COMPARA

Optical Character Recognition (OCR) editing and preliminary markup instructions

Ana Frankenberg-Garcia (27/06/2007)

For a text to become searchable in an electronic corpus, it has to be in digital form. If the text is not available in digital form, it has to be either typed into a computer or scanned and submitted to an OCR program. The scanner works like a photocopier, i.e., it takes a photograph of the text. An OCR program transforms the text that has been photographed by the scanner into text that you can work with using a word processor. After a text has been through an OCR program, you can add or delete words, change fonts, etc. The first stage of preparing texts for COMPARA described here involves (1) correcting OCR problems, (2) removing parts of the text that are not needed, and (3) introducing a number of tags that are necessary. All three things can be done at the same time, as you go over the digital text on your computer screen with the printed text near at hand.

I. The first thing about OCR editing is to make sure you can see the text very clearly on your computer screen. It is also important to ensure that your word processor doesn't do any automatic changes to the formatting of the text.

1. Open file in Word
2. *Maximize* your document so that it takes up the whole screen.
3. Go to *view* and select *normal*.
4. Select all (Ctrl A), change *font* to Courier New or OCR, and change *font size* to 14.
5. Go to *file* and *page setup*. Set *paper size* to *portrait* (apply to *whole document*).
6. Go to *file* and *page setup*. Set *left* and *right margins* to 1 cm (apply to *whole document*).
7. Go to *format* and *autoformat.* Select *options.* Make sure the boxes saying "replace straight quotes with smart quotes" and the one saying "replace hyphens with dashes" are **unchecked.**

II. Now you're ready to start cleaning OCR mistakes, removing unwanted parts of the text and adding some corpus annotation marks.

1. Sit in a comfortable position and have the hard copy of the text next to you and ready for constant reference.
2. Check exactly which pages of the text are to be included in COMPARA and remove the rest. For example, if the extract needed is from page 3 to 17, you may also get on your screen pages 2 and 18, which are not needed. Simply delete them.
3. Read through the text on your screen, and follow the instructions below as you do so:

Extra-linguistic material

Delete all pictures, diagrams, chapter numbers, page numbers, etc. Remove any extra-linguistic formatting (e.g., in some books, the first word or line of every chapter is in capital letters – this should be rewritten in small letters, like the rest of the chapter).

Line breaks

Make sure every paragraph in the text begins with a carriage-return line break. If they don't, insert one by pressing *enter*. Remove all other line breaks in the text. The ¶ symbol on your tool bar helps you see where the line breaks are.

In poetry and sometimes in other types of text, there may be new lines without there being any punctuation to suggest end of sentence or paragraph. For example:

```
      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.
```

In cases such as these, delete the carriage returns (i.e., line breaks) and insert <br> at the point where the author changes line, e.g.:

```
      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.
```

OCR correction

When you compare the OCR output on your screen with the printed text next to you, you will notice that the OCR output does not always give you what you see in the printed text. Correct the OCR problems with your word processor so that the text on your screen matches the printed edition. Some typical OCR problems are:

mil-r6is	instead of	mil-réis
Nicdau	instead of	Nicolau
tomava	instead of	tornava
1 (number 1)	instead of	I (capital I) – and vice-versa
0 (zero)	instead of	O (capital O) – and vice-versa

If the error is recurrent, you can use the *replace* function of Word (Ctrl H) to speed things up. Do not use *replace all*, as this can inadvertently replace other parts of the text.

Spelling

If you are working with an old edition, update the spelling so as to make it conform to current orthographic norms.

e.g change "êle" to "ele"
"fácilmente" to"facilmente"

Spelling should only be updated. Do not change the spelling of different varieties of Portuguese or English. Note that in Brazilian Portuguese the *umlaut* (ü) is still used in words like *seqüestro*, *bilíngüe,* etc. Words such as *idéia* e *vôo* also take accents, but words in the pretérito perfeito do not need them (e.g., *amamos* instead of *amámos*).

Misprints

Some printed editions may contain misprints, which you are to correct. Only obvious misprints should be corrected (spelling and grammar mistakes that appear to have been introduced by authors or translators should be left alone).

