3 USING TEXTBRIDGE

After you have installed TextBridge as described in Chapter 2, you are ready to use the application to turn paper documents and on-line page images into usable text files.

At your direction, TextBridge can scan a document, or read on-line **TIFF** images, and perform **optical character recognition** (OCR).

TextBridge can also display pages in a **Preview window**, where you can view, **zoom**, and **zone** the page before processing.

During OCR, you can **verify** recognized words as correct, or fix recognition errors. By verifying text, you teach TextBridge to improve recognition accuracy for the rest of the job.

After TextBridge performs OCR, you can direct the application to convert the recognized text to a **text format** that you can use with your favorite word processing, desktop publishing, spreadsheet, database, or other application.

This chapter describes and provides step-by-step procedures for using the main TextBridge application.

+ For information about the TextBridge Application Server and OCR Printer, see Chapter 4.

Specifically, this chapter covers the following topics:

- About TextBridge preferences
- Scanning and converting a document
- Recognizing and converting on-line TIFF files
- Previewing pages before processing
- Verifying text during recognition

ABOUT TEXTBRIDGE PREFERENCES

TextBridge optical character recognition software has a built-in set of **preferences**, settings that control the OCR process. For most good-quality office documents, you can use TextBridge default preferences to provide excellent character recognition.

However, to fine-tune the OCR process for a variety of documents, TextBridge enables you to define preferences on a job-by-job basis. Before starting OCR, you can click the Preferences button in the Main dialog. This displays the Preferences dialog (Figure 3–1).

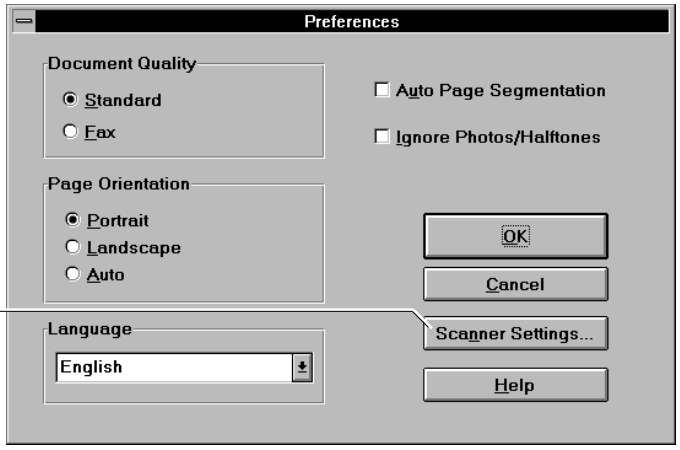

Figure 3–1. Preferences dialog

Specify one or more of the preferences to control how TextBridge will process your document(s). Table 3–1 describes the options available to you.

Note When you change a preference, it remains in place until you change it again, even when you exit TextBridge and run the program again later. This lets you lock in place the preferences that are appropriate for documents you process most often. Of course, you can always change one or more preferences at any time.

Table 3–1. Preferences

Preference	Function
Document Quality	• Choose **Standard** for most documents. • Choose **Fax** for on-line, fax-quality TIFF images (200x100 or 200x200 dots per inch), or if you are scanning hard copy faxes.
Page Orientation	• Click **Portrait** for most typical portrait-oriented office documents. • Click **Landscape** for wide documents that are scanned in sideways, and thus must be rotated in memory by 90-degrees before recognition begins. • Click **Auto** if your document contains a mixture of page orientations (for example, mostly portrait pages, with large rotated tables on landscape pages). Note that you may also want to click Auto if you are recognizing TIFF files and are not sure how the page image is oriented in the file. With Auto selected, TextBridge performs a pre-processing step to determine the orientation of the page. Thus, overall processing speed is slower with this option turned on.
Language	This category provides a pop-up menu of the TextBridge recognition languages that you have installed on your system. TextBridge can perform highly accurate optical character recognition on documents in German, French, Italian, and Spanish, as well as English. Select the primary language of the document to be recognized (for example, German).
Auto Page Segmentation	• Leave **off** for single-column documents. • Click the checkbox **on** for documents with two or more columns of text. With page segmentation on, TextBridge performs a pre-processing step to analyze the shape and location of text blocks on the page, so they are output in the right order. Note that, in the converted text file, TextBridge outputs text in **galley** (single-column) format.

Table 3–1. Preferences (cont.)

Preference	Function
Ignore Photos/ Halftones	Click the **Ignore Photos/Halftones** checkbox on if the document to be recognized contains photographs or halftones (screened photographs). Otherwise, TextBridge tries to recognize the halftone dots as text, reducing recognition accuracy and speed.
Scanner Settings	Click the Scanner Settings button to display the Scanner Settings dialog, in which you can specify scanner-specific controls. • **Use Automatic Document Feeder** is available only for scanners that have attached automatic document feeders (ADF). When you click the checkbox on, TextBridge is instructed to scan and recognize pages from the ADF. • **Brightness** enables you to control the amount of light your scanner shines on a page as it scans it, thus affecting the lightness or darkness of the scanned page image. You can adjust brightness to compensate for different original documents and improve the recognition process. + If you are using an HP scanner with HP AccuPage, Auto is the correct brightness setting. • **Resolution** lets you control the number of dots per inch (dpi) at which the page(s) will be scanned. In general, for best character recognition results, specify the highest resolution, up to 400 dots per inch, that your scanner allows. • **Page Size** lets you control the size of the area the scanner will scan. In general, specify the smallest size that will accommodate the size of your original pages: **Letter** (8.5-by-11 inches or 21.59-by-27.94 centimeters); **Legal** (8.5-by-14 inches or 21.59-by-35.56 centimeters); or **A4** (8.27-by-11.69 inches or 21-by-29.70 centimeters)

SCANNING AND CONVERTING A DOCUMENT

One of the tasks that TextBridge enables you to accomplish is scanning a hard copy document into an on-line text file. The document can comprise one page or many pages.

During this process, TextBridge scans the first page, performs OCR on it, and saves the recognized text in a temporary file. It repeats the process for each subsequent page, appending the recognized text to the temporary file.

When you inform TextBridge that the entire document has been scanned, TextBridge displays the Save As dialog. In this dialog, you specify the output file name, disk and directory, and the text format to save it in.

Note The following procedure assumes that the scanner is properly connected to your PC and is powered on and ready. The procedure also assumes that TextBridge is installed and running.

To scan and convert a document, complete the following steps:

1. **Load the page(s) into your scanner.**

   Depending on your scanner, you can load a stack of pages in the automatic document feeder (ADF), or a single page on the scanner's flatbed.

2. **From the Main dialog, click Scanner in the Input From box (Figure 3–2, next page).**

3. **To save an image of each page to a TIFF file, click the Save Page Images check box.**

   + Page images are stored as TIFF files with **CCITT** Group 3 compression. See Chapter 4 for more information about Save Page Images.

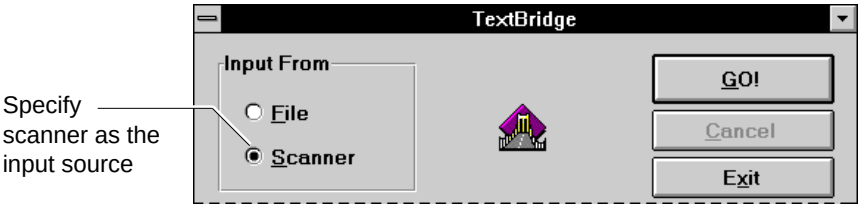

Figure 3–2. Scanner as input source

4. **Optionally, click the Preferences button to further define the scanning and OCR process.**

   Use Table 3–1 as a guide. With Preferences, you can specify such things as page orientation, recognition language, and so on. When you are done specifying preferences, click OK to return to the Main dialog.

5. **Click GO! in the Main dialog to begin the scanning and recognition process.**

   If you are using a TWAIN scanner, the **native user interface** of the source driver appears. Here you can control the scanning process.

   If you are using an ISIS scanner, the page(s) in the scanner are scanned and recognized automatically.

   When you dismiss the native UI (TWAIN), or the page(s) are all scanned (ISIS), TextBridge displays the Add More Pages dialog (Figure 3–3).

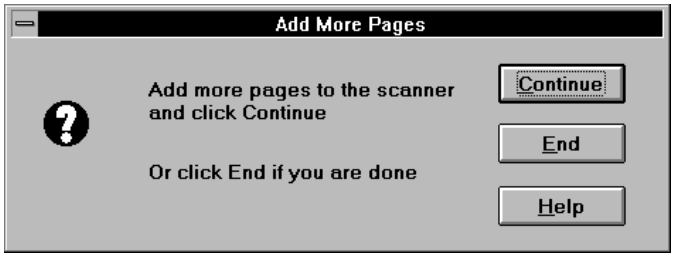

Figure 3–3. Add More Pages dialog

6. **To scan more pages, go to step 7. To end scanning and OCR, go directly to step 8.**

7. **Prepare your scanner, then click Continue.**

   If you are using a TWAIN scanner, the native UI reappears, and you can resume scanning and recognition. If you are using an ISIS scanner, TextBridge automatically resumes scanning and recognition. When you are finished, TextBridge again displays the Add More Pages dialog (refer to Figure 3–3). Proceed from step 6.

8. **Click End in the Add More Pages dialog.**

   TextBridge now displays the Save As dialog (Figure 3–4).

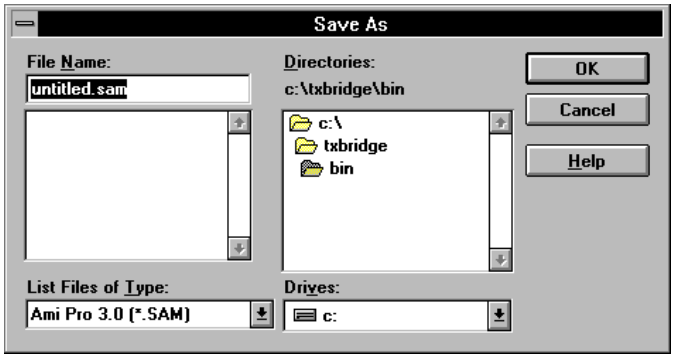

Figure 3–4. Save As dialog

9. **Specify the output file information.**

   - Specify the file name, destination disk and directory.
   - Also specify the output text format in the Save File as Type pop-up menu.
   - Click OK.

   TextBridge converts the recognized text to the specified format, saves the converted text to the specified file, and closes the Save As dialog. The Main dialog remains, ready for the next job.

RECOGNIZING AND CONVERTING TIFF FILES

If you have one or more TIFF files that contain page images, you can use TextBridge to produce a text file from them. To TextBridge, an on-line TIFF image file is similar to one or more scanned page images.

Some TIFF files contain more than one page image. These multiple-page TIFF files are most commonly generated from **fax modems**. TextBridge is equally capable of processing both single- and multiple-page TIFF files.

TextBridge can process TIFF (`*.TIF`) files in the resolutions and formats described in Table 3–2:

Table 3–2. Supported TIFF Resolutions and Formats

Resolutions*	TIFF Formats
100-by-200 200-by-100 200-by-200 300-by-300 400-by-200 400-by-400	Uncompressed (Intel header) CCITT-3 (Intel header) CCITT-4 (Intel header) Uncompressed (Motorola header) CCITT-3 (Motorola header) CCITT-4 (Motorola header) Intel FAXability

* All TIFF files processed to one document must be of the same resolution. If TextBridge has processed several files that are 200-by-200 dpi, for example, and the next file that it encounters is 300-by-300 dpi, the program generates an error message, then displays the Save As dialog box so you can save the recognition results up to that point to an output document.

At your direction, TextBridge opens the TIFF file(s), reads the image data into computer memory, performs OCR on the image(s), and then saves the recognized text to a formatted text file.

Note The following procedure assumes that TextBridge is properly installed on your PC and running.

To recognize and convert one or more on-line TIFF files to a text file, use the following procedure:

1. **From the Main dialog, click File in the Input From box (Figure 3–5).**

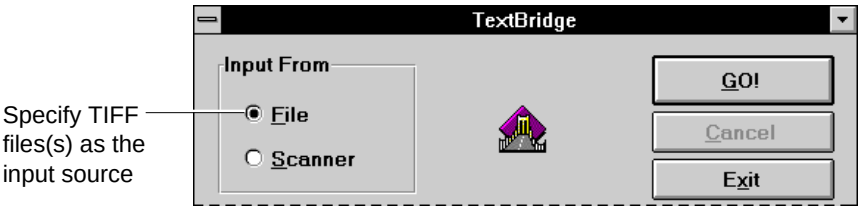

Figure 3–5. Input From File

2. **Optionally, specify Preferences to control the recognition process.**

   - Click the Preferences button to display the Preferences dialog (refer to Figure 3–1).
   - Change one or more of the preferences in place. Refer to Table 3–1 for information.
   - When you are done specifying preferences, click OK to return to the Main dialog.

3. Click GO! in the Main dialog.

TextBridge displays the Open dialog (Figure 3–6).

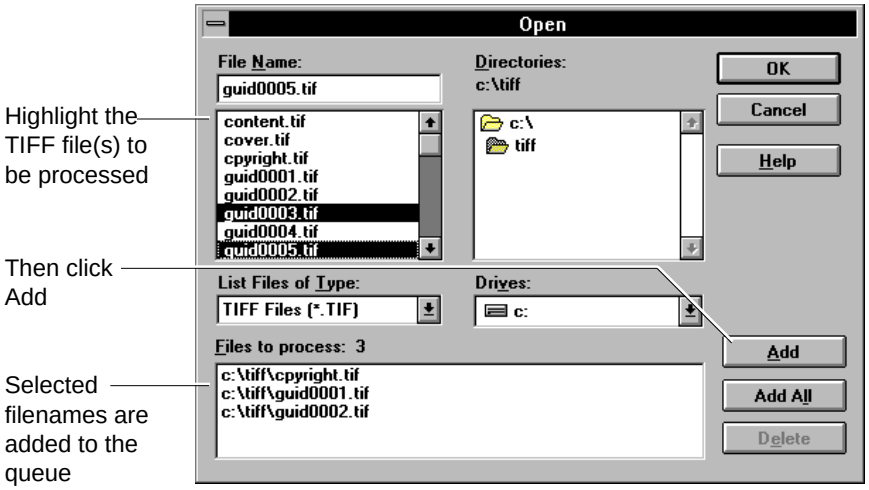

Figure 3–6. Open dialog

4. Specify the TIFF file(s), then click OK.

+ To select a single file, click on it in the File list box, then click Add to add it to the **Files to process** list. Or, simply double-click the file.

  To select a range of files, point to the first file in the range, click, hold, and drag the mouse downward in the File list box. Or, click the first file in a range, then Shift-click (hold down the Shift key and click the mouse) on the last file in the range. With the file range selected, click Add to add the files to the process list.

  To select TIFF files that are not in sequence, click the first file, then Control-click (hold down the Control key and click the mouse on) subsequent files in the file list.

+ To add all files in the current directory to the process list, click the Add All button.

  To add files from another directory, change to the other directory, and add the files as noted here. Note that the full pathnames of the files are added to the process list.

  To delete files from the process list, you can select one at a time, and click the Delete button. Or, you can select a sequence of files, or a non-sequential group of files in the same manner as described above, and click the Delete button.

  Files are processed in the order in which they are added to the queue. As you add files, the number of files in the process list is tracked by a counter (Figure 3–7).

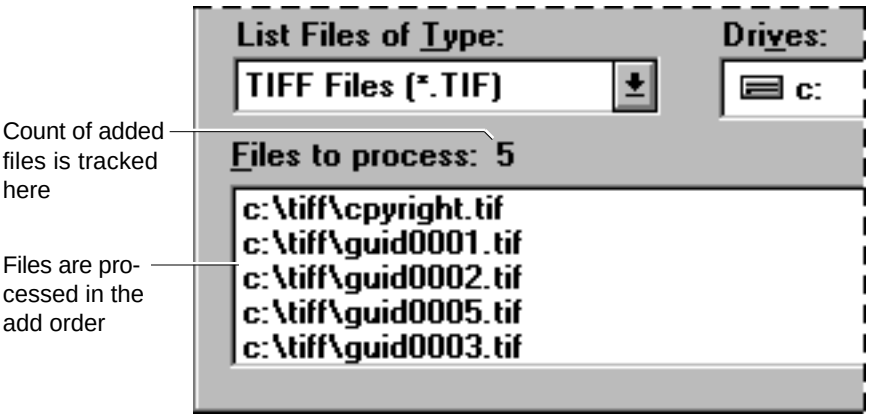

Figure 3–7. Multiple TIFF files to be processed

After you add the file(s) and click OK in the Open dialog, TextBridge performs OCR on the page image(s). When OCR is complete, TextBridge displays the Save As dialog (refer to Figure 3–4).

5. **In the Save As dialog, specify the output file information.**

   - Specify the file name, destination disk and directory.
   - Also specify the output text format in the Save File as Type pop-up menu.
   - When done, click OK.

   TextBridge converts the recognized text and saves the converted text to a file of the given name, disk and directory location. It then displays the Main dialog, ready for the next job.

PREVIEWING PAGES BEFORE PROCESSING

To view a page before processing, or to define areas of a page to process, TextBridge provides the Preview window (Figure 3–8).

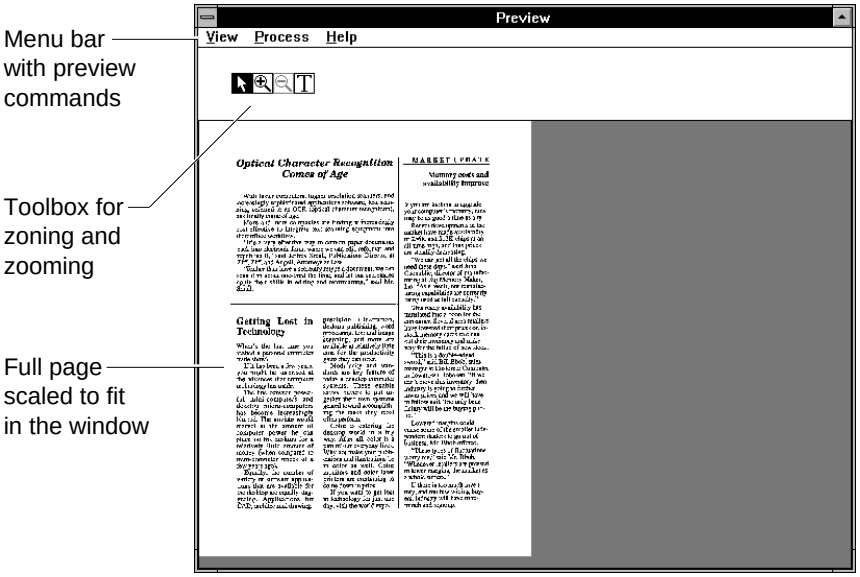

Figure 3–8. Preview Window

In the Preview window, you can use the Zoom tools in the Preview toolbox to zoom in and out on the page at several magnification levels:

Zoom In

Zoom Out

With the Zoom tools, you can magnify a page to full **resolution**, zoom out to fit the page entirely in the window, or display the image somewhere in between.

To define a portion of the page to be processed, you can use the Text Zone tool to draw up to 127 rectangular **zones** around specific areas on the page:

Text Zone

Zones are numbered to show the order in which you created them and the order in which contained text will be output in the finished file. Overlapping zones are opaque; the topmost zone "owns" any common text it shares with the underlying zone(s).

The Select Zone tool lets you select any of the zones you have created:

Select Zone

A selected zone is identified by solid corner **handles**. After you select a zone, you can move it, resize it, delete it, or put it in front or in back of another zone.

For processing purposes, you can adjust the zones page-by-page or have the zones take effect for all pages of the document. When the job is complete, TextBridge automatically clears the zones.

Note The following step-by-step procedure assumes that, if you are using a scanner, it is properly connected to your PC, powered on and ready. The procedure also assumes that TextBridge is properly installed and running.

To preview a document before processing it, use the following procedure:

1. **If you are scanning, load the page(s) into your scanner, then go to step 2. If you are processing a TIFF file, start at step 2.**

2. **In the Main dialog, define the input source.**

   - Select either File or Scanner in the Input From box (refer to Figures 3–2 and 3–5).
   - Click the Preview box in the Main dialog.

3. **Optionally, click the Preferences button to further define the scanning and OCR process.**

   Use Table 3–1 as a guide. When you are done with preferences, click OK to return to the Main dialog.

4. **Click GO! in the Main dialog.**

   If you are scanning with a TWAIN scanner, the native UI for the scanner appears, and you can execute scanning from this interface. If you are scanning with an ISIS scanner, TextBridge automatically scans a page from the scanner.

   If you are reading the page image(s) from one or more TIFF files, TextBridge first displays the Open dialog (refer to Figure 3–6). In the Open dialog, specify the name(s) of the TIFF file(s), and their directory and drive location, then click OK.

   TextBridge opens the Preview window, and displays the first scanned or on-line page image (refer to Figure 3–8).

5. Optionally, zoom the page image.

The page first appears zoomed out to fit in the window. To quickly zoom to full resolution, you can pull down the View menu and choose the **Zoom Max** command. To zoom in on the page image by steps, click the Zoom In icon from the Preview toolbox, and click on the page display.

When the page is zoomed in, scroll bars appear to let you shift the display horizontally and vertically (Figure 3–9).

To zoom all the way out, pull down the View menu and choose the **Zoom Min** command. To zoom out in steps, click the Zoom Out icon, and click on the page display.

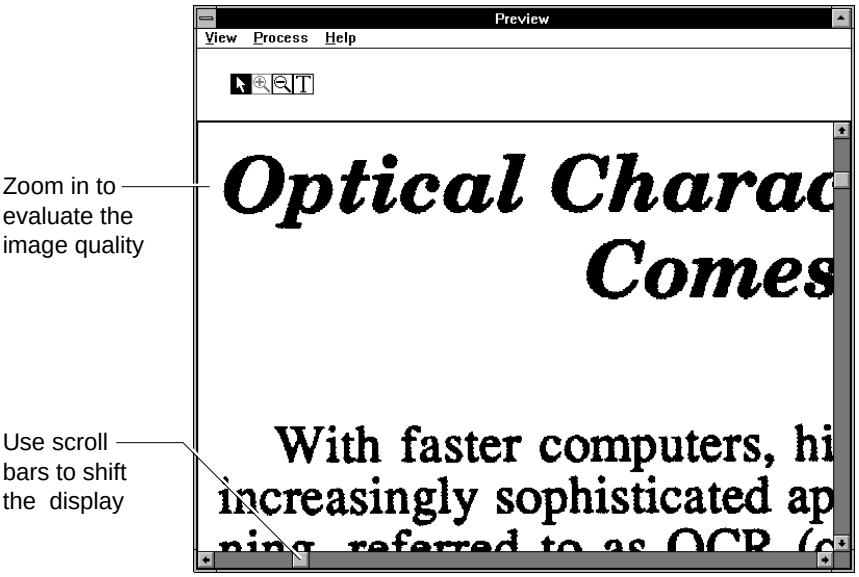

Figure 3–9. Previewed page zoomed in

6. **Optionally, create one or more zones to define the page area(s) to be processed.**

   Up to 127 zones can be created, as follows:

   - Click the Text Zone tool, point to a corner of the area to be zoned, click and hold the left mouse button, and drag the mouse. A zone rectangle appears as you are moving the mouse.
   - When the zone is fully sized as you intend, release the mouse.

     + The zone number appears in the upper left corner. Handles appear on the zone rectangle for resizing purposes (Figure 3–10).

       The zone number indicates the order in which recognized text blocks will be output in the finished text file.

       To **resize** a zone, click and hold on a corner handle and drag the mouse.

       To **move** a zone, click and hold inside the zone, and drag the mouse.

       To **position** a zone relative to another zone, make sure the zone is selected (handles are black), then pull down the View menu and select the **Move to Front** or **Move to Back** command.

       Overlapping zones are opaque, meaning that any image area shared by more than one zone is output as part of the topmost zone.

       To **delete** a zone, pull down the View menu and choose the Clear Zone command (or simply press the Delete key).

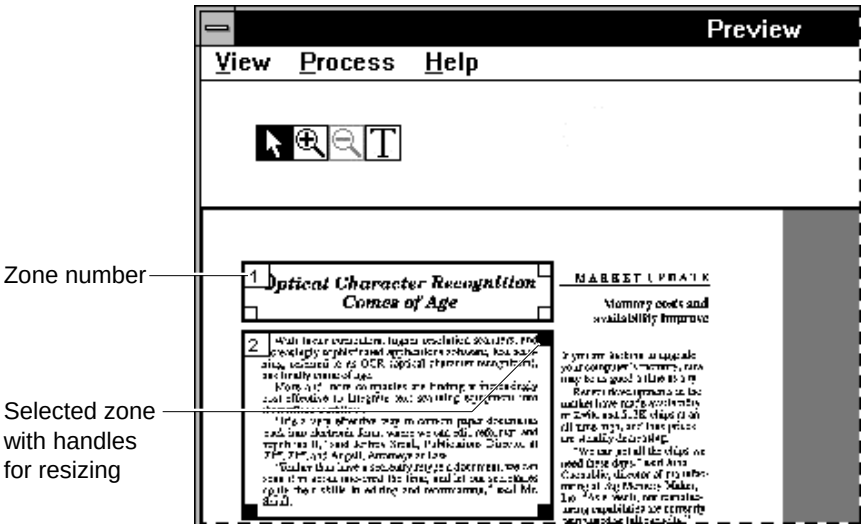

Figure 3–10. Zones in Preview

7. When you are done previewing the page, start the OCR process.

Pull down the Process menu and select This Page or All Pages to start/resume OCR.

+ Select All Pages to close the Preview window and process all pages of the job to the zones in place.

  Select This Page to preview and process every page individually. TextBridge will process the first page, scan, or read, and display the next page in the Preview window. TextBridge then goes into idle mode. Proceed from step 5 to preview the now-displayed page.

  When you have previewed the last page of the document, select the All Pages command to close the Preview window and have TextBridge finish OCR of the document.

When OCR is completed, TextBridge displays the Save As dialog (refer to Figure 3–4).

8. **Specify the output file information.**

   - Specify the file name, destination disk and directory.
   - Also specify the output text format in the Save File as Type pop-up menu.
   - When done, click OK.

   TextBridge converts the recognized text and saves the converted text to a file of the given name, disk and directory location. The Main dialog remains, ready for the next job.

VERIFYING TEXT DURING RECOGNