Historical Corpora*, in: *Proceedings of the Fourth Linguistic Annotation Workshop*, Stroudsburg PA, The Association for Computational Linguistics, 217–221.

Fiormonte, Domenico (2012), *Towards a Cultural Critique of the Digital Humanities*, in: Manfred Thaller (ed.), *Controversies Around the Digital Humanities*, Historical Social Research/Historische Sozialforschung, special issue, 59–76.

Goldfarb, Charles F. (1999), *Future Directions in SGML/XML*, in: Wiebke Möhr/Ingrid Schmidt (edd.), *SGML und XML*, Berlin/Heidelberg, Springer, 3–25. http://dx.doi.org/10.1007/978-3-642-46881-0_1.

Grover, Claire/ Matthews, Michael/Tobin, Richard (2006), *Tools to Address the Interdependence between Tokenisation and Standoff Annotation*, in: *Proceedings of the 5th Workshop on NLP and XML: Multi-Dimensional Markup in Natural Language Processing*, Stroudsburg PA, Association for Computational Linguistics, 19–26. url: http://dl.acm.org/citation.cfm?id=1621034.1621038

Hayles, N. Katherine (2004), *Print is Flat, Code is Deep: The Importance of Media-specific Analysis*, Poetics Today 25, n. 1, 67–90.

Hockey, Susan/Marriott, Ian (1980), *The Oxford Concordance Program. Version 1.0. Users' Manual*, Oxford, Oxford University Computing Centre.

Ide, Nancy/ Priest-Dorman, Greg/Véronis, Jean (1996); *Corpus Encoding Standard*, http://www.cs.vassar.edu/CES/.

Kirschenbaum, Mathew G. (2010), *What is Digital Humanities and what's it doing in English Departments?*, Association of Departments of English Bulletin 150, 1–7.

Koch, Peter/Oesterreicher, Wulf (2007, [1]1999), *Lengua hablada en la Romania: español, francés, italiano*, Madrid, Editorial Gredos.

Mateus, Maria Helena Mira (1968), *Informática e Linguística: a Mecanografia nos Estudos da Linguagem*, Revista de Portugal, série A 33, 217–232.

Mateus, Maria Helena Mira (1974), *Glossário da Vida e Feitos de Júlio César, tradução portuguesa quatrocentista de Li Fet des Romains*, Boletim de Filologia 23, 1–80.

Mateus, Maria Helena Mira (2010), *Vida e Feitos de Júlio César*, vol. 3, Lisboa, Fundação Calouste Gulbenkian.

McCarty, Willard (2004), *Modeling: A Study in Words and Meanings*, in: Susan Schreibman/Raymond George Siemens/John Unsworth (edd.), *A Companion to Digital Humanities*, Malden MA, Blackwell Publishing Ltd., 254–270. http://www.digitalhumanities.org/companion/.

McGann, Jerome J. (2014), *A New Republic of Letters: Memory and Scholarship in the Age of Digital Reproduction*, Cambridge MA/London, Harvard University Press. http://search.ebscohost.com/login.aspx?direct=true&scope=site&db=nlebk&db=nlabk&AN=706819.

McEnery, Tony/ Wilson, Andrew (2001, [1]1996), *Corpus Linguistics*, Edinburgh, Edinburgh University Press.

Murdoch, Brian O. (1971), *The Production of Concordances from Diplomatic Transcriptions of Early Medieval German Manuscripts: Some Comments*, in: Roy A. Wisbey (ed.), *The Computer in Literary and Linguistic Research*, Cambridge, Cambridge University Press, 35–44.

Parkinson, Stephen (1983), *Um Arquivo Computorizado de Textos Medievais Portugueses*, Boletim de Filologia 28, 241–252.

Poisson, Esther (2011), *Le Trésor de la langue française informatisé: une ressource d'une valeur insoupçonnée*, Correspondance 17, n. 1. url: http://correspo.ccdmd.qc.ca/Corr17-1/Capsule.html.

Prista, Luís/Albino, Cristina (1996), *Filólogos Portugueses entre 1868 e 1943: Catálogo da Exposição organizada para o XI Encontro Nacional da Associação Portuguesa de Linguística, Faculdade de Letras de Lisboa, 1995*, Lisboa, Colibri.

Ricœur, Paul (1986), *Du texte à l'action. II, Essais d'herméneutique*. Paris, Éditions du Seuil.

Robertson, Benjamin J. (2015), *The Grammatization of Scholarship*, Amodern 1. url:http://amodern.net/article/the-grammatization-of-scholarship/.

Roncaglia, Aurelio (1975), *Principi e applicazione di critica testuale*, Roma, Bulzoni Editore.

Sahle, Patrick (2014), *Criteria for Reviewing Scholarly Digital Editions, Version 1.1.*, IDE, Institur für Dokumentologie und Editorik, http://www.i-d-e.de/reviews/criteria-version-1-1.

Schmidt, Desmond (2010), *The Inadequacy of Embedded Markup for Cultural Heritage Texts*, Literary and Linguist Computing 25, 3, 337-356.

Sousa, Maria Clara Paixão de (2013a), *Texto Digital: uma Perspectiva Material*, Revista da ANPOLL 35, 17–60.

Sousa, Maria Clara Paixão de (2013b), *A Filologia Digital em Língua Portuguesa: alguns Caminhos*, in: Maria Filomena Gonçalves/Ana Paula Banza (edd.), *Património Textual e Humanidades Digitais: da Antiga à Nova Filologia*, Évora, CIDEHUS, 113–138.

Steriade, Donca (1995), *Underspecification and Markedness*, in: John A. Goldsmith (ed.), *The Handbook of Phonological Theory*, Oxford, Blackwell, 114–74.

Stiegler, Bernard (1998, [1]1994), *Technics and Time,1, The Fault of Epitheus*, Stanford CA, Stanford University Press.

TEI Consortium (2015), *Guidelines for Electronic Text Encoding and Interchange*. [last modified 2015]. http://www.tei-c.org/P5/.

Timpanaro, Sebastiano (2005, [1]1963), *The Genesis of Lachmann´s Method*, Chicago, Chicago University Press.

Wall, Larry/Christiansen, Tom/Orwant, Jon (2004), *Programming Perl*, s.l., O'Reilly Media Inc., 2004.

Xavier, Maria Francisca/Brocardo, Maria Teresa/Vicente, Maria da Graça (1995), CIPM: Um Corpus Informatizado do Português Medieval, in: *Actas do X Encontro Nacional da Associação Portuguesa de Linguística*, 599–612, [Lisboa], APL.