If the text you are working on contains misprints, create a separate text document to keep a record of your corrections. This document should take the following form:

```
111, 10: Tinhas as faces: Tinha as faces
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
145, 31: fragmentou-e: fragmentou-se
```

The number on the first column is the page number on the printed edition where the misprint was found. The number on the second column is the line number where the misprint is or begins. The text on the third column is the misprint. The text on the fourth column is your correction. Save this ducument in text format, following the instructions on point 5 of the present document.

Remember that only misprint corrections need be recorded in this separate file. OCR problems and spelling updates should not be recorded.

Chapter titles

Insert the tag <chaptitle> before chapter titles and subtitles and the tag </chaptitle> after them. If the titles are capitalized, rewrite them such that only the first letter of the title or of any proper names within it remains capitalized, for example:

Change

```
LOOKING-GLASS INSECTS
```

To

```
<chaptitle> Looking-glass insects </chaptitle>
```

Hyphens, dashes, *travessões* and bullets

Rewrite travessões and dashes as double hyphens (--), and make sure hyphens and bullets represented by the single mark (-) .

Quotation marks and apostrophes

Change all double quotes into («) to open and (») to close. Change all single quotes to grave accent (`) to open and acute accent (´) to close. Rewrite all apostrophes as single, non-directional quotation marks (').

Translators' notes

Whenever there are translators' notes, remove the mark that identifies the note (usually a superscript number or an asterisk) and insert the full text of the note (which is usually at the bottom of the page or at the end of the chapter) exactly the point of its mark. Insert **<tnote >** where the note begins and **</tnote>** where the note ends, for example:

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.

Authors' notes

If there are any authors' notes, change the mark that identifies the note (usually a superscript number or an asterisk) into <marca num=1> for the first note, <marca num=2> for the second note, and so on. Then insert the full text of the note (which is usually at the bottom of the page or at the end of the chapter) immediately after the sentence in which the note appears and surround it with **<anote >** and **</anote>** tags. Thus the text

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).

Should be marked as follows:

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?

Remember to preserve the exact punctuation of the notes and notice that there may be translators' notes inside authors' notes.

Highlighted text

Corpus texts are stored in plain text format, so all highlighting (usually italics, but also bold, underlining, indentation, capital letters or different font) that is not purposefully preserved will be lost. In COMPARA, we do not preserve the physical appearance of highlighted text in printed editions, but we do preserve the semantic information underlying it with tags for titles, named entities, foreign words, emphasis and voice. These should follow the criteria outlined below. Remember to do this only for titles, foreign words and so on that have been set off by special formatting in the printed edition. When there is no special formatting do not insert any tags.

Titles

Insert <title> at the beginning and </title> at the end of both real and fictional titles of books, newspapers, magazines, films, plays, television programmes, songs, etc. For example:

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

Note that this mark only identifies titles *cited* in the corpus texts, and not titles or sub-titles of the texts themselves, which, as already mentioned, should be marked <chaptitle>.

Named entities

Insert <named> at the beginning and </named> at the end of **proper names** used for shops, products, companies, brand marks, people, doctrines, etc. For example:

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>.

Foreign words and expressions

Insert <foreign> at the beginning </foreign> at the end of words in a language other than the main language of the text. For example:

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

The <foreign> tag should *not* be used for proper names like *Macbeth*, but should be used for proper names that are made of or include common nouns, like *Bouvard et Pécuchet*, which is considered foreign because the French conjunction *et* can give rise to the translation *Bouvard and Pécuchet* (En) and *Bouvard e Pécuchet* (Pt) . Likewise, in a Portuguese text, *Benson and Hedges* is considered foreign because the English conjunction *and* can give rise to the translation *Benson e Hedges* . Note, however, that a name like *Luís de Camões*, which contains the Portuguese preposition *de*, cannot be marked <foreign> because the name cannot give rise to the translation **Luís of Camões.*

Voice

Insert <voice> at the beginning and </voice> at the end of citations and changes of voice in the narrative, indicating that a character is thinking, reading or writing a letter, reminiscing, or that the voice of another character is intruding. For example:

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> 0 vendedor disse que podia conseguir um outro em duas ou três semanas,

Emphasis

Insert <emph> at the beginning </emph> at the end of words or expressions within a sentence that have been highlighted for emphasis. For example,

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.

Because this tag involves a certain amount of subjectivity, <emph> can only be used if the highlighted text is not <title>, <named> or <foreign>. So if, for instance, you think a foreign word might be in italics not just because it is foreign, but also because it is being emphasised, mark only <foreign>. For example,

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

Likewise, although there can be <emph> *within* a <voice> segment, <voice> and <emph> cannot overlap completely. So even if you feel the example below should be both <emph> and <voice>, you should mark only <voice> (see notes on overlaps).

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,

Important notes concerning highlighted text tags

Only insert the above tags if the text has been *highlighted* in the printed edition. Do not mark any titles, proper names, foreign words, emphasis and changes of voice that have not been highlighted in italics or otherwise. Compare:

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

Text (a) should include tags for foreign around the word *jeans*, but no tags should be used around the same word in text (b).

Quotation marks

Quotation marks are *not* to be considered highlighted text. Do not put any tags around titles, proper names, foreign words, emphasis and changes of voice that come in between quotation marks. In the example below, for instance, no tags should be inserted around the word *things*:

It became one my «things» -- things I can't decide, can't forget, can't leave alone

Quotation and punctuation marks adjacent to highlighted text

If there are any quotation or punctuation marks adjacent to the highlighted text, leave them outside the tags. For example:

With the greatest respect, complete <foreign> **cojones </foreign>.**

(full-stop outside)

```
«How's your <foreign>Angst </foreign>?»
```

(question mark and quotation mark outside)

```
«<foreign>Louche </foreign>» foi o veredicto de Amy
```

(double-quotes outside tags)

Capital letters

After putting tags around capitalized titles, foreign words (and so on), change the capital letters to small letters. If there are any proper nouns (e.g. titles and named entities), use "title case", i.e., leave the first letter of each word in capitals. For example, in:

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

Mark:

```
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 that the capital letters used in acronyms (e.g. UNESCO, NATO, etc.) do not indicate that the text is highlighted. Therefore, acronyms should only be tagged if in addition to being in capitals they are also in italics, bold, different font, etc. Whatever the case, acronyms should never be rewritten in small letters.

Lists

If you find a list of titles, foreign words, named entities, etc., use separate tags for each element of the list. For example:

```
<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.
```

Overlaps

There can be tags for <title>, <named>, <foreign> and <emph> inside <voice> segments. For example:

```
<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> )
```

The tag for <foreign> can overlap with <title> and <named>. For example:

```
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
```

The tag for <emph> can be inserted within other tags but cannot not overlap completely with them (see above criteria for marking emphasis)

Order of tags
The order of the tags is not important, as long as the first one to open is the last one to close, the second to open is the second last to close, and so on… For example, both <title><foreign>…</foreign></title> and <foreign><title>…</title></foreign> are correct. But <title><foreign>…</title></foreign> is incorrect.

III. Having a break

Remember to save your text every few minutes. If you have to stop for a rest, remember to mark your place on the book (page x) as well as your place on the screen. A good way of marking your place on screen is to insert a *** mark. The next time you open your document, go to *edit,* then *find* ***.

IV. When you finish

Spelling check
When you finish going over the whole text, select the whole text (Ctrl A), go to “tools”, then “language”, and “set language”. Choose the language of your text, i.e., British English, or American English, or Brazilian Portuguese, or European Portuguese, etc. After that, go back to “tools” and run a “spelling-check” to detect any OCR problems you’ve missed. Do not use the grammar-check option and use the spelling checker **only** to detect OCR problems. Disconsider all other spelling-checker suggestions. In other words, the spelling checker is to be used very carefully, and only to locate OCR problems that you've missed.

Final reading
The spelling checker cannot detect errors like *tomava* instead of *tornava*, or *ai* instead of *aí*. This means you must print the text, give it a final read and correct whatever is still needed.

V. Saving

After you've finished going through the whole text, save the files in text format with the text's letter&number code (ask me or see http://www.linguateca.pt/COMPARA/Code.html), followed by *ocr* and extension *pt* for texts in Portuguese and *en* for texts in English. For example:

PBMA3ocr.pt (for a text in Portuguese)
PBMA3ocr.en (for the same text in English)

If the text contains misprints, the file with a record of the corrections made should be saved in text format with the text's letter & number code followed by *erros,* followed by extension *pt* or *en*. For example:

PBMA3erros.pt

Send all files to Ana Frankenberg.

Acknowledgements
COMPARA is developed in the scope of the Linguateca project, jointly funded by the Portuguese Government and the European Union (FEDER and FSE) under contract ref. POSC/339/1.3/C/NAC..