CONCEITO MATERIAL DE "TEXTO DIGITAL": UM ENSAIO

TOWARDS A MATERIAL CONCEPT OF "DIGITAL TEXTS"

Maria Clara Paixão de Sousa
Doutora em Lingüística pela Unicamp
Universidade de São Paulo
mariaclara@usp.br

RESUMO

A circulação da escrita no ambiente digital renova a relevância de algumas das tarefas tradicionais da crítica textual – em particular, a tarefa de refletir conceitualmente sobre as cadeias de difusão dos textos. Este artigo explora os desafios assim colocados para o campo, partindo para isso de uma reflexão sobre a natureza material do texto digital. Irei sugerir que a característica diferenciadora do texto digital, nesta perspectiva puramente material, é a inclusão de uma etapa de processamento artificial da linguagem na sua cadeia de difusão. Esta etapa singulariza o texto digital e determina suas condições de produção e transmissão, tornando-o um objeto teórico singular para a crítica textual.

1. Singularidade do "texto digital"

Neste artigo irei sugerir que o "texto digital" diferencia-se das demais formas de texto pela inclusão de etapas de processamento artificial da linguagem em sua cadeia de difusão. É dessa singularidade que extrairemos um conceito de "texto digital", e é nela que poderemos observar as implicações teóricas e metodológicas do trabalho com este tipo de texto. Essa abordagem parte de uma breve reflexão em torno da seguinte pergunta: qual a característica fundamental a distinguir o "texto digital" frente às demais formas de "texto" quanto ao seu processo de produção, circulação e recepção?

Para começar a explorar esta pergunta, vamos pensar o "texto" como uma ponte no espaço-tempo: um registro de enunciados produzidos num ponto do espaço e do tempo, que podem ser recebidos num ponto diferente do espaço e do tempo. Para construir essas pontes que chamamos textos, as culturas humanas inventaram técnicas de representação da linguagem, e técnicas de registro e transmissão dessa representação. Nessa perspectiva estritamente material, se quisermos compreender o lugar do "texto digital" entre os demais tipos de textos precisamos compreender os processos envolvidos nas duas pontas: de um lado, os sistemas simbólicos de representação da linguagem; de outro lado, as tecnologias envolvidas na difusão dos sistemas simbólicos.

Quanto à primeira ponta, a representação da linguagem tem funcionado por meio de sistemas de correspondência simbólica entre informação lingüística e sinais gráficos - isto é: da escrita. Os principais sistemas de escrita conhecidos diferenciam-se sobretudo quanto ao nível de informação linguística a ser representado – idéias (na escrita "analítica", por exemplo a escrita ideogramática chinesa) ou sons (na escrita "fonética", por exemplo a alfabética ocidental). A "História da Escrita", neste primeiro plano, é a história das diferentes maneiras encontradas pelas diferentes culturas humanas para operar diferentes sistemas simbólicos capazes de representar a linguagem.

De outro lado está a questão das tecnologias envolvidas na difusão da informação simbólica representada pela escrita. Aqui saímos do plano dos sistemas simbólicos, e entramos no plano estritamente material das tecnologias inventadas pelo homem para estabelecer as correspondências simbólicas dentro de cada sistema e propagá-las no tempo e no espaço. Nesse aspecto, por milênios a humanidade valeu-se de uma mesma tecnologia fundamental, que vamos chamar de "lógico-sensorial".

Essa tecnologia depende de dois fatores apenas: o primeiro é a mente humana e sua capacidade lógica; o segundo é a disponibilidade dos sinais de um sistema simbólico frente ao sistema perceptual humano. Tipicamente, a apreensão dos sinais dos sistemas de escrita se dará por percepção visual, e a informação simbólica visualmente percebida será decodificada graças às capacidades cognitivas humanas e ao conhecimento cultural partilhado de determinado sistema de representação. Portanto: nesta tecnologia, para que o processo de codificação e decodificação da informação aconteça, basta que os sinais a serem codificados e decodificados estejam passíveis de apreensão pelo sistema sensorial humano, tipicamente a visão. Diante disso, as diferentes "técnicas de escrita" historicamente desenvolvidas pelo homem dirigiram-se a tornar a informação simbólica aparente e transportável.

Assim é que para registrar e transportar a escrita de um ser humano até o outro, diferentes culturas inventaram diferentes artefatos "carregadores de sinais gráficos", cuja construção envolveu diferentes modos de chegar a um mesmo objetivo: inscrever sinais gráficos aparentes em um suporte capaz de levar adiante esses sinais. Isso pôde ser feito, por exemplo, graças a um instrumento (graveto) que inscrevesse os sinais em um suporte maleável (argila); um instrumento duro (cinzel) que rasgasse os sinais sobre um suporte duro (pedra); um instrumento (pena; lápis; giz; tipo de chumbo...) que transferisse pigmentos (tinta; pó de grafite...) por sobre suportes absorventes (cascas de árvore; peles de animais; papel...). Podemos reconhecer diversas etapas de desenvolvimento e aperfeiçoamento de técnicas inventadas com o propósito básico de "carregar

informação codificada". A "história da escrita", neste sentido material estrito, é a história das transformações técnicas desse processo de transportar enunciados pelo espaço e pelo tempo. Algumas dessas transformações foram revolucionárias, em particular a introdução de instrumentos mecânicos que puderam substituir a mão do homem na tarefa de estampar os sinais gráficos nos suportes, que veio a ampliar de modo inédito a capacidade de reprodução de um "mesmo texto" para um grande número de leitores, revolucionando as culturas de escrita (EISENSTEIN, 1998).

Entretanto, mesmo a introdução da escrita mecânica configura-se fundamentalmente como uma etapa técnica da evolução de uma mesma tecnologia: a fabricação de artefatos que carreguem informações lingüísticas codificadas até seus de-codificadores humanos. Em todas essas etapas, o sistema de estabelecimento das correspondências simbólicas, em si, funcionou da mesma maneira fundamental: a correspondência lógico-sensorial, na qual os seres humanos apreendem uma informação aparente (sinais gráficos) graças a seu sistema perceptual (a visão, o talvez também o tato) e decodificam-na graças a suas capacidades cognitivas. Essencialmente, na escrita manuscrita ou na mecânica, todo o processo de codificação e decodificação da informação se dá no plano do ser humano – os artefatos da escrita tradicional não participam deste processo, apenas levam as informações codificadas até seu decodificador.

No caso do "texto digital", estaremos diante de algo inteiramente diverso. Neste caso, não apenas a forma de levar a informação codificada é singular, mas – fundamentalmente – o processamento da informação a ser codificada e decodificada é outro, uma vez que envolve, além da correspondência "lógico-sensorial" humana, etapas de correspondência lógica artificial.

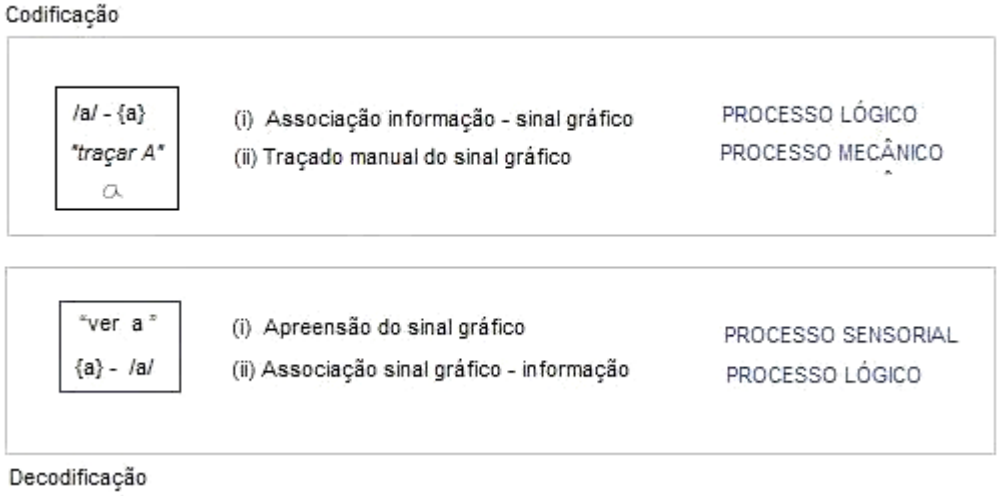
Vamos ver de perto como isso funciona, examinando um exemplo simplificado: em essência, o que fazemos quando "escrevemos à mão"? Traçamos sinais gráficos num suporte (uma folha de papel, por exemplo), com um instrumento (uma caneta, por exemplo) e matéria aparente (tinta, por exemplo). "Traçamos sinais gráficos" – ou seja, desenhamos formas que (sabemos) *significam alguma coisa*, ou seja, carregam informação

semântica mínima – desenhamos símbolos. Por exemplo, no caso de uma escrita alfabética (i.e., um sistema em que a correspondência lógica se dá entre os sinais gráficos e os fonemas de uma língua), de tipo "romano", posso desenhar a seguinte figura, se quero registrar o símbolo para o som /a/:

Este desenho ficará registrado no suporte ("folha de papel"), e poderá assim ser levado até a outra ponta do processo, onde estará o receptor do meu texto - o meu leitor, que "decodificará" a informação contida no desenho acima imediatamente, pelo mero contato visual. Isso é: supondo ser este meu leitor alguém que domina a escrita alfabética e conhece o alfabeto romano, ele reconhecerá visualmente este desenho (como um dos símbolos do alfabeto romano, "a"), e saberá que ele representa o som /a/.

Termina aí a cadeia de codificação e decodificação, que pode ser representada da seguinte forma esquemática:

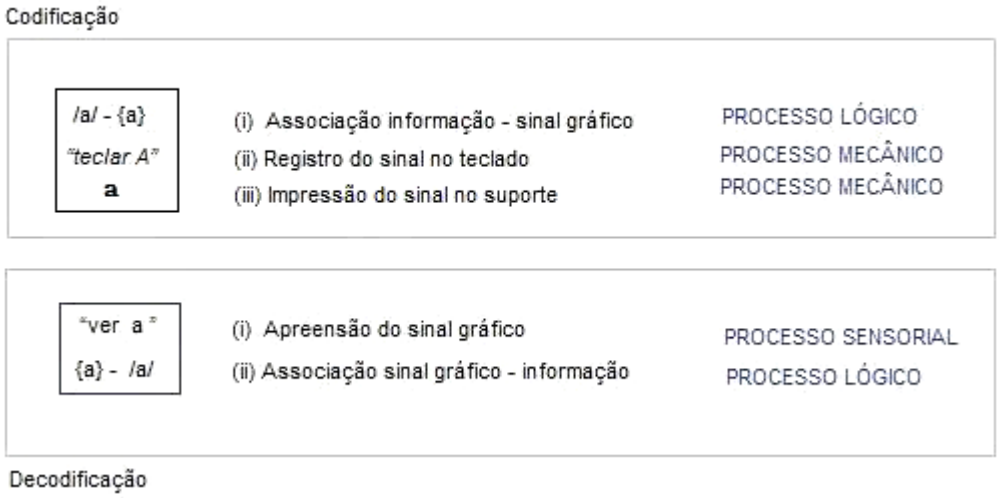
(1) Resumo dos processos na escrita manual



É verdade que o processo de registro do sinal gráfico poderia ser diferente – por exemplo, poderia contar com o auxílio de um instrumento mecânico, como a haste de uma máquina de escrever. Entretanto, os dois casos (manual e mecânico) evolvem apenas duas etapas de processos lógicos: a

escolha de um sinal gráfico a ser registrado, e a interpretação desse sinal gráfico – ambas as etapas realizadas por humanos:

(2)Resumo dos processos na escrita mecânica



Vejamos agora como o processamento do texto digital difere deste quadro, pensando em como a operação "escrever 'a'" se daria com o auxílio de num computador.

À primeira vista, a operação parece ser semelhante à que descrevemos acima para a máquina de escrever – afinal, ao usar o computador, temos diante de nós um conjunto de teclas, cada uma com o desenho de um símbolo; a cada vez que apertamos uma tecla, a imagem daquele símbolo aparece na tela - com as máquinas, também apertávamos uma tecla, e o símbolo correspondente aparecia no papel que estava envolto num cilindro à nossa frente.

Essa semelhança é, evidentemente, ilusória. Na máquina de escrever, entre a nossa ação de pressionar a tecla e o aparecimento do símbolo no papel, operava um processo mecânico (ou seja: a tecla fazia mover uma haste em cuja ponta havia um desenho em relevo do símbolo; a ponta da haste pressionava uma fita com tinta, e o desenho do símbolo se imprimia no papel – como se a haste fosse um prolongamento da nossa mão, simplesmente). No computador, entre a nossa ação de pressionar a tecla e o aparecimento do símbolo na tela opera um processo matemático, ao longo do qual a informação é recodificada de modo a aparecer como informação gráfica na tela. Ao pressionar a tecla do

computador, não acionamos um mecanismo físico – e sim ativamos um comando de programação que irá participar do processo de codificação e decodificação das informações. Neste caso, teríamos na cadeia algumas etapas lógicas adicionais:

(3) Resumo dos processos na escrita digital

Assim, na construção do "texto digital", entre a codificação e decodificação humanas, interfere o processamento artificial da informação. O que torna um texto "digital", de fato, não é simplesmente a técnica de registro do sistema simbólico, mas fundamentalmente a tecnologia envolvida na construção das correspondências entre símbolos e informação lingüística: há uma diferença lógica, para além da material, entre o texto no meio digital e os outros textos. O processamento digital inclui uma etapa adicional de codificação de informação - e essa etapa, notemos, é externa à mente do produtor e do receptor do "texto", algo inédito frente à tecnologia anterior.

Nesse sentido, podemos começar a explorar as possibilidades de compreender o que singulariza o texto digital entre as diferentes formas de "texto" inventadas pelas culturas de escrita ao longo da história. Não estamos diante de um incremento na evolução técnica na longa linhagem que vai das tabuletas de argila até a prensa mecânica, de mais uma técnica de registro e transporte do sistema simbólico.

Estamos, isso sim, diante de uma nova tecnologia de processamento do sistema simbólico, uma nova tecnologia de escrita, que envolverá novas maneiras de se combinar codificação e transporte da codificação.

Já vimos mais acima que as "tecnologias de escrita" envolvem dois momentos fundamentais: de um lado, a forma de estabelecimento das correspondências simbólicas; de outro lado, a forma de fixação e transporte das correspondências estabelecidas. Evidentemente, as técnicas para registrar e transportar as correspondências simbólicas são dependentes da forma pela qual essas correspondências se estabelecem. Na tecnologia "lógico-sensorial", o estabelecimento das correspondências exclusivamente com base nas capacidades sensoriais e lógicas humanas determinou técnicas de difusão fundamentalmente dedicadas a tornar os símbolos da escrita *aparentes*. As evoluções dessas técnicas representaram aperfeiçoamentos na forma de registro e transporte da informação visual codificada pelas mentes humanas, e não modificaram a tecnologia de codificação das informações, que permaneceu dependente apenas de um codificador humano com capacidade de percepção sensorial e capacidade de produzir e compreender os símbolos de um sistema de escrita.

Já no texto digital, esse processo é mediado por etapas de processamento artificiais que participam da codificação e que, evidentemente, não poderão fundar-se no princípio da aparência, da disponibilidade para o sistema sensorial. Para serem processadas artificialmente, as informações do sistema de escrita precisam se transformar em dados numéricos. Adiante iremos discutir essa etapa de processamento, para defender que o conceito do texto digital deve partir delas.

2. Funcionamento do texto digital

2.1 Aspectos básicos

O funcionamento da mediação artificial da informação na construção (e na difusão) do texto digital é contingência de um fato simples e primordial: os processamentos digitais se dão com base em códigos binários, mas os produtores e receptores humanos não lidam diretamente com códigos binários. Assim, a operação de base do processamento digital

é a transformação da informação gerada pela lógica humana em informação binária (e depois, reversamente, a transformação da informação binária em informação legível pela lógica humana).

Na base de todo texto digitalmente processado está um sistema de equivalências a que se denomina "codificação de caracteres" – e que consiste, fundamentalmente, em correspondências entre símbolos e sequências numéricas. Há atualmente dois sistemas de correspondências principais: o *American Standard Code for Information Interchange,* ASCII (desenvolvido desde a década de 1960) e o *Unicode Transformation Format*, UTF (desenvolvido desde a década de 1990); abaixo, por exemplo, está parte da tabela de correspondências ASCII*:*

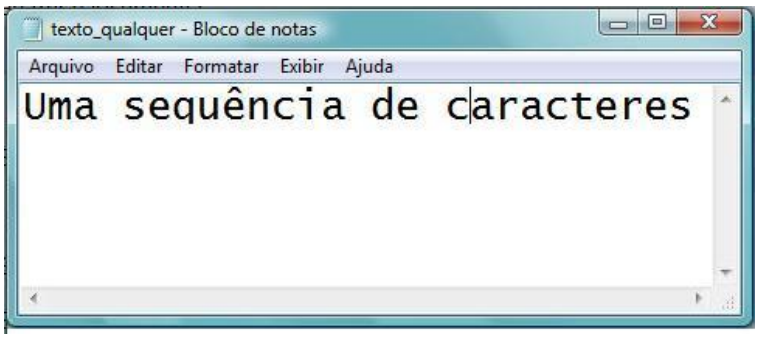
Tabela 1: Correspondências ASCII (@ - D)

Binário	Decimal	Hexa	Glifo
0100 0000	64	40	@
0100 0001	65	41	A
0100 0010	66	42	B
0100 0011	67	43	C
0100 0100	68	44	D

Todo e qualquer arquivo digital de texto que circula hoje no mundo estará codificado por um desses sistemas de correspondência de caracteres. Naturalmente, a informação é efetivamente transmitida no meio digital (de um computador a outro, de um arquivo de computador a outro arquivo no mesmo computador, ou de um computador para uma impressora) sob forma de sequências de códigos como os da primeira coluna da tabela acima, sendo a correspondência entre os códigos na primeira coluna e os glifos na última coluna estabelecida já em uma etapa de pós-processamento.

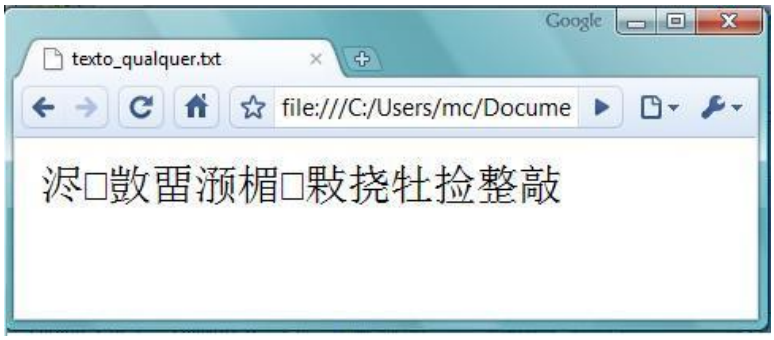
Tanto é assim que uma mesma sequência de caracteres de entrada pode receber diferentes visualizações a depender da programação do aplicativo de saída final, como ilustram as figuras (1) a (3).

(Figura 1) Codificação ANSCII



(Figura 2) Codificação Windows 1252 - Coreano

(Figura 3) Codificação UTF-16



As figuras (1) a (3) representam diferentes visualizações de um mesmo arquivo de texto simples (.txt), no qual uma sequência de caracteres ("*Uma sequência de caracteres*") foi registrada e gravada como ANSCII, no aplicativo "Bloco de Notas". (1) é a visualização do arquivo no próprio aplicativo "Bloco de Notas"; (2) é sua visualização em um navegador (Google Chrome) com a codificação de caracteres programada para "Ocidental – Windows 1272"; (3) é a visualização do mesmo arquivo no mesmo navegador, mas com a codificação de caracteres programada para "UTF-16". Ou seja: (2) e (3) são diferentes versões produzidas por um programa de visualização para um mesmo arquivo de texto simples.

Essas imagens sugerem uma pergunta interessante: ainda que (1),(2) e (3) correspondam a um mesmo arquivo digital, podemos dizer que são "o mesmo texto"?

Do ponto de vista da leitura humana, (1), (2) e (3) representam, claramente, “textos diferentes”. Tanto é assim que um leitor humano com conhecimento do alfabeto romano, por exemplo, consegue ler perfeitamente a escrita das Figuras (1) e (2), mas não a da Figura (3). De um ponto de vista estritamente computacional, entretanto, (1), (2) e (3) representam, claramente, o “mesmo texto”: (2) e (3) são diferentes versões de uma mesma entrada de escrita (1) – uma mesma sequência de caracteres, apenas decodificada por diferentes vocabulários.

Podemos tomar a contradição assim formada como fio condutor a nos guiar pelo caminho inicial de uma exploração conceitual do “texto digital”, para propor mais adiante que seu conceito deve incluir a dimensão da leitura humana e a dimensão do processamento artificial.

O processamento artificial funciona como uma cadeia complexa de composições e decomposições de informação binária, que “quebram” a corrente daquilo que, nas mentes humanas, chamamos de “leitura”. Depois de “quebradas” – ou seja, depois de decompostas em pedaços mínimos de informação binária – as informações precisarão ser recompostas em linguagem legível na interface. No caso do processamento de textos, e tomando ainda o exemplo dos caracteres: os caracteres decompostos em códigos binários precisam ser recompostos sob forma de caracteres, e então sob forma de glifos humanamente legíveis.

Entre a "transformação" dos caracteres em códigos binários (numa ponta), e a "transformação" do código binário em caracteres (na outra ponta), operam os programas denominados "compiladores".

Os compiladores são programas ou conjuntos de programas que "compilam" as múltiplas informações codificadas numericamente nos arquivos em informações humanamente legíveis, fundamentalmente operando traduções de códigos binários complexos em códigos-objeto mais simplificados.

Para se ter uma idéia das operações envolvidas numa compilação, remeto ao esquema básico da figura (5), reproduzida de (REFS); para descrições detalhadas mais adequadas, remeto à bibliografia pertinente produzida no âmbito da Ciência da Computação – por exemplo, (http://www.di.uminho.pt/~jcr/XML/publicacoes/teses/phd-jcr/src/c508.htm).

Figura (4)

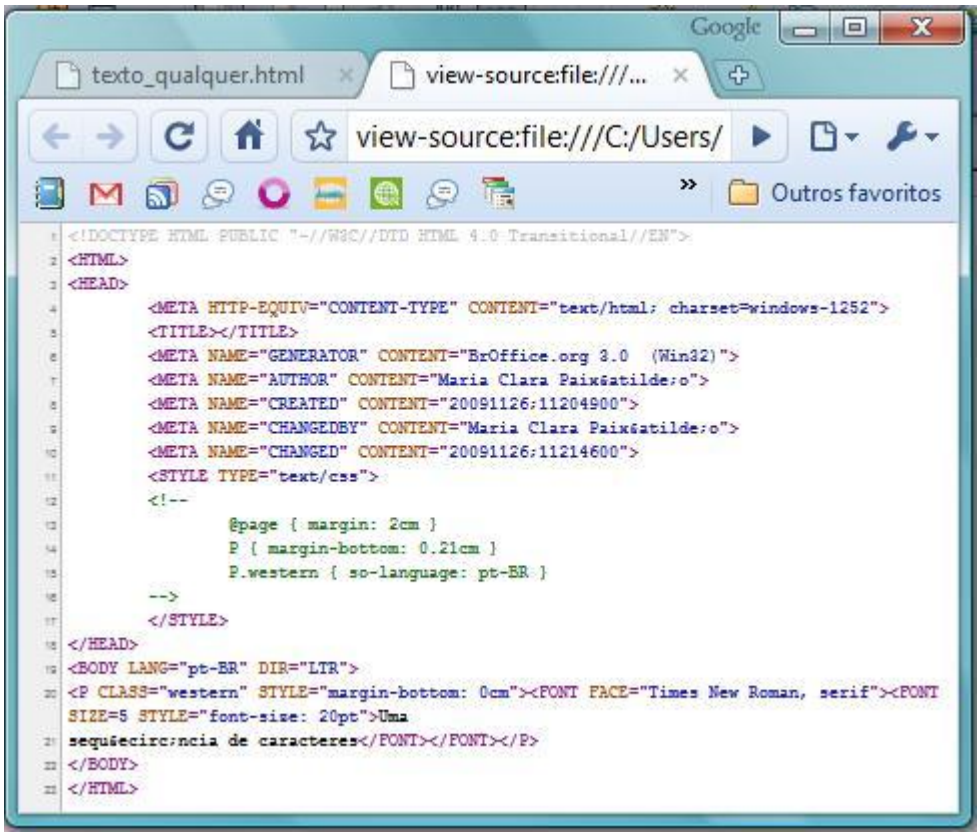
Neste momento, não nos importará tanto compreender *como* essa compilação acontece – mas sim onde e quando ela acontece.

Notemos que a compilação pode ser local ou remota; acessível ou inacessível. Como exemplo de compilação local e inteiramente inacessível, podemos citar aquela oferecida pela grande maioria dos chamados "processadores de texto" ou "editores de texto" comerciais: "Word", da Microsoft; "Open Office" da Sun; etc. Com maior ou menor grau de abertura do código de programação, e maior ou menor grau de sofisticação de compilação sustentada, esses aplicativos têm em comum o fato de manipularem a informação codificada em um módulo de programação embutido no aplicativo – e ao qual o "editor humano" não tem acesso.

No lado oposto, estão as compilações remotas e parcialmente acessíveis: podemos citar aquela que se oferece sob a denominação de "navegador" ou "browser" – a ferramenta de acesso à "internet"; um navegador é um compilador (e ao

mesmo tempo, um visualizador). Os navegadores, por serem compiladores remotos, permitem uma maior visibilidade da compilação – ao menos, nessas ferramentas, um usuário médio pode ter uma mínima percepção de que o texto que ele lê é na realidade composto por um conjunto de informações artificialmente codificadas que podem se transformar, se modificar, no processo de sua compilação. Por exemplo, qualquer navegador comercialmente disponível hoje permite a visualização dos "códigos-fonte" dos arquivos visualizados. O código-fonte nada mais é que a etapa final da compilação, um código-objeto simplificado, que sofrerá apenas uma última manipulação artificial antes de ser visualizado pelo leitor humano. Na figura (5) temos o exemplo do código-fonte do arquivo .html gerado a partir do mesmo documento de texto visualizado nas figuras (1) a (3) mais acima.

Figura (5)



Nesta figura podemos observar um último elemento importante na compreensão do texto digital: as anotações do código-fonte. Notemos, a esse respeito, que para que a compilação funcione, o material a ser compilado precisa oferecer informações codificadas legíveis para o compilador. Essa

codificação pode conter instruções mais ou menos complexas: da informação mínima num arquivo de texto (como já vimos, a a sequência de caracteres) até uma estruturação lógica completa do documento. Isso inclui, por exemplo, informações relativas à formatação, disposição do texto em uma unidade espacial, etc. Essas informações funcionam como instruções para os compiladores porque estão anotadas nos arquivos em determinada linguagem de anotação legível por determinados compiladores. No exemplo da Figura (5), um texto anotado em linguagem HTML (Hypertext Markup Language), vemos a anotação de instruções como "abrir um parágrafo" (o código <P>...</P>), "tamanho da fonte" (o código <FONT SIZE=5>...</FONT>), entre outros.

Assim, todo texto produzido em meio digital que contenha alguma informação para além da simples sequência de caracteres (de quebras de linha a correspondências complexas do tipo bases de dados) é um texto que encerra processos de anotação em alguma linguagem artificial, para poder, desta forma, participar daquela etapa "intermediária" de processamento lógico artificial que se interpõe ao processamento natural – humano – da informação.

Podemos agora abordar mais detidamente dois pontos já sugeridos de forma introdutória.

O primeiro deles é o problema da transparência dos processamentos: como já foi mencionado, a etapa artificial de processamento pode permanecer inteiramente opaca ao produtor e ao receptor humano do texto digital. De fato, mesmo a codificação de caracteres, variável-chave e processo básico para o processamento artificial da informação, é um estágio normalmente que só costuma se tornar evidente quando um dos elos da corrente de codificação deixa de funcionar corretamente, fazendo com que notemos que "*por trás do texto*" opera um processo lógico-matemático. O processo, entretanto, opera sempre. E de um ponto de vista estritamente computacional, como vimos, é o fato de um arquivo conter este tipo de informação codificada (em ASCII, UTF, etc) que o torna arquivo um documento de texto.

Podemos então voltar à pergunta sugerida mais acima: visualizações de um mesmo arquivo digital que se apresentem diferentes para a leitura humana ainda podem ser

consideradas como "o mesmo texto"? Voltando à figura (5) e às figuras (1) a (3) mais acima, podemos imaginar que ao ser colocado diante de um navegador que produzisse a visualização em (3), o produtor de um texto inicialmente visualizado como (1) provavelmente exclamaria: "*Mas não foi isso que eu escrevi! Este não é o meu texto!*"... Tudo o que ele fez foi pressionar teclas e fazer aparecerem letras na tela, sem se dar conta de que o que ele estava compondo, de fato, era um código a ser processado por uma máquina, e que pode ser visualizado como (1) e (2) ou como (3).

Assim, de um lado, a sequência codificada de caracteres constitui materialmente o texto (é o que, de fato, se constrói na ação da escrita) – e portanto, não pode ser dissociado do conceito "texto". De outro lado, já que a apresentação do texto constitui o que os leitores humanos processam, ela tampouco pode ser dissociada do conceito de texto.

Podemos sair deste labirinto conceitual aceitando o problema nos seus dois planos inseparáveis, abordando os "textos digitais" como camadas de informação matemática e informação humana, as quais, combinadas, formam o que percebemos como "o texto". Definiremos então o texto "digital" como o texto cujo processo de difusão envolve a codificação de informação por linguagens artificiais, e que se constitui materialmente como informação linguística codificada matematicamente e apresentada com a forma de escrita humanamente legível.

2.2 Breve tipologia

Buscamos, até este ponto, um conceito de "texto digital" que tomasse a singularidade do processamento artificial da informação como o ponto central. A definição assim esboçada permite elaborar uma tipologia interessante do texto digital, como se sugere aqui. Notemos, antes de tudo, que a definição procura abarcar a propriedade de "digital" na dimensão do processo e na dimensão do produto. Isso nos permite considerar textos parcialmente processados no meio digital – por exemplo, produzidos por intermédio de um aparelho eletrônico, mas recebidos em suporte não-eletrônico

(como um texto impresso a partir de um arquivo de computador); veremos, adiante, porque isso seria desejável.

Além da consideração da forma de difusão predominante do texto (parcialmente digital; globalmente digital), nossa tipologia procura sistematizar o funcionamento do processamento artificial da informação, em especial considerando o grau de manipulação explícita possível da anotação das informações. Combinando esses dois aspectos, compõe-se o critério norteador da tipologia: o grau de transparência do processamento artificial da informação.

2.2.1 Processos e Formatos mais transparentes

Os processos de codificação de informação mais transparentes estão tipicamente associados aos formatos de processamento digital global – i.e., aos textos produzidos em meio digital e programados para circularem no meio digital até a etapa de recepção. Este grupo de textos cuja cadeia de difusão está globalmente inserida no meio digital corresponde, de fato, aos textos reconhecidos pelo senso comum como "digitais" e assim cotidianamente denominados – ou seja: os textos que são "lidos no computador", por exemplo por meio da internet.

O hipertexto é o representante mais evidente e mais característico desta classe de textos globalmente digitais. Incluem-se na "família hipertexto" os diferentes formatos associados às linguagens de marcação padronizadas pelo consórcio WWW: o HTML ("HyperText Markup Language"), o XML ("eXtensible Markup Language") e o XHTML ("eXtensible HyperText Markup Language").

Podemos salientar duas propriedades que distinguem essa família de textos – propriedades estreitamente relacionadas entre si, por sinal. A primeira propriedade é a independência em relação a aplicativos particulares. De fato, a concepção do hipertexto tem por base a possibilidade de disseminação em qualquer compilador ou interpretador remoto. Notemos, portanto, que o fator "processamento global" e o fator "independência de aplicativos" estão ligados: é por serem idealizados para o processamento global em meio digital que os textos desta família não podem depender de interpretadores localizados particulares. A

grande ferramenta para a leitura, nestes casos, é um compilador e visualizador remoto – aquilo a que chamamos navegador ou "browser", como já vimos. Existem hoje diversos navegadores preparados para compilar e tornar humanamente legíveis os códigos html, por exemplo, e o espírito do desenvolvimento desta linguagem é tornar-se legível por qualquer um deles. Assim, o padrão da linguagem de anotação é regulado por um consórcio mundial aberto, o www (REF).

Ao considerar que o hipertexto é independente com relação a aplicativos particulares, não estamos esquecendo que, atualmente, há um grande número de aplicativos que permitem editar hipertextos. O ponto, aqui, é salientar que o hipertexto pode ser construído dentro ou fora desses aplicativos, por qualquer editor humano conhecedor do código (que é aberto), em qualquer aplicativo que possa processar um documento de texto simples.

Essencialmente, importa salientar o fato de que a codificação da informação, no hipertexto, configura-se como relativamente mais transparente para o produtor e para o receptor do texto. Mais adiante levantaremos algumas consequências desta propriedade.

2.2.1 Processos e Formatos mais opacos

Os formatos de textos digitais com o menor grau de transparência nos processos de codificação estão tipicamente associados ao processamento parcial no meio digital.

Mais acima já ressaltamos que queremos incluir nesta tipologia os textos parcialmente processados no meio digital; e a relevância disso se torna clara já de partida quando consideramos as classificações tradicionais da crítica textual para o texto não-digital. De fato, a classificação material analisa tradicionalmente produção e difusão, remetendo ao processo global da produção dos artefatos, não apenas ao artefato final. Por exemplo: tradicionalmente, um artefato do tipo "reprodução mecânica de um manuscrito" (uma fotografia, um microfilme, etc.) continua sendo classificado como um "manuscrito" para fins de estudo filológico, e não como um "texto mecânico". Assim, até o ponto em que um texto parcialmente produzido em meio

digital chegue a se apresentar como o artefato "folha de papel com sinais gráficos estampados", algumas das contingências importantes da análise do texto digital se aplicarão a ele. Logo abaixo voltaremos a este ponto.

Notemos que, reversamente, um artefato do tipo "reprodução digital de um manuscrito ou texto impresso" (por fotografia digital ou escanerização) tampouco deve ser considerado um texto digital, mas apenas e simplesmente, como a reprodução digital de um texto manuscrito ou texto impresso. Noutros termos: nossa definição e nossa tipologia deixam de fora textos cujo processamento não envolve codificação da sequência de caracteres, embora sejam produzidos sob a aparência de um arquivo digital - caso da fotografia digital de um texto impresso ou manuscrito.

Assim, falamos, aqui, dos textos produzidos num ambiente digital, mas pensados para serem recebidos primordialmente fora do ambiente digital, por exemplo impressos. Este grupo inclui portanto a maioria dos formatos que podem ser produzidos nos aplicativos de processamento de textos mais usados (.doc, .odt, etc). Esses formatos apresentam três características, presentes em maior ou menor grau, e relacionadas entre si: a remissão a meios não-eletrônicos; a obliteração do processamento artificial; e a dependência do aplicativo.

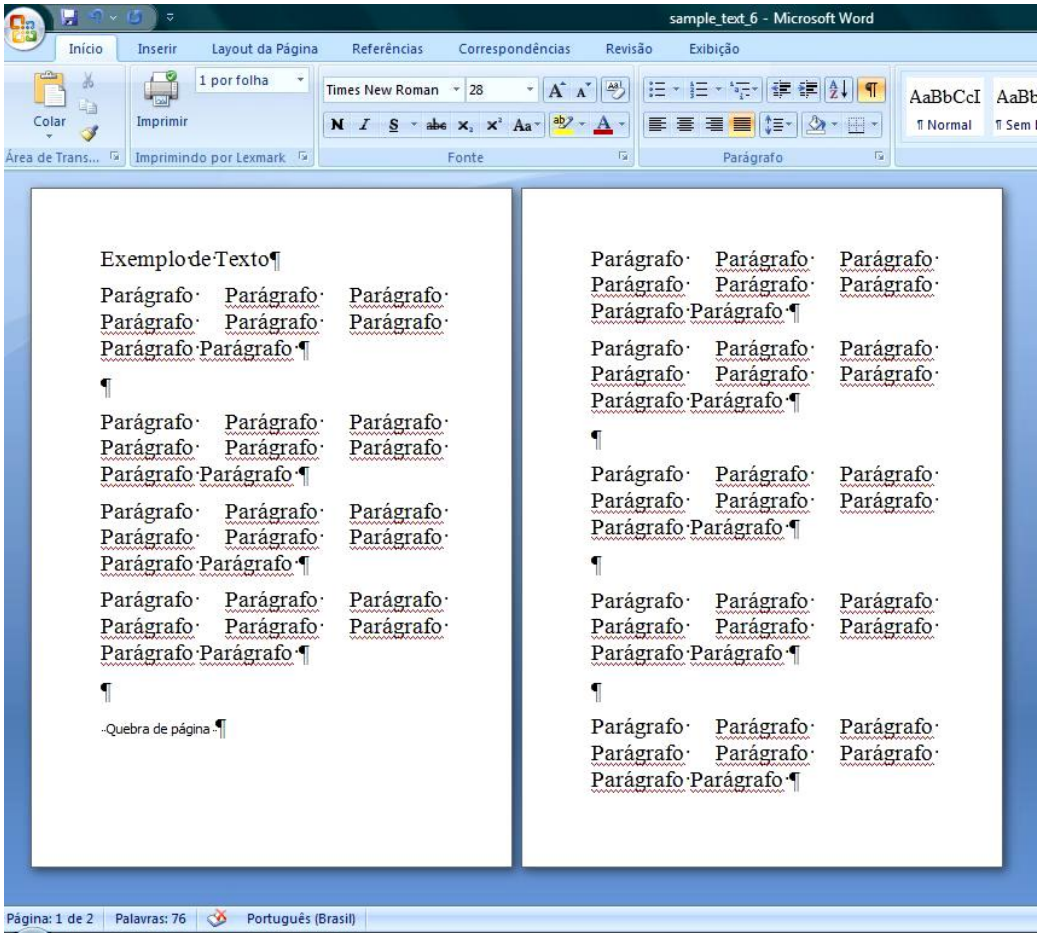
Falamos de "remissão a meios não eletrônicos" para descrever o fato de que nesses formatos de texto, o objetivo final é produzir artefatos o mais próximos possível aos textos convencionais, tipicamente em papel. Esta característica está estreitamente relacionada aos problemas de difusão neste grupo de textos, como veremos.

Um bom exemplo da questão da remissão a meios não eletrônicos é a idéia de "página". A "página", naturalmente, é uma unidade espacial intrinsecamente ligada ao mundo do papel. Entretanto, os processadores de texto atuais costumam exibir uma tela com uma "página" na qual o usuário vai "escrever" o texto. Esses espaços que reconhecemos como "páginas" na tela dos nossos processadores são, evidentemente, representações visuais fabricadas por códigos; e os usuários permanecerão insensíveis a este fato enquanto a representação funcionar sem problemas. Apenas em

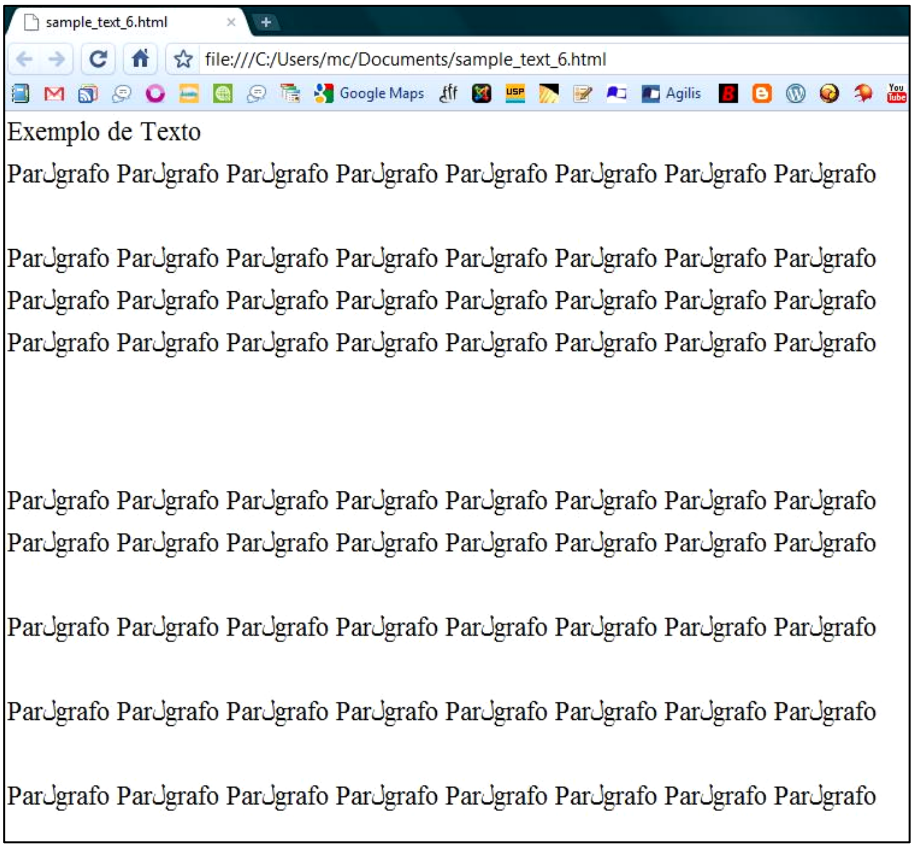
algum momento de "erro de reprodução" o usuário se dará conta de que a "página" que "*estava ali*" já não está – caso típico daquele momento em que se envia um texto de um processador para uma impressora, ou de uma máquina para outra, e a "paginação" se altera inteiramente.

Embora a página seja um exemplo dos mais evidentes, importa lembrar que este processo de representação codificada das categorias visualmente perceptíveis (e sua obliteração, como veremos) envolve toda a construção gráfica dos textos nos processadores comuns. A figura (6) abaixo mostra a tela de um processador "Word" da marca Microsoft, exibindo um documento com "duas páginas", nas quais uma sequência de caracteres está disposta em diferentes "parágrafos". A Figura (7) mostra o mesmo arquivo visualizado em um navegador (Google Chrome); notem-se os sumiços das "páginas", a corrupção dos caracteres, e a disposição alterada dos "parágrafos":

(Figura 6)



(Figura 7)



Já neste ponto se revela a segunda característica dos documentos neste grupo: a obliteração dos processamentos artificiais. De fato, nos processadores mais modernos, todo o esforço é concentrado em tornar a produção do texto o mais intuitiva possível para o usuário – quase como se ele estivesse diante de uma máquina de escrever. Na verdade, a remissão a meios não-eletrônicos ou (se quisermos) "dependência do papel" parece estar na raiz do esforço de obliteração dos processamentos eletrônicos. É deste modo que hoje, qualquer pessoa letrada pode fazer uso de um processador eletrônico de textos, sem necessidade de conhecer programação matemática. Este é certamente um progresso desejável. Notemos, entretanto, que quanto mais "intuitivo" parecer um processo de construção de texto num processador, mais sofisticada será a codificação matemática que o possibilita - e portanto: mais elevado será o risco potencial de corrupções do texto. Assim, os formatos ligados a aplicativos de processamento encerram uma complexidade elevada de processo lógicos artificiais intermediários opacos ao usuário.

Chegamos então à terceira característica: sobre esses processos lógicos intermediários, o usuário tem pouco controle, ou até, nenhum controle. Em diferentes graus, os formatos processados por aplicativos são dependentes dos aplicativos – gerando, por exemplo, aquele problema bem conhecido da impossibilidade de abrirem-se determinados formatos em determinados processadores. Nesta questão da dependência do aplicativo, esse tipo de formato pode ainda ser separado em formatos de código fechado e formatos de código aberto – nesses últimos, o código pode ser conhecido e manipulado por especialistas. Mas nos dois casos, a codificação (i) não se dirige à manipulação do usuário médio e (ii) está embutida em compiladores privativos aos aplicativos de codificação. A seguir comentamos algumas consequências dessas propriedades.

2.2.2 Resumo e implicações

Em resumo, sugerimos uma tipologia de formatos de textos digitais guiada pelo critério básico da transparência da codificação, compondo uma escala composta por três aspectos:

(i) o grau de independência da linguagem de anotação com relação a aplicativos de processamento;
(ii) a abertura da linguagem de anotação;
(iii) a possibilidade de manipulação da anotação pelo editor.

Podemos formar então o seguinte quadro esquemático, com alguns exemplos de tipos de formatos mais conhecidos:

	Formato		Independente	Aberto	Manipulável
+ transparência -	Extensible Markup Language	.xml	+	+	+
	Hypertext Markup Language	.html	+	+	+
	Open Document Text (Sun)	.odt	-	+	-
	Document (Microsoft)	.docx .doc	-	-	-
	Portable Document File (Adobe)	.pdf	-	-	-

O primeiro e mais evidente impacto dessa escala de transparência é o problema da longevidade dos formatos. De fato: quanto menos transparente for um tipo de texto (menos independente, menos aberto, menos manipulável), menor será sua longevidade como arquivo digital. Evidentemente, o acesso e o processamento das informações contidas em arquivos de tipo opaco só podem ser realizados por meio dos aplicativos particularmente dedicados a esses arquivos; isso significa que, se tais aplicativos deixarem de estar disponíveis, aqueles arquivos podem se tornar inacessíveis e improcessáveis. É interessante notar que embora boa parte dos usuários de processadores de textos eletrônicos já tenha, em algum momento, passado pela frustrante experiência de "perder" textos por conta da obsolescência de alguns aplicativos comerciais, a maioria desses usuários continua confiando aos novos aplicativos a possibilidade de acessar os arquivos armazenados digitalmente. Entretanto, a única forma segura de se conservar um arquivo de texto processado por aplicativos deste tipo – ironicamente – é imprimindo seu conteúdo e armazenando esta cópia em papel. Esta solução, entretanto, tira muito do sentido de se usar o meio eletrônico como ferramenta de armazenamento e gerenciamento da informação.

Uma segunda consequência prática desse problema da transparência dos formatos remete às potencialidades do meio digital como ferramenta para o trabalho mais especializado de edição de textos, como a edição filológica. Para a filologia, o trabalho de edição é uma ferramenta central de trabalho – torna-se desejável, aí, o maior controle possível sobre as informações a serem codificadas em um texto escrito, e o maior grau de confiança possível em sua preservação. Conforme já salientei em outras oportunidades (Paixão de Sousa, 2005, 2007), é nos formatos mais transparentes, de codificação manipulável, que o trabalho filológico encontra uma ferramenta adequada de trabalho.

Por fim, há um terceiro aspecto interessante a ser discutido a partir desta tipologia: ela coloca em primeiro plano as questões envolvidas na difusão do texto digital, permitindo

um tratamento teórico interessante dessa cadeia do ponto de vista da crítica textual, como se sugere a seguir.

2.3 A Cadeia de difusão do texto digital

Para serem espalhados no tempo e no espaço, os textos precisam ser reproduzidos. No processo de sua reprodução, os textos "se alteram", "se transformam", "mudam", "movem" – esta é a contingência fundante da arte e técnica da crítica textual (Blecua, 1983; Cambraia, 2005).

As transformações que um "texto" pode sofrer no processo de suas reproduções são o fato central deste campo de estudos que tradicionalmente se dedicou a compreender os processos de transmissão dos textos escritos e desenvolveu técnicas para recuperá-los das alterações produzidas nessa transmissão. A crítica textual definiu uma tipologia de fatores de interferência nas cadeias de reproduções, sejam fatores relativos à deterioração material (i.e., "exógenos"), sejam fatores relativos ao próprio processo de reprodução (i.e., fatores "endógenos"), de modo a reconhecer, para cada processo de produção e reprodução, as diferentes limitações e possibilidades de alteração, resultando em diferentes graus de complexidade de difusão.

Nesta seção, tentaremos compreender como o texto digital se compara aos demais tipos de texto quanto ao problema da difusão.

Notemos, antes de tudo, que a crítica textual tradicional se ocupou sobretudo da difusão manuscrita, na qual se reconhece o grau máximo do potencial de interferência de fatores endógenos na cadeia de reproduções (é o caso dos famosos "erros de cópia"). De fato, se considerarmos o caso da difusão do texto manuscrito de perto, veremos que cada instância de reprodução manuscrita precisa incluir todas as etapas do processo de produção, configurando uma cadeia de difusão de máxima complexidade. Vamos lembrar o processo resumido no quadro (1) da seção 1: o lado da "codificação da informação" incluía duas etapas, (i) associação informação – sinal gráfico; (ii) traçado do sinal gráfico. Ora, numa cadeia de reprodução manuscrita, cada instância de reprodução precisa incluir também esses dois passos.

Noutros termos, a cópia manual de um texto não difere essencialmente da composição manual de um texto. Comparemos isso com a reprodução mecânica: nesse caso, cada "cópia" pode prescindir de algumas das etapas do processo de composição. Num texto impresso, a composição da matriz se distingue fundamentalmente da reprodução da matriz uma vez que enquanto a primeira envolve processos lógicos e mecânicos, a segunda consiste em um processo puramente mecânico.

Em resumo, quanto mais complexa a cadeia de reproduções, mais vulnerável à alteração. E de fato: a importância do processamento lógico na reprodução será proporcional ao risco de erro de reprodução (e por isso, é infinitamente menos provável uma prensa mecânica reproduzir mal um original que um copista humano).

Pensemos, agora, na cadeia de reproduções digital, e veremos que nela o problema da "cópia" ganha um significado interessante.

Vamos antes pensar: como se dá a reprodução em meio digital, e como ela se compara à reprodução manual e à reprodução mecânica, que já examinamos esquematicamente? Novamente (como nos aspectos relativos à construção do texto), o processo de reprodução digital parece, à primeira vista, ser análogo ao processo mecânico, mais que ao manual. Abrimos um computador, e lemos um texto; abrimos outro computador, e podemos ler a cópia deste mesmo texto – aparentemente, a "máquina-computador" está reproduzindo esse texto de modo análogo à "máquina-prensa". Mas novamente a analogia é ilusória. Se examinarmos a reprodução digital pelo mesmo prisma que examinamos a reprodução manual e a mecânica – pelo prisma das etapas envolvidas na reprodução, em comparação com as etapas envolvidas na produção da "matriz" – veremos que na reprodução digital, a cada reprodução diversas etapas da produção precisam ser repetidas; não apenas etapas mecânicas, mas também (crucialmente) etapas de processamento lógico.

Aqui voltamos a encontrar o problema da transparência dos processamentos, obrigando-nos a procurar despir o processo de reprodução digital de forma crua.

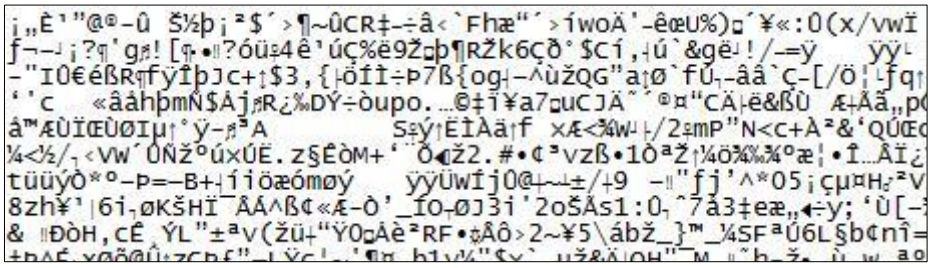
Tomemos o processo de reprodução que se dá naquilo a que chamamos "internet". Todo texto que lemos nas nossas telas de computador quando "*acessamos a internet*" é, naturalmente, uma cópia: uma reprodução visualizável de um código-fonte que está armazenado em algum computador com o qual o nosso computador "*entrou em rede*". Notemos: não é que diretamente estejamos lendo o texto que está em outro computador: estamos lendo uma versão daquele texto, produzida por uma programação embutida na nossa máquina (a programação compiladora do "navegador"). Esse fato fica claro quando pensamos que um "*mesmo texto*" pode estar sendo lido por milhares de pessoas num mesmo ponto do tempo, em diferentes pontos no espaço, e ainda assim apresentar diferentes características de aparência para cada leitor – já que cada leitor estará lendo uma versão daquele texto por intermédio de uma programação potencialmente diferente.

Este ponto precisa ser salientado: no meio digital, os textos são sempre difundidos como cópias produzidas por programações, ou seja, cópias que incluem etapas de processamento lógico. Na situação mais característica de difusão do texto digital – a difusão remota, i.e., os textos recebidos via rede de computadores – aquilo a que os leitores têm acesso é uma versão do código-fonte remotamente armazenado. O que se armazena num "servidor" (i.e., a máquina remota à qual às máquinas interligadas terão acesso) são arquivos com os códigos-fontes, programados para aparecerem como textos quando forem processados por leitores artificiais ("navegadores") para se tornarem legíveis aos leitores humanos. Mas o mesmo acontece no acesso local: a sequência de caracteres que conseguimos ler na tela dos nossos computadores, nos arquivos que armazenamos ali, e que "*nós mesmo escrevemos*", são também cópias: cópias do código-fonte que de fato está armazenado, e que se traduz para nossa leitura humana por um programa (por exemplo, por um editor de texto) a cada vez que abrimos o arquivo.

Novamente, esse processo de reprodução é opaco para o leitor médio – a consciência de que aquilo que ele lê não é exatamente aquilo que ele "*escreveu*" só emerge nos momentos de erro de cópia. Todo usuário de computadores já passou pela situação em que um texto que havia sido perfeitamente

composto (com determinados títulos, tabelas, quebras de páginas...) na versão (a) de um editor eletrônico, ao ser aberto na versão (b), passa por uma alteração exasperante como esta ilustrada na figura (8) (que representa o que se visualiza ao se abrir um arquivo .txt composto com a a sequência de caracteres "*Exemplo de Texto*" no aplicativo Word 2007, salvá-lo como .docx, e tentar abri-lo novamente no aplicativo Bloco de Notas):

(Figura 8) "Exemplo de texto" – Codificação ANSII, arquivo .docx aberto em aplicativo Bloco de Notas



Como já discutimos, os momentos em que os processos não funcionam como o esperado nos revelam que há algo operando entre nossa escrita e a apresentação final do texto no meio digital; mas esse "algo" (a mediação artificial da programação que intervém entre as etapas do nosso processamento lógico humano) está operando sempre, na composição e na cópia digital. Voltamos então à afirmação mais acima sugerida: no meio digital o problema da "cópia" ganha um significado interessante.

De fato, a cópia digital, num certo sentido, se aproxima mais da cópia manual que da cópia mecânica, pois o processo de cópia digital inclui o re-processamento de etapas lógicas, de modo análogo à cópia humana, e ao contrário da cópia mecânica. A diferença, aqui, é que serão as etapas lógicas artificiais que precisam se repetir. O ponto importante dessa comparação entre o processo de reprodução digital e os processos manual e mecânico é notarmos que a reprodução digital encerra um grande potencial de alterações por cópia, que remetem às etapas de processamento lógico envolvidas na difusão digital.

Futuras abordagens teóricas sobre o texto digital do ponto de vista da crítica textual precisarão, assim, levar em

conta a programação e codificação matemática como um estágio na cadeia de difusão deste tipo de texto.

3. Por fim

Ainda não chegamos ao momento em que o "texto digital" pode revelar seu potencial completo, nem enquanto artefato a ser desenvolvido tecnicamente, nem enquanto objeto a ser explorado conceitualmente.

Na sua dimensão de artefato, o texto digital está ainda amarrado a *idéias de papel*, tais como "páginas", "linhas"... Parecemos estar ainda testemunhando uma fase de transição, curiosamente evocadora das primeiras décadas que se seguiram ao advento da imprensa, quando o grande objetivo dos melhores impressores era replicar, o mais fielmente possível, os textos manuscritos (CHARTIER, 2001; EISENSTEIN, 1998).

Da mesma forma, enquanto objeto teórico, o texto digital parece ainda não se apresentar de modo independente ao horizonte dos campos que tradicionalmente se ocuparam ao estudo da difusão dos textos e suas questões correlatas.

É verdade que, no plano metodológico, podemos observar nos anos recentes um desenvolvimento interessante da relação entre estes campos tradicionais e o texto digital. Diferentes iniciativas de pesquisa ao redor do mundo vêm buscando uma aproximação entre as técnicas tradicionais da edição filológica e as técnicas de anotação e processamento digital; em especial, a anotação por linguagens de marcação como o XML, por meio da qual inúmeras camadas de informação podem ser condensadas em um mesmo documento, têm permitido a captura e a análise de diferentes planos de interesse para os estudos do texto: a organização gráfico-visual, a estrutura linguística (léxico, morfologia, sintaxe, semântica), as marcas de diferentes estágios de difusão – verticalmente em um mesmo texto, ou horizontalmente em relação a outros textos. É mesmo possível vislumbrarmos, nessa exploração horizontal e vertical dos textos, o coroamento dos sonhos do crítico textual em tempos imemoriais - no seu ofício incansável de anotar as diferenças entre os diversos testemunhos de uma tradição,

perseguindo as marcas dos editores passados, desenhando os sinais para os editores futuros.

Para chegarmos, entretanto, a uma revelação conceitual do texto digital, precisamos ainda ultrapassar alguns desafios.

Antes de tudo, seria necessário o abandono definitivo da noção do ambiente "digital" como um novo tipo de "suporte" de texto, como se o processamento digital se equiparasse aos avanços técnicos da tecnologia anterior - a imprensa, ou a máquina de escrever. Essa noção impede a observação da característica singular do texto digital: a inclusão de etapas lógicas artificiais de processamento de informação. Além disso, a filologia e a crítica textual, se desejarem abraçar este novo objeto "texto digital", precisariam ampliar seu escopo de interesses de modo a incluir o fenômeno do processamento artificial da linguagem, passando a considerá-lo como um campo afim, da mesma forma como consideraram, classicamente, a paleografia, a diplomática, ou a codicologia.

Este artigo procurou contribuir para este cenário futuro de uma sistematização conceitual do texto digital, fundamentalmente chamando a atenção para sua singularidade frente aos demais tipos de texto. Localizamos esta singularidade na etapa do processamento da informação, e não simplesmente em seu registro ou transporte. Propusemos assim uma definição inicial na qual o texto digital constitui-se materialmente na combinação entre informação linguística artificialmente processável e informação linguística humanamente processável, em sucessivas camadas que são percebidas como "texto".

Para finalizar este ensaio, quero lembrar as palavras de Rosa Borges de Carvalho, para quem as diferentes formas de "filologia" ao longo da história têm em comum a propriedade de tomarem o texto como "...*um sistema de raízes que pode ser escavado nele próprio*" e fazerem, daí, aflorarem as mais diversas abordagens (CARVALHO, 2003) – para sugerir que, na espiral lógica das camadas de informação encerradas nos textos digitais, um potencial renovador aguarda os campos que se dedicam à escavação desse objeto infinito, o texto.

REFERÊNCIAS

BLECUA, Alberto. Manual de crítica textual. Madrid: Castalia, 1983 [1987].

CARVALHO, Rosa Borges Santos (2003). "A Filologia e seu Objeto: Diferentes perspectivas de estudo". Philologus - Revista do Círculo Fluminense de Estudos Filológicos e Lingüísticos, ano 9, n. 26, Rio de Janeiro. http://www.filologia.org.br/revista/artigo/9(26)03.htm

CHARTIER, R (2001). Cultura Escrita, Literatura e História. Porto Alegre: Artmed.

EISENSTEIN, Elizabeth (1998). A Revolução da cultura impressa. São Paulo: Ática.

IDE, Nancy and ROMARY, Laurent (2000): "XML Support for Annotated Language Resources". Linguistic Exploration: Workshop on Web-Based Language Documentation and Description. Dec 12 - Dec 15, 2000, University of Pennsylvania.

LESK, Michael (1997). "Hypertext". Practical Digital Libraries. New York: Morgan Kaufmann.

PAIXÃO DE SOUSA, Maria Clara (2006). "Edições Críticas Eletrônicas: Fundamentos e Diretrizes". www.ime.usp.br/~tycho/participants/psousa/memorias/critical_hyper/ece.html

TRIPPEL, Thorsten and PAIXÃO DE SOUSA, Maria Clara (2006). "Building a historical corpus for Classical Portuguese: some technological aspects". Papers from the V International Conference on Language Resources and Evaluation, Genoa: LREC.

TBACHP (2006). "Tycho Brahe Annotated Corpus of Historical Portuguese". <http://www.ime.usp.br/~tycho/corpus>

W3C (1997a). "Extensible Markup Language". http://www.w3.org/XML.

W3C (1997b). "HyperText Markup Language".http://www.w3.org/MarkUp.