Post-OCR workflows for text

ABBYY FineReader offers a variety of options for saving text after recognition is completed. If you are not saving a single selected page or all pages in a longer document into a single file, ensure that the desired range is selected before you save the text.

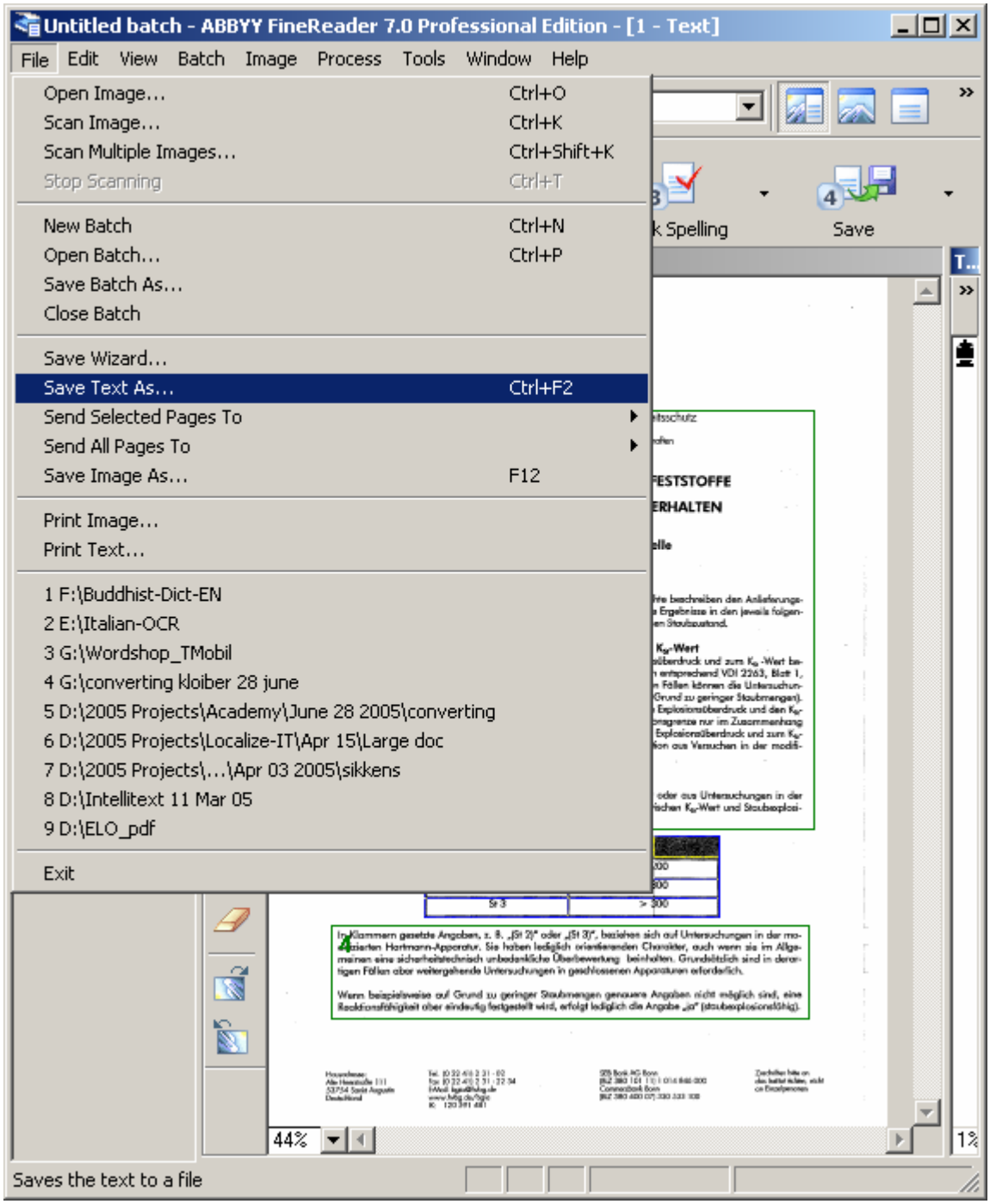

Fig 1: "Save Text As..." menu option in ABBYY FineReader

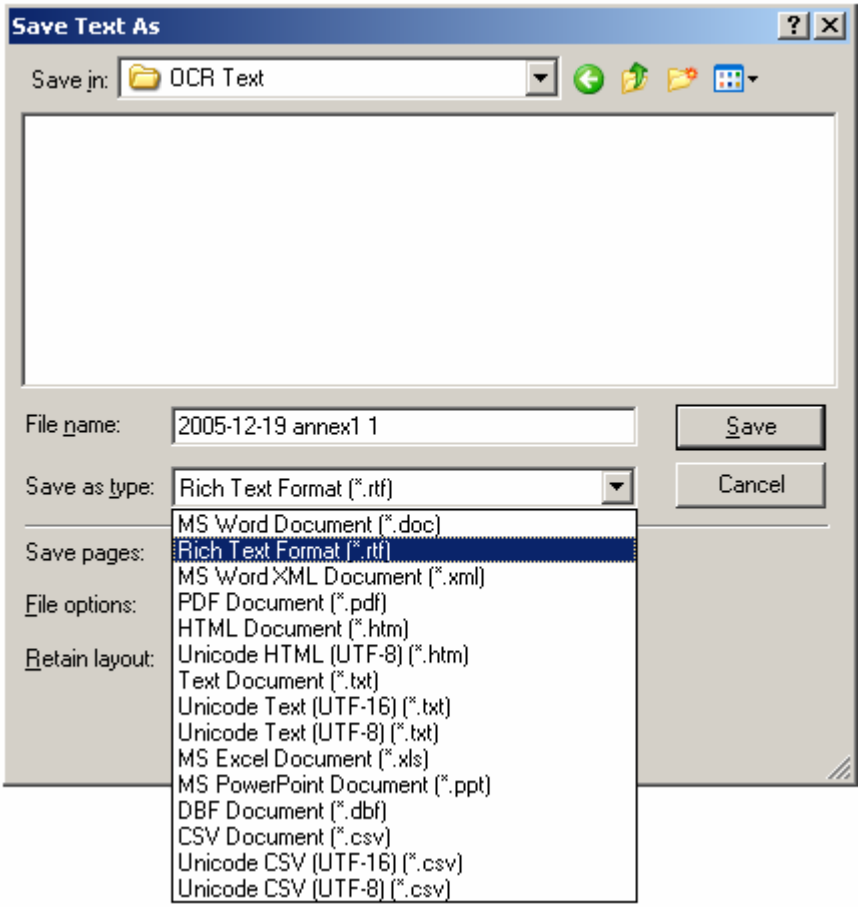

Fig 2: "Save Text As" dialog in ABBYY FineReader with file format options displayed

RTF and text formats are probably the most useful for translation work. A text format will strip all formatting and omit graphics. If you need to save layout, text formatting or graphics together with the text, then RTF is the proper choice.

When saving to RTF, ensure that the proper selection has been made under the **file options** in the Save dialog. The **retain layout** selection list has three basic options:

- saving the full layout
- saving the font, size and style information
- stripping all the font properties to save straight text

The option to save graphics is governed by the checkbox below the retain layout selection menu. The **Format Settings…** button allows you to specify useful things like retention of font color (a bad idea if white text is present!), automatic removal of optional hyphens (select this or remember to do it later in a text processing program)

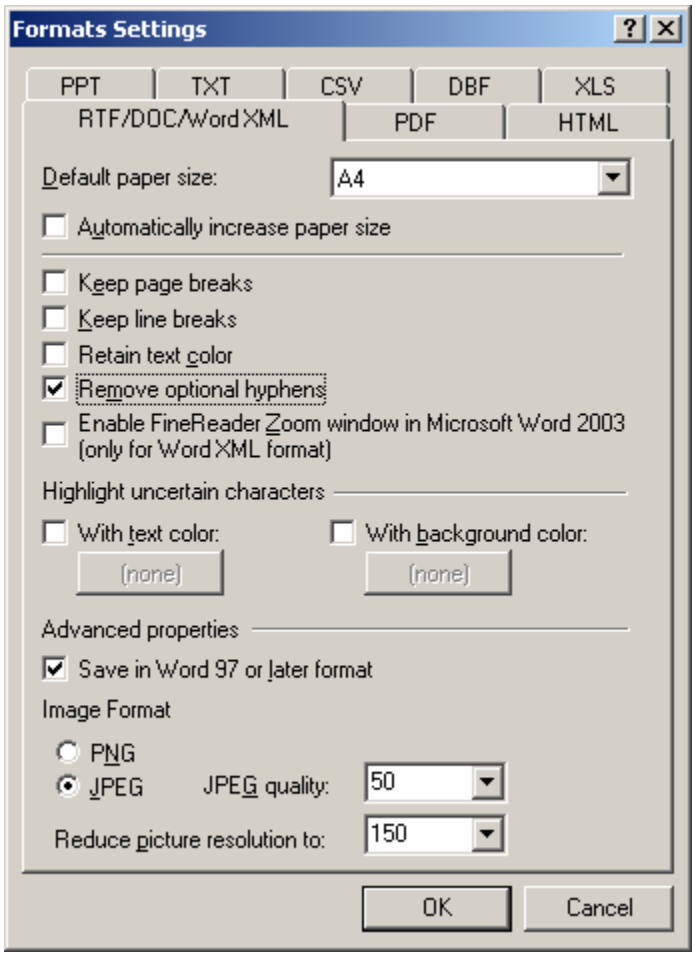

Fig 3: "Format Setting" subdialog of the "Save Text As" dialog in ABBYY FineReader with typical recommended options

Make absolutely sure that the checkbox labeled "**Automatically increase paper size**" is **not** selected in this subdialog. This is a dangerous feature that often leads to subtle, difficult to fix corruption of the OCR text documents. Select the other options appropriate to you needs.

Retaining the full layout (save text option)

This option is useful for recreating the original PDF layout in the saved RTF document. Magazine articles and brochures, for example, look very much like the original document, and with a little adjustment the OCR text can often be made to look quite good for distribution purposes. This is a popular option for press clippings, advertisements or charts and graphs with label text. Using the full layout option also makes it easier for graphic artists that must re-do an original layout to assign the elements to their correct locations. Frequent disadvantages of this layout are:

- Lots of adjustment needed to font properties to reduce tags/codes when working with translation memory tools such as Déjà Vu or Trados

- Crazy mixed layout features like text boxes and columns mixed in a multi-column layout
- Subtle layout errors that require great dexterity to fix

Retaining the font formatting (save text option)
This option is good if the results of a full format save are simply to crazy to use but characteristics like header formatting, bold and italics are useful to keep in order to read the text better or to correlate the OCR text with the original document with less difficulty. This option also often requires adjustment of font characteristics if the original documents were scanned in such a way that the OCR software interprets fractional point size differences in the type between pages or pages sections or other distortions occur.

Stripping all formatting (save text option)
This option is useful if the quality of the original document would yield a crazy mix of font properties with the other two save options or if font sizes would make on-screen reading of the text difficult. It is usually a simple matter to apply any desired font formatting after the text is saved.

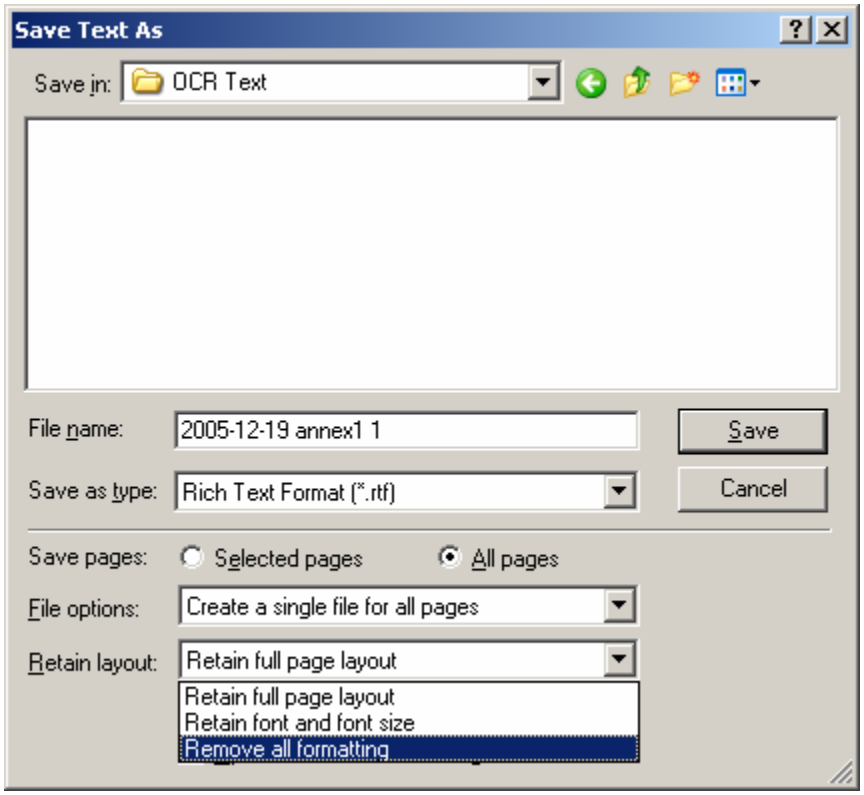

Fig 4: Saving the text without any font formatting or layout

This option still preserves table structures and any graphics (if the option to save pictures has been selected).

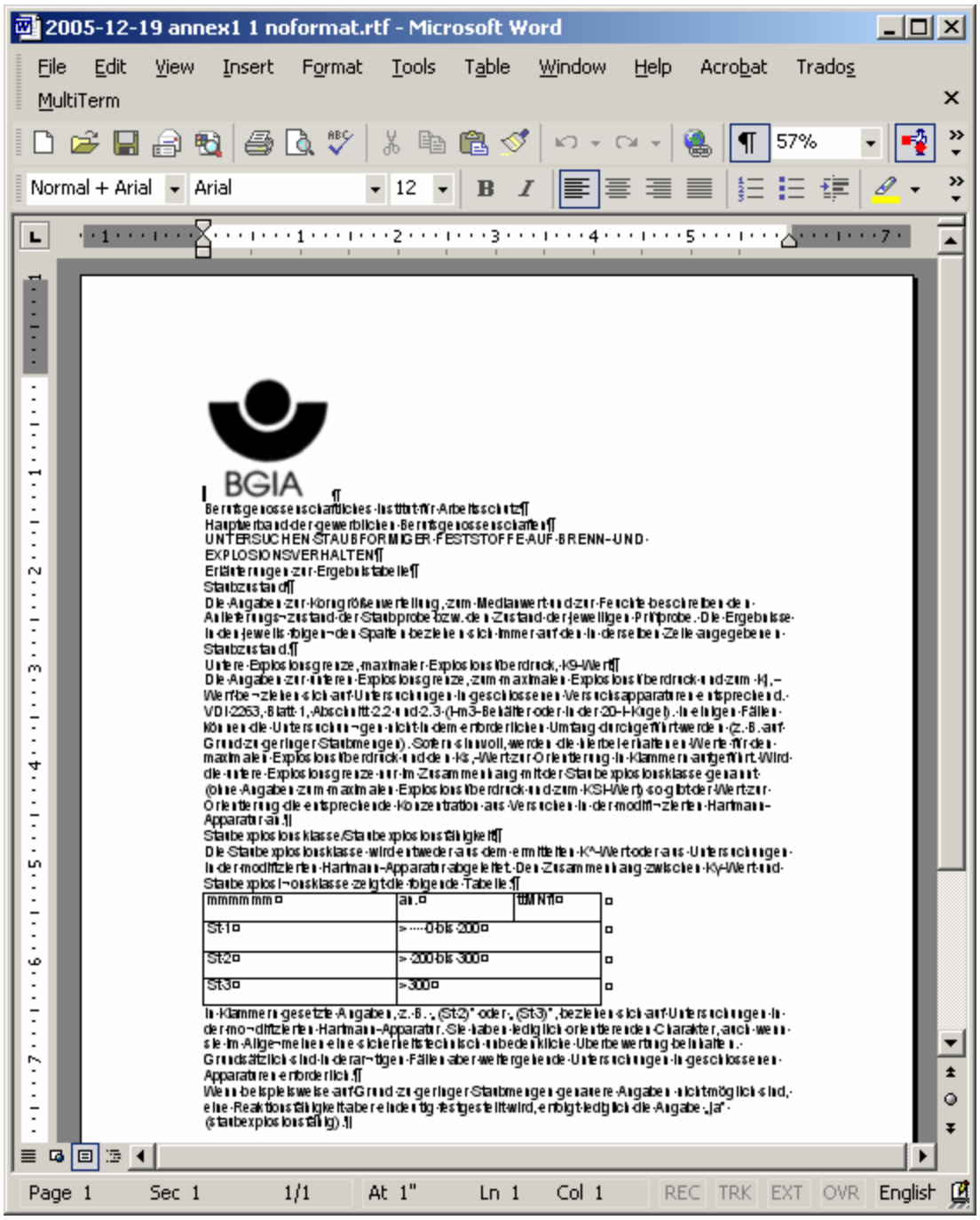

Fig 5: OCR text saved with neither layout nor formatting. Recognition errors (such as incorrect table formatting due the bad quality of the source scan) have not yet been manually corrected.

After the OCR text has been saved, use the checklist below to ensure that it is ready to be translated and, if desired, processed efficiently with translation memory tools.

Post-OCR quality assurance checklist for text to be translated

Quality characteristic	**OK**	**N/A**
Layout and text flow (order) correct and suitable for editing?		
Scale and spacing set to 100% and Normal under Format > Font > Character Spacing tab (reduces codes/tags for TM)		
All text to be translated present in the OCR (no missing words, garbage lines, etc.)		
Superfluous line breaks ands carriage returns corrected (especially for text in a sentence that flows between pages!)		
Optional hyphens and other disruptive elements removed?		
Doubled line breaks substituted by paragraph breaks (to prevent segmentation problems in Trados)		
Headers and footers placed correctly? (scan once and embed in header and footer)		
Have all formatted tables been corrected to use MS Word tables or tabs (instead of lots of spaces)?		
Have all superfluous spaces (double spaces, etc.) been removed by search and replace?		
Have any specified row heights been turned off in the table properties in MS Word? (This keeps all segments visible in Trados and avoids text wrapping trouble if the translation is “too long”)		
Entire document read through and spell-checked (If not, will corrections be made to the original in DV?)		
Source document quality suitable for delivery of the source to the customer? (If not, will the source be re-generated from corrections in the TM tool?)		

Nota bene: for search and replace in MS Word, the following are useful shortcuts

^-	optional hyphen
^p	paragraph mark
^t	tab
^l	manual line break

If too many sentences are broken up by paragraph marks (for example after saving a PDF as text from Acrobat Reader) or manual line breaks, do the following:

1. Separate all headers from body text by double paragraph marks
2. Search and replace all double paragraph marks by a "placeholder text" that will not occur in the document, such as “KKKKKK”.
3. Search and replace all the remaining paragraph marks with spaces (type a single space in the replacement text field)
4. Revert all the placeholder text to paragraph marks (or double paragraph marks as desired) using search and replace.