TRANSLATING PDF FILES – IF WE HAVE TO: TOOLS, TIPS AND TECHNIQUES FOR CONVERTING AND TRANSLATING PDF FILES

Tuomas Kostiainen (tuomas@jps.net)
www.FinnTranslations.com

---

1. Converting PDF Files to Editable Text

Text-based *vs.* Image-based PDF files in Adobe Reader/Acrobat

- **Text-based PDF files**
  - Possible to select (and copy) text word by word and letter by letter
  - Possible to search text
  - Characters look smooth and clear
  - PDF documents created from other applications are typically text-based
- **Image-based PDF files**
  - Possible to select only a rectangular area and copy it as an image (not as text)
  - Select Tool cursor is a cross-hair (instead of an arrow) and clicking the document highlights the whole page blue
  - Search function does not find any text
  - Faxed and scanned documents are typically image-based
- **Searchable image PDF files**
  - Can be created in Adobe Acrobat (and in some other PDF tools) from an image-based PDF file
  - Looks exactly like the image-based PDF file but the OCR-recognized text is "hidden" behind the image and can be searched, selected and copied as text
  - The accuracy of the recognized text depends on the clarity of the image
  - The OCR function in Adobe Acrobat 9 is not very good but it is better in versions X and XI

Adobe Reader (ver. X/XI)

- Copying and pasting using Clipboard
  - Column select mode (Alt), selecting whole page (4 clicks), selecting all (Ctrl+A; "all" can be either whole page or whole document depending on the Page Display setting: View > Page Display > Single Page vs. all other settings), copying with or without formatting (right click menu; if available)
- Saving as a text (txt) file (File > Save as Other > Text)
- Paragraph mark problem with both methods > practical only for small amount of text
- Saving as as a Word or Excel file online; a fee-based service (File > Save as Other > Word or Excel Online...)
- Text can be saved, selected and copied only from text-based PDF files and searchable image files (not from normal image-based PDF files)

Adobe Acrobat (ver. 9/X/XI)

- File > Export; File > Save As (Word, Excel, html, xml, etc.); File > Save As Other (Word, Excel, html, xml, etc.)
- Right click menu options:
    - **Acrobat ver. 9:** Copy, Copy As Table, Save As Table, Open Table in Spreadsheet
    - **Acrobat ver. X/XI:** Copy, Copy With Formatting, Export Selection As
- Tables can be tricky to convert with any of the above methods
- Conversion settings: Edit > Preferences > Convert From PDF > [select file type] > Edit Settings... [*see **Figures 1 and 2***]
- More info (Adobe online Help): http://tinyurl.com/r9jn9y (ver. 9), http://tinyurl.com/convverx (ver. X), http://tinyurl.com/convverXI (ver. XI)
- Includes an OCR function that allows converting image-based PDF files into searchable images (see above "Searchable image PDF files")
    - **Acrobat ver. 9:** Document > OCR Text Recognition
    - **Acrobat ver. X/XI:** Tools > Text Recognition
- Text can be selected and copied from text-based PDF files and searchable image files (but not from normal image-based PDF files without converting them to searchable images first)
- Image-based files can be saved directly as text-based Word/Excel files in Acrobat X/XI (File > Save As; File > Save As Other)

OCR (Optical Character Recognition) Tools for PDF Conversion

- ABBYY FineReader (www.abbyy.com)
- PDF Transformer (by ABBYY)
- OmniPage (www.nuance.com/imaging/products/omnipage.asp)
- PDF Converter (by Nuance)
- PDF to Excel: www.pdftoexcelonline.com (free online tool)
    - Not a perfect tool but free
    - Remember document confidentiality when using online tools

Using ABBYY PDF Transformer

- Simple – just a few clicks [*see **Figure 3***]
- PDF conversion and PDF creation
- Can convert both text-based and image-based PDF files
- Advanced Options for Word files: Original layout, Text flow, Keep pictures
- Advanced Options for Excel files: Ignore text outside tables, Convert numeric values to numbers
- Recommended fonts for Chinese, Japanese, Hebrew and Thai

Using ABBYY FineReader

- Full-range OCR program for scanning, PDF/image file conversion and PDF creation [*see **Figure 4***]
- Can convert both text-based and image-based PDF files; includes several features/settings that can help improve conversion results when converting image-based files
- Open an image file (PDF, TIF, etc.) or scan a document
- Read file using the OCR tool
- Image can be edited (Edit Image button)

- Several general and file-specific options for perfecting the output (Tools > Options) [*see* ***Figure 5***]
- Output can be previewed and edited on the Text Window (right side)
- Check spelling (allows you to verify words that the OCR program misread or did not recognize) and make other editing corrections on the Text Window
- Save as Word, Excel, etc. file

2. Post-editing Converted Files in Word/Excel

- Often additional formatting or "cleaning" is needed to get rid of incorrect formatting, unncessary spaces and hidden tags
- Paste Special in Word/Excel can be useful
- Reset character formatting (Ctrl+Space) in Word can also be useful (select text first)
- CodeZapper (a set of Word macros; efficient way to get rid of unnecessary "rogue" tags in Word files after conversion; for details, see http://asap-traduction.com/CodeZapper)
- TransTools Document Cleaner (http://www.translatortools.net/word-doccleaner.html)
- AutoUnbreak for deleting unnecessary line breaks (http://tinyurl.com/autounbrk)
- ASAP Utilities for Excel (http://www.asap-utilities.com)

3. Translating PDF Files Using CAT Tools

- Trados Studio, Wordfast Pro, memoQ and Fluency offer support for PDF files [*see* ***Figure 6 and 7***] but handle the conversion process differently
- Not very practical in most cases – a better solution would be to use a good PDF conversion tool to convert the file to Word format. For more info, see http://tradoshelp.wordpress.com/2010/05/17/
- If you use your CAT tool's integrated conversion function, check that the conversion settings are appropriate

4. PDF File Password Protection and Permission Control

- In Adobe Reader/Acrobat: File > Properties > Security [*see* ***Figure 8***]
- Two types of password protection: One requires a password to open the file and the other one restricts copying, editing and printing of the file. These two protection types require separate passwords and can be used independently of one another.
- If needed, there are several tools available for unprotecting PDF files (search "unprotect PDF" in the Internet and you'll find out)

5. Creating TMs from PDF Files Using Alignment Tools

- LogiTerm AlignFactory (www.terminotix.com) [*see* ***Figure 9***]
    - Also available as a FREE online version: YouAlign by LogiTerm; limited selection of languages (www.youalign.com)
- ABBYY Aligner, limited selection of languages (http://www.abbyy.com/aligner) [*see* ***Figure 10***]
    - "If you want to improve the quality of texts extracted from PDF files, use ABBYY FineReader or ABBYY PDF Transformer for conversion."
- Both tools support also many other file types

6. Additional PDF-related Links and Info

- http://acrobatusers.com/
- www.adobe.com/support/
- www.planetpdf.com
- http://desktoppub.about.com (search for “PDF”)
- Translator’s Tool Box by Jost Zetzsche (www.internationalwriters.com/toolbox): new chapter on PDF files
- ABBYY 20% discount code (“KOSTIAINEN”) at www.abbyyusa.com

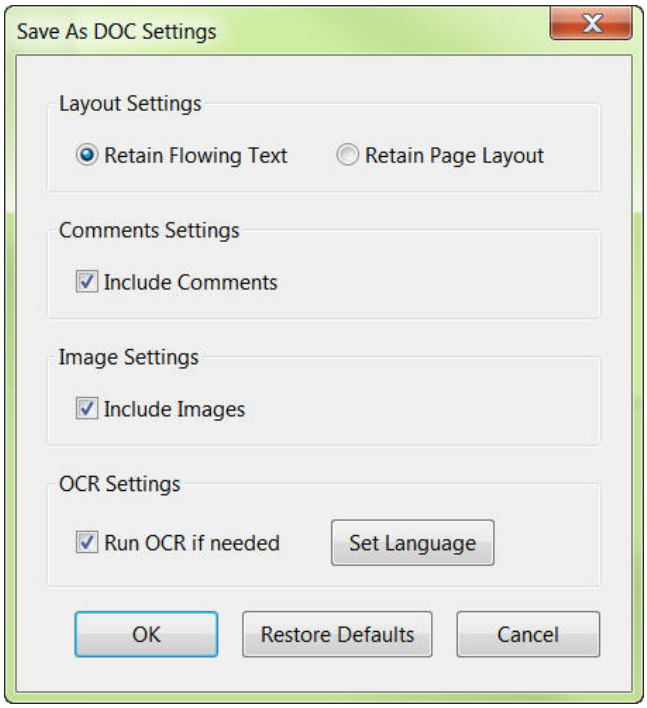

Figure 1. Adobe Acrobat X/XI: PDF to DOC conversion settings

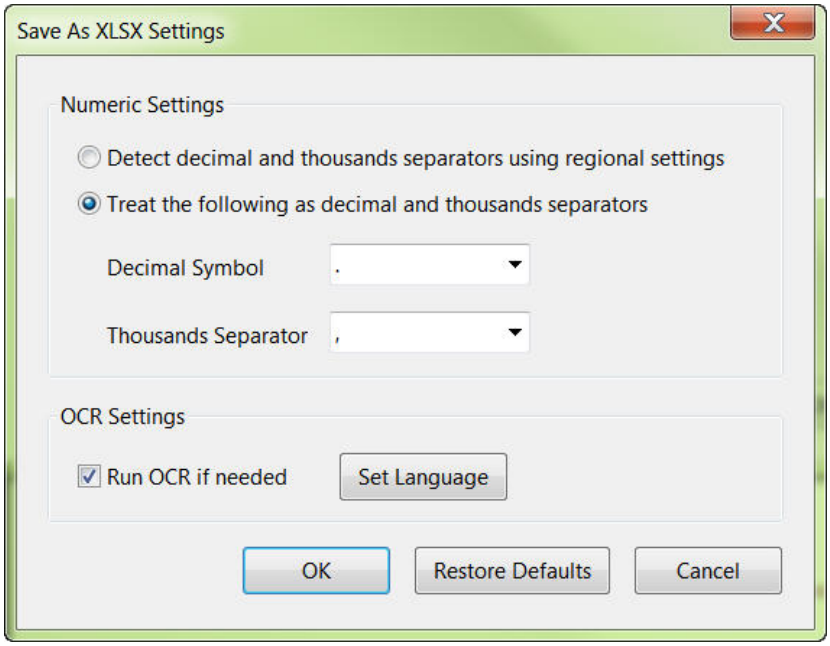

Figure 2. Adobe Acrobat X/XI: PDF to Excel (XLSX) conversion settings

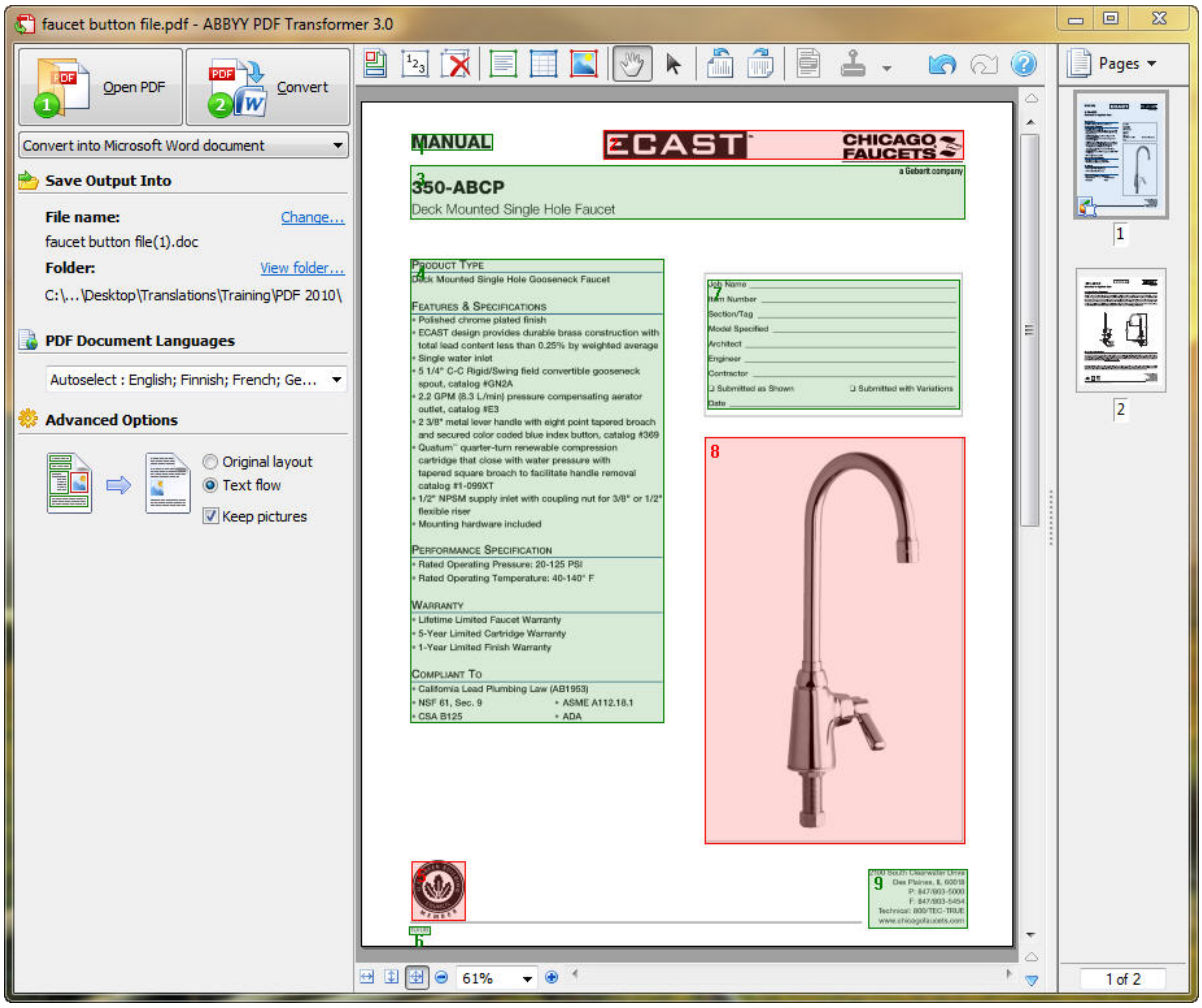

Figure 3. ABBYY PDF Transformer: Converting a PDF file to Word format.

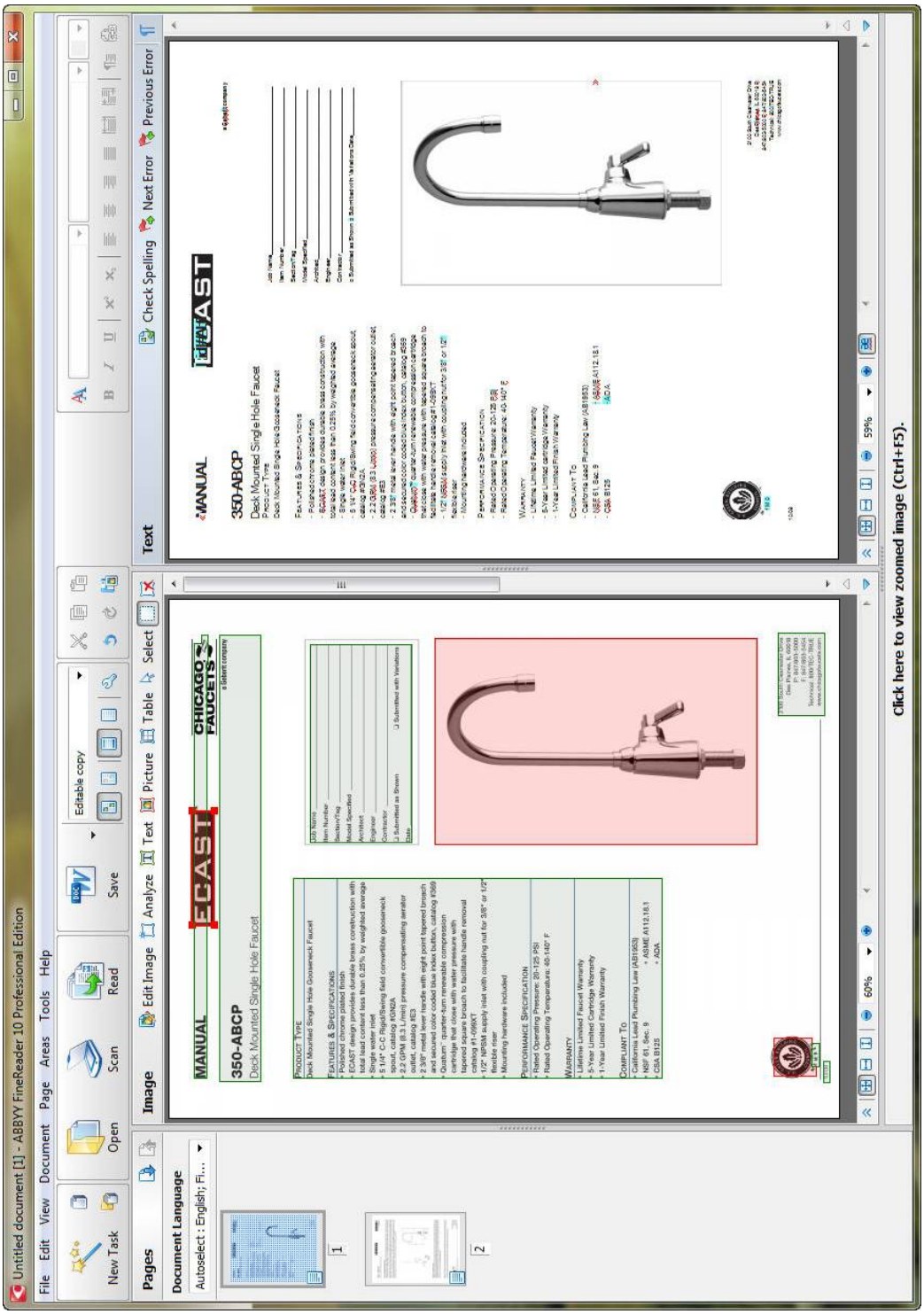

Figure 4. ABBYY FineReader: Converting a PDF file to Word format.

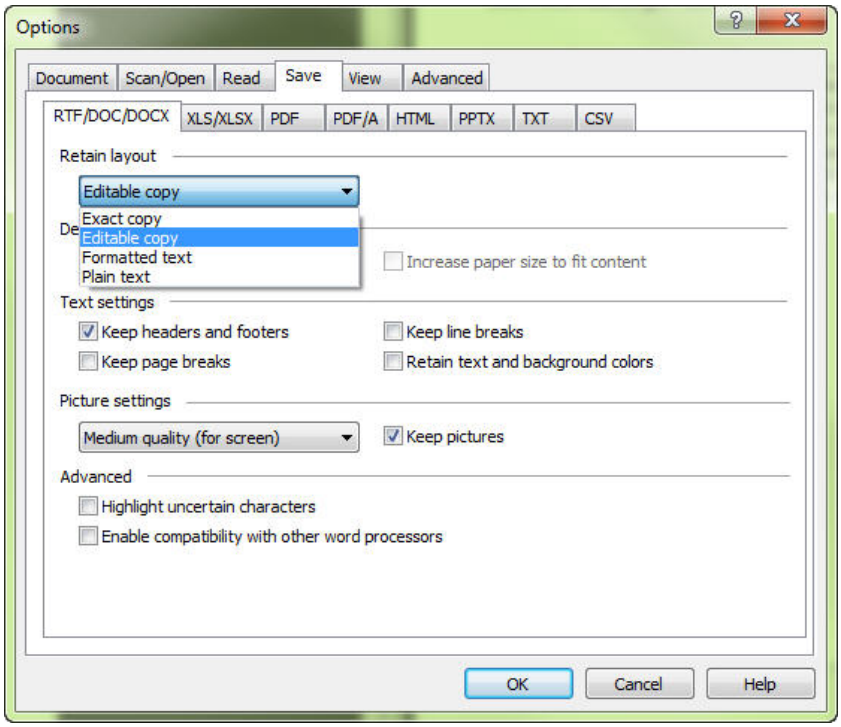

Figure 5. ABBYY FineReader: Available layout options when converting PDF files to Word format.

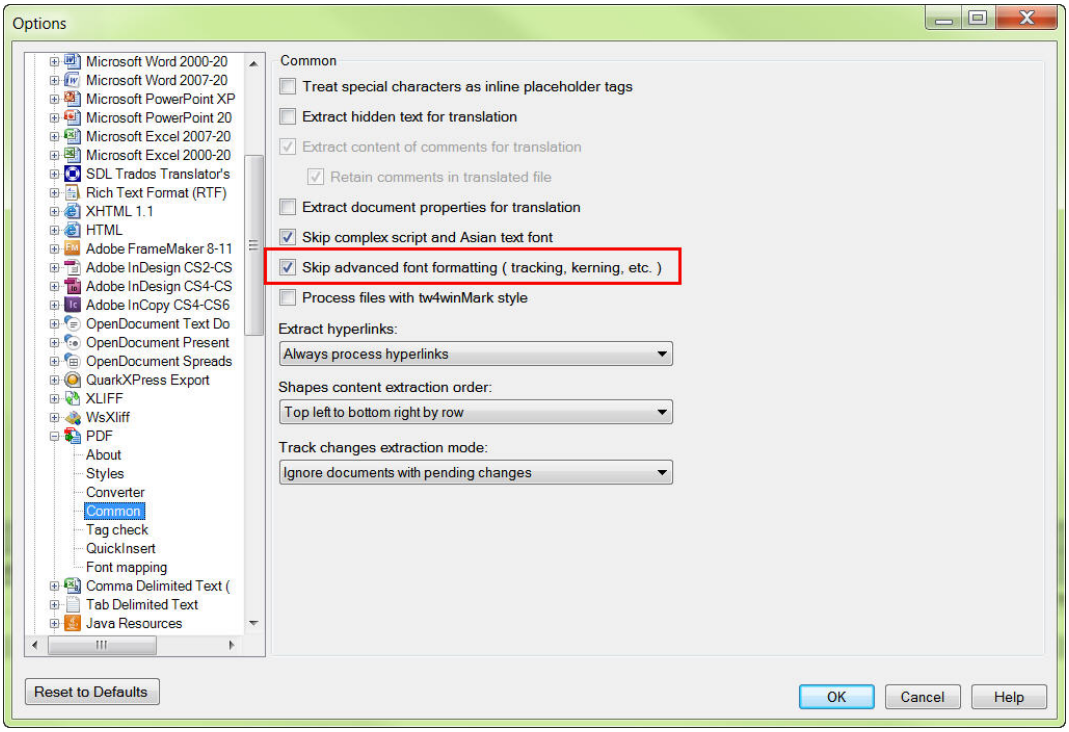

Figure 6. Trados Studio 2011: PDF filter “Common” settings.

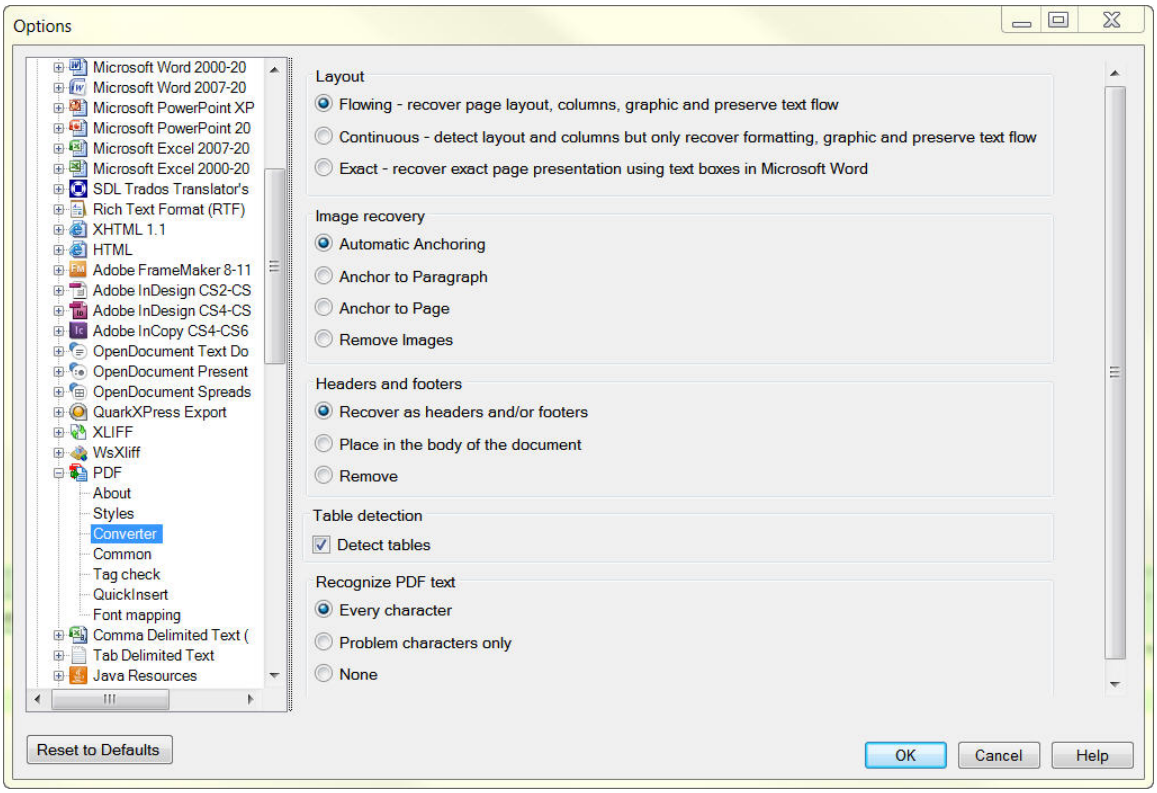

Figure 7. Trados Studio 2011: PDF filter “Converter” settings.

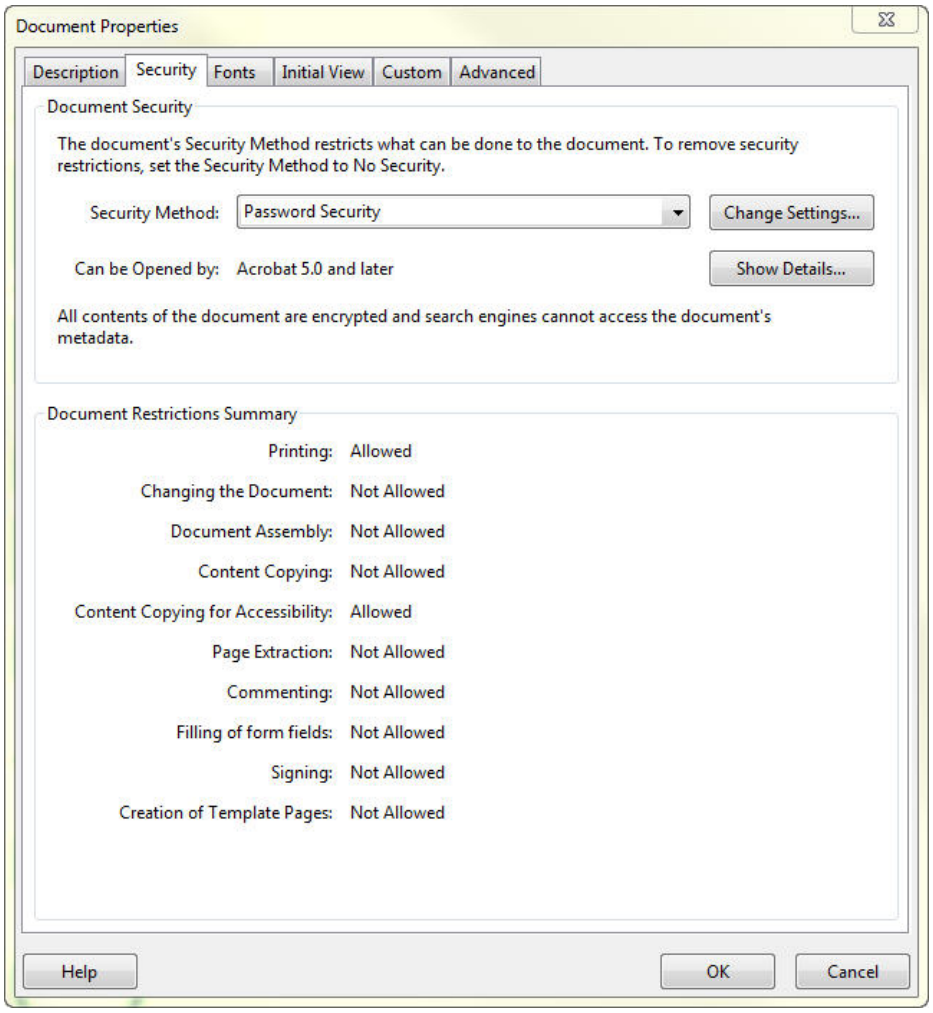

Figure 8. Adobe Acrobat: Security Settings dialog box

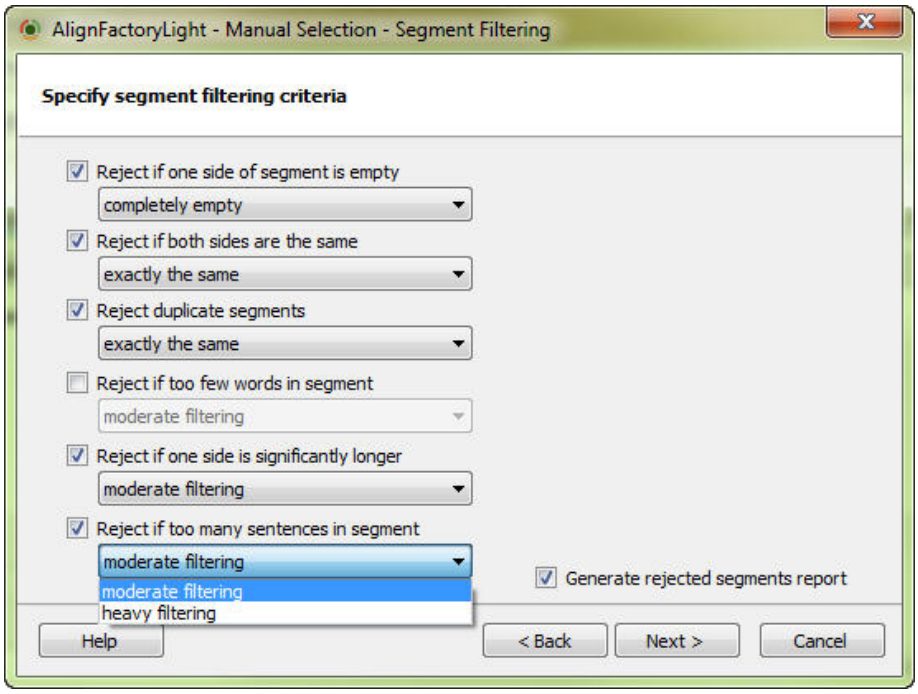

Figure 9. AlignFactoryLight: Available filtering criteria for file alignment.

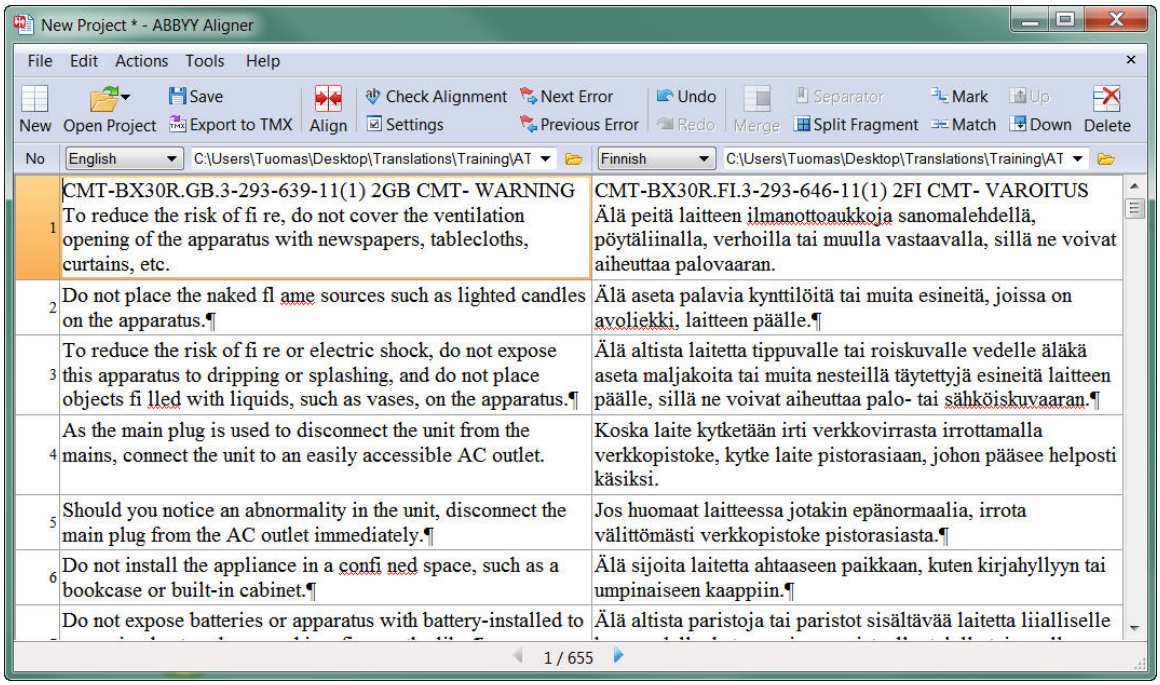

Figure 10. ABBYY Aligner 2.0: User interface.