COMPARA

Instruções para a revisão do Reconhecimento Óptico de Caracteres (ROC) e etiquetagem preliminar

Ana Frankenberg-Garcia
Primeira versão: 19 de Setembro de 2002 [última actualização: 4 de Novembro de 2007]

Para que um texto possa ser pesquisado num corpus *electrónico, tem de estar em formato digital. Se o texto não está disponível em formato digital, tem de ser passado a computador ou digitalizado através de um* scâner *e submetido a um programa de ROC. O* scâner *funciona como uma fotocopiadora, ou seja, tira uma fotografia do texto. Um programa de ROC transforma o texto que foi fotografado pelo* scâner, *em texto que pode ser trabalhado num processador de texto. Depois de um texto passar por um programa de ROC, é possível acrescentar ou apagar palavras, mudar o tipo de letra, etc. A primeira etapa de preparação de textos para o COMPARA aqui descrita consiste em (1) corrigir problemas típicos de ROC, (2) retirar as partes do texto que não são precisas, e (3) introduzir no texto algumas etiquetas necessárias. As três tarefas podem ser feitas ao mesmo tempo, ao rever o texto digitalizado no seu ecrã acompanhado do texto impresso ao seu lado.*

I. O primeiro ponto a ter em conta relativamente à revisão do ROC é certificar-se que consegue ver nitidamente o texto no ecrã do computador. Também é importante assegurar-se que o processador de texto não altera automaticamente a formatação do texto.

1. Abra o ficheiro em Word.
2. Deve *Maximizar* o documento de modo a que ocupe todo o ecrã.
3. Vá ao menu *Ver* e seleccione *Normal*.
4. Seleccione tudo (Ctrl+T), mude o *Tipo de Letra* para Courier New ou OCR, e mude o *Tamanho de Letra* para 14.
5. No menu *Ficheiro,* escolha *Configurar Página*. Na opção *Tamanho do Papel,* escolha *Orientação Vertical* (aplique a *Todo o Documento*).
6. No menu *Ficheiro,* escolha *Configurar Página*. Na opção *Margem,* defina 1 cm para as margens *Esquerda e Direita* (aplique a *Todo o Documento*).
7. No menu *Formatação, escolha Formatação Automática.* Certifique-se que as opções "substituir aspas direitas por aspas curvas" e "substituir caracteres de símbolo (-) por (–)" **não** se encontram seleccionadas**.**

II. Agora pode começar a corrigir os erros do ROC, retirar as partes do texto que não são precisas e acrescentar algumas etiquetas necessárias para o COMPARA.

1. Sente-se numa posição confortável e tenha um exemplar do texto impresso junto de si, para que o possa consultar a todo o momento.
2. Verifique exactamente que páginas do texto deverão ser incluídas no COMPARA e apague as restantes. Por exemplo, se o extracto que é preciso vai da página 3 à página 17, podem aparecer no ecrã as páginas 2 e 18, que não são precisas. Deve simplesmente removê-las.
3. Leia o texto no ecrã e, ao fazê-lo, siga as orientações abaixo.

Elementos extra-linguísticos

Apague todas as imagens, diagramas, numeração de capítulos, numeração de páginas, etc. Retire qualquer formatação extra-linguística (ex.: em certos livros, as primeiras palavras ou linha de cada capítulo estão em letra maiúscula, devendo estas serem reescritas em letras minúsculas de forma a que fiquem igual ao resto do capítulo).

Quebras de linha

Certifique-se que os parágrafos do texto se encontram separados por mudança de linha. Se não estiverem, separe-os carregando na tecla *Enter*. Retire todas as outras quebras de linha existentes no texto. O símbolo ¶ que se encontra na barra de ferramentas permite que visualize as quebras de linha.

Na poesia e por vezes noutros tipos de texto, poderá encontrar mudanças de linha sem que haja qualquer tipo de pontuação que indique o fim de uma frase ou parágrafo. Por exemplo:

```
      IDENTIFICADA A VÍTIMA

      trata-se do ex-major do Exército Luís Dantas Castro que em Dezembro
passado se tinha evadido do Forte da Graça, em Elvas, onde aguardava
julgamento por participação num abortado golpe militar

      e isto não é mais que a patada do mau defunto.
```

Nestes casos, apague a marca de parágrafo, ou seja, a quebra de linha, e introduza
 no local onde o autor muda de linha, por exemplo:

```
      IDENTIFICADA A VÍTIMA <br> trata-se do ex-major do Exército Luís
Dantas Castro que em Dezembro passado se tinha evadido do Forte da Graça,
em Elvas, onde aguardava julgamento por participação num abortado golpe
militar <br> e isto não é mais que a patada do mau defunto.
```

Correcção do ROC

Ao comparar o texto que aparece no seu ecrã com o texto impresso ao seu lado, verá que o reconhecimento óptico de caracteres nem sempre corresponde ao que consta na edição impressa. Deve corrigir os problemas de ROC com processador de textos de modo a que o texto do ecrã fique igual ao texto impresso. Alguns problemas típicos do ROC são:

mil-r6is	em vez de	mil-réis
Nicdau	em vez de	Nicolau
tomava	em vez de	tornava
1 (number 1)	em vez de	I (I maiúsculo) – e vice-versa
0 (zero)	em vez de	O (O maiúsculo) – e vice-versa

Se o erro for recorrente, pode utilizar a função *Substituir* do Word (Ctrl L) para tornar o processo mais rápido. Não utilize o comando *Substituir tudo,* pois poderá inadvertidamente fazer alterações em partes do texto que não deve mudar.

Ortografia

Se estiver a trabalhar numa edição antiga, actualize a ortografia de forma a que esta fique de acordo com as normas ortográficas em vigor.

Por exemplo: altere "êle" para "ele"
"fácilmente" para "facilmente"

A ortografia deve apenas ser actualizada. Não altere as convenções ortográficas das diferentes variantes do português ou do inglês. Repare que, em português do Brasil, o trema (ü) é usado em palavras como *seqüestro*, *bilíngüe,* etc., palavras como *idéia* e *vôo* são também acentuadas, mas não é obrigatório o uso de acento agudo em formas do pretérito perfeito (*amamos* em vez de *amámos* está correcto).

Erros tipográficos

Algumas edições impressas podem conter erros tipográficos, que deverá corrigir. Só corrija os erros de tipografia óbvios (os eventuais erros de ortografia e gramática que pareçam ser da responsabilidade do autor ou do tradutor devem ser preservados tal como estão).

Se o texto com o qual está a trabalhar contiver erros tipográficos, crie um documento texto separado com um registo das suas correcções. Este ficheiro deve ter a seguinte forma:

```
125, 20: impedi-a: impedia-a
125, 26: engadonhos: enfadonhos
129, 14: proclamda: proclamada
137, 8: espraiva-: espraiava-
137, 31: com se de repente: como se de repente
144, 23: trabalho, com: trabalho, como
```

O número da primeira coluna corresponde ao número da página da edição impressa onde encontrou o erro. O número da segunda coluna corresponde ao número da linha onde está ou onde começa o erro. O texto da terceira coluna corresponde ao erro. Na quarta coluna consta a correcção. O registo do erro (e da sua correcção) nas colunas 3 e 4 pode incluir palavras adjacentes ao erro em si, para facilitar a identificação do mesmo, como no caso dos dois últimos exemplos acima. Guarde esse ficheiro em formato texto, conforme as instruções no ponto 5 deste documento.

Lembre-se de que só os erros tipográficos devem ficar registados neste ficheiro separado. Os problemas de ROC e a actualização ortográfica não devem ser registados.

Títulos de capítulos

Os títulos e subtítulos de capítulos devem ser etiquetados com <chaptitle> antes do título e </chaptitle> depois. Se estiverem em letras maiúsculas, devem ser passados para minúsculas (apenas a primeira letra da primeira palavra e a primeira letra dos nomes próprios devem continuar em letra maiúscula). Por exemplo:

Mude
```
LOOKING-GLASS INSECTS
```
Para
```
<chaptitle> Looking-glass insects </chaptitle>
```

Hífens, traços, travessões e *bullets*

Mude os travessões e os traços para hífens duplos (--). Os hífens e os sinais tipo *bullets* devem ser representados por hífens curtos (-).

Aspas e apóstrofos

Altere todas as aspas duplas (“) e (”) para os caracteres («) e (»). Altere todas as aspas simples (‘) e (’) para acento grave (`) e agudo (´). Substitua todos os apóstrofos curvos (’) por apóstrofos rectos (').

Notas do Tradutor

Quando existirem notas do tradutor, retire o sinal que identifica a nota (normalmente um número superior à linha ou um asterisco) e insira o texto integral da nota (que está, normalmente, no fim da página ou no fim do capítulo) precisamente no ponto onde se encontrava a marca. Insira <tnote> onde a nota começa e </tnote> onde a nota acaba, por exemplo:

```
ele revelou-me o seu interesse por Gosse <tnote> Edmund William Gosse
(1849-1928), crítico inglês </tnote> e pela sociedade literária inglesa dos
finais do século passado.
```

Notas de Autor

Se existirem notas de autor, altere o sinal que identifica a nota (normalmente um número superior à linha ou um asterisco) para <marca num=1> na primeira nota, <marca num=2> na segunda nota, e assim sucessivamente. De seguida, insira o texto integral da nota (que está, normalmente, no fim da página ou no fim do capítulo) imediatamente a seguir à frase que remete à nota, inserindo as etiquetas **<anote >** no início e **</anote>** no fim. Assim, o texto

```
Filomena. Ou Mena. Filomena Joana Vanilo* Athaide (segundo os arquivos
daquele Depósito Disciplinar) de 23 anos, solteira, que por autorização
superior visitou o major Dantas Castro nas datas tais e tais e nas
condições de vigilância determinadas pelo Regulamento, Elvas, Forte da
Graça, tantos de tal. Otero: A que propósito é que uma merda destas vem em
ofício confidencial?

* Van Niel, e não Vanilo. A mãe de Mena, já falecida, era filha de
comerciantes sulafricanos (correcção, a lápis, do inspector Otero).
```

deve ser etiquetado da seguinte forma:

```
Filomena. Ou Mena. Filomena Joana Vanilo <marca num=1> Athaide (segundo os
arquivos daquele Depósito Disciplinar) de 23 anos, solteira, que por
autorização superior visitou o major Dantas Castro nas datas tais e tais e
nas condições de vigilância determinadas pelo Regulamento, Elvas, Forte da
Graça, tantos de tal. <anote> Van Niel, e não Vanilo. A mãe de Mena, já
falecida, era filha de comerciantes sulafricanos (correcção, a lápis, do
inspector Otero). </anote> Otero: A que propósito é que uma merda destas
vem em ofício confidencial?
```

Deve preservar a pontuação exacta das notas, e lembre-se que podem existir notas do tradutor dentro das notas do autor.

Texto tipograficamente saliente

Os corpora armazenam textos em formato txt, perdendo-se toda a formatação tipográfica das edições impressas (negrito, itálico, sublinhado, tamanho ou tipo de letra diferente, tamanho de margem diferente) que não seja propositadamente assinalada. No COMPARA, nós não preservamos o aspecto físico da edição impressa, mas guardamos a informação semântica subjacente com etiquetas para identificar títulos, entidades mencionadas, palavras estrangeiras, ênfase e voz. Isto deve ser feito de acordo com os critérios delineados abaixo. Note-se que só devem ser etiquetados os títulos, as palavras estrangeiras etc. que se encontram tipograficamente salientes na edição impressa. Não insira nenhuma etiqueta se não houver formatação que a saliente.

Títulos

Insira <title> no início e </title> no fim de títulos de livros, jornais, revistas, filmes, peças de teatro, programas de televisão, canções, notícias, etc., tanto reais como fictícios. Por exemplo:

```
We'd been to an early-evening showing of <title>Reservoir Dogs</title> .
```

Tenha em atenção que esta etiqueta identifica apenas os títulos *citados* nos textos e não os títulos ou sub-títulos dos próprios textos, que, conforme descrito anteriormente, devem ser etiquetados com <chaptitle>.

Entidades mencionadas

Insira <named> no início e </named> no fim de **nomes próprios** utilizados para identificar lojas, produtos, empresas, marcas, pessoas, doutrinas, etc. Por exemplo:

```
He stayed at the <named>Hotel Paris</named>

me puseram a alcunha de <named> Bolinha </named> quando estava na tropa

passou-me uma receita de <named> Valium </named>.
```

Palavras e expressões estrangeiras

Insira <foreign> no início e </foreign> no fim de palavras que estejam numa língua diferente da língua principal do texto. Por exemplo:

```
But the white bear, <foreign> thalassarctos maritimus </foreign>, is the
aristocrat of bears...
```

A etiqueta <foreign> *não* deve ser utilizada para nomes próprios como *Macbeth*, mas deve ser utilizada para nomes próprios que são constituídos por ou incluem nomes comuns, tal como *Bouvard et Pécuchet,* que é considerado *foreign* porque a conjunção francesa *et* pode dar origem às traduções *Bouvard and Pécuchet* (en) e *Bouvard e Pécuchet* (pt). Da mesma forma, num texto português, *Benson and Hedges* é considerado *foreign* porque a conjunção inglesa *and* pode gerar a tradução portuguesa *Benson e Hedges.* Note-se, porém, que um nome próprio como *Luís de Camões*, que contém a preposição portuguesa *de*, não pode ser marcado <foreign> porque a sua tradução inglesa jamais seria **Luís of Camões.*

Voz

Insira a etiqueta <voice> no início e </voice> no fim de citações e de mudanças de voz na narrativa, ou seja, mudanças que indicam que determinada personagem está a pensar, ler, escrever uma carta, recordar ou que a voz de outra personagem se intromete no texto. Por exemplo:

```
The fox stopped and turned his head to look at Vic for a moment, as if to
say, <voice>Yes?</voice> and then proceeded calmly on his way, his brush
swaying in the air behind him.

«To understand a message is to decode it. Language is a code. <voice>But
every decoding is another encoding. </voice> If you say something to me…»

The station is plastered with notices saying that platforms will be closed
one minute before the advertised departure times of trains «<voice> in the
interests of punctuality and customer safety </voice>», but he could have
let me through without endangering either.

Eu disse que queria o carro. <voice> Eu tinha de ter o carro. </voice>  O
vendedor disse que podia conseguir um outro em duas ou três semanas,
```

Ênfase

Insira <emph> no início e </emph> no fim de palavras ou expressões dentro de uma frase que tenham sido salientadas por razões de ênfase. Por exemplo:

```
intimate, bitter and incessant <emph> boredom </emph> which prevents me

fui ver ao dicionário <emph> parafernália </emph>

acaba por se esquecer de ter medo, até que acaba por verificar que não há
<emph> de que </emph> ter medo.
```

Por envolver uma certa dose de subjectividade, a etiqueta <emph> só pode ser utilizada se o texto tipograficamente saliente não for <title>, <named> ou <foreign>. Caso considere que uma palavra estrangeira possa estar em itálico não só por ser estrangeira mas também por estar enfatizada, coloque apenas a etiqueta <foreign>. Por exemplo:

```
«<foreign>Au contraire</foreign>, as Amy would say…»
```

Da mesma forma, apesar de a etiqueta <emph> poder estar *dentro* de um segmento etiquetado com <voice>, <voice> e <emph> não podem sobrepor-se totalmente. Assim, mesmo que ache que o exemplo abaixo indicado é tanto <emph> como <voice>, deve apenas colocar a etiqueta <voice> (ver nota sobre co-ocorrência de etiquetas).

```
Eu disse que queria o carro. <voice> Eu tinha de ter o carro. </voice>  0
vendedor disse que podia conseguir um outro em duas ou três semanas,
```

Notas importantes acerca das etiquetas de texto tipograficamente saliente

Só devem ser utilizadas as etiquetas acima em torno de títulos, palavras estrangeiras etc. que se encontrem tipograficamente salientes na edição impressa. Não insira nenhuma etiqueta se

não houver formatação de saliência (itálico, negrito, maiúsculas, tipo ou tamanho de letra diferente ou tamanho de margem diferente). Compare:

a.“quando vou de *jeans* e blusão de cabedal”
b. “vestido com camiseta e jeans…”

A palavra *jeans* deve ser etiquetada <foreign> no exemplo (a), mas a mesma palavra deve ficar sem etiquetas no exemplo (b).

Aspas
As aspas não são consideradas marcadores de texto saliente. Por isso, não ponha etiquetas nos títulos, palavras estrangeiras, ênfase, mudança de voz e entidades mencionadas que estiverem entre aspas. Por exemplo, no excerto abaixo a palavra *boas* não deve ser etiquetada:

```
Na coluna de coisas «boas», escrevi:
```

Aspas e sinais de pontuação adjacentes ao texto com tipografia saliente
Quando houver aspas ou sinais de pontuação adjacentes ao texto tipograficamente saliente, deixe-os do lado de fora das etiquetas. Por exemplo:

```
With the greatest respect, complete <foreign> cojones </foreign>.
```
(ponto final por fora)

```
«How's your <foreign>Angst </foreign>?»
```
(ponto de interrogação e aspas por fora)

```
«<foreign>Louche </foreign>» foi o veredicto de Amy
```
(aspas por fora)

Maiúsculas
Após etiquetar as palavras tipograficamente salientes assinaladas com maiúsculas (em vez de itálico, negrito, letra diferente, etc.), mude as letras maiúsculas para minúsculas. Caso se trate de nomes próprios (títulos e entidades mencionadas), deixe a primeira letra de cada palavra em maiúsculas. Por exemplo, em:

```
Parked near the hospital was a large white Peugeot hatchback: it was
painted with blue stars, a telephone number and the words AMBULANCE
FLAUBERT
```

Marque:

```
Parked near the hospital was a large white Peugeot hatchback: it was
painted with blue stars, a telephone number and the words <named><foreign>
Ambulance Flaubert </foreign> </named>
```

Note que as maiúsculas usadas nos acrónimos (por exemplo, UNESCO, FAO, ONU, etc.) não devem ser consideradas como "tipograficamente salientes". Assim sendo, só devem ser etiquetados os acrónimos que estiverem em negrito, itálico, tamanho de letra diferente, etc. Etiquetados ou não, os acrónimos nunca devem ser passados para letras minúsculas.

Listas
Se encontrar uma lista de títulos, palavras estrangeiras, entidades mencionadas, etc., insira etiquetas separadas para cada elemento da lista. Por exemplo, marque:

```
<foreign> Urutus </foreign>, <foreign> jararacas </foreign>, <foreign>
cascavéis </foreign>, <foreign> jararacuçus </foreign>, <foreign>
surucutingas </foreign>, <foreign> cotiaras </foreign> -- I saw these and
many other serpents in the slides that Melissa projected during her talk.
```

Co-ocorrência de etiquetas
É possível existirem etiquetas <title>, <named>, <foreign> e <emph> dentro de segmentos etiquetados com <voice>. Por exemplo:

```
<voice>It would be good if she came <foreign>tout de suite</foreign>.
</voice>

Ocredito ( faz-me lembrar aquela canção  <voice> « <emph> Ocredito </emph>
que por cada gota de chuva que cai nasce uma nova flor » </voice> )
```

<foreign> pode co-ocorrer com <title> e <named>. Por exemplo:

```
uma espécie de versão inglesa do <title><foreign>Twin
Peaks</foreign></title>

A pair of them come at once to mind: <title><foreign> Bouvard et Pécuchet
</foreign></title>, where Flaubert sought to enclose...

sugeriam que tinha pedido um favor em troca do <named><foreign> Benson and
Hedges </foreign></named>

como uma senhora da <foreign><named> Belle Époque </named></foreign>
ajustaria seu vestido
```

Finalmente, a etiqueta <emph> pode ser inserida dentro de outras etiquetas, mas não pode haver concordância total do seu âmbito com essas (recorde os critérios para etiquetar palavras com ênfase, explicitados mais acima).

Ordem das etiquetas
A ordem das etiquetas não é importante desde que a primeira a ser aberta seja a última a ser fechada, e que a segunda a ser aberta seja a penúltima a ser fechada, e assim sucessivamente. Por exemplo, ambas as alternativas <title><foreign>…</foreign></title> e <foreign><title>…</title></foreign> estão correctas. Mas a marcação <title><foreign>…</title></foreign> está incorrecta.

III. Para descansar
Guarde o documento várias vezes enquanto estiver a trabalhar. Se tiver de parar para descansar, assinale onde parou, tanto no livro (página x) como no ecrã. Para assinalar, no texto digital, o ponto onde parou para descansar, pode inserir três asteriscos (***). Quando voltar a abrir o documento, deve ir ao menu *Editar* e depois seleccionar *Localizar* ***.

IV. Revisão final
Correcção ortográfica

Quando acabar a preparação do texto, seleccione o texto todo (Ctrl+T), vá ao menu *Ferramentas*, depois seleccione *Idioma* e *Definir Idioma*. Escolha a língua do texto, ou seja, Inglês (Reino Unido), ou Inglês (E.U.A.), ou Português (Brasil), ou Português (Portugal), etc. Depois vá ao menu *Ferramentas* e escolha *Ortografia e Gramática.* Desactive a correcção gramatical, e utilize a correcção ortográfica **apenas** para detectar problemas de ROC que escaparam a um primeiro escrutínio. Ignore todas as sugestões do corrector que não tenham a ver com ROC. Por outras palavras, o corrector ortográfico tem de ser usado com muito cuidado, e só para localizar problemas de ROC que escaparam.

Leitura final
O corrector ortográfico não detecta erros como *tomava* em vez de *tornava*, ou *ai* em vez de *aí*. Por isso é preciso imprimir o texto e fazer uma leitura final, corrigindo o que faltar.

V. Guardar
Depois de acabar a revisão do texto todo, use o código alfanumérico do COMPARA para dar nome ao ficheiro (pergunte-me se não souber, ou veja http://www.linguateca.pt/COMPARA/Codigo.html) seguido de *ocr*, e guarde-o em formato texto com extensão *pt* para os textos em língua portuguesa e extensão *en* para os textos em língua inglesa. Por exemplo:

PBMA3ocr.pt (para um texto em versão portuguesa)
PBMA3ocr.en (para o mesmo texto em versão inglesa)

Se o texto contiver erros tipográficos, o ficheiro com os registos das correcções feitas deve ser guardado em formato txt com o código alfanumérico da obra, seguido de *erros,* seguido da extensão *pt* ou *en*, conforme o caso. Por exemplo:

PBMA3erros.pt

Envie os ficheiros para Ana Frankenberg

Agradecimento
O COMPARA é financiado pelo programa POSC através da Linguateca projecto POSC 339/1.3/C/NAC, executado pela FCCN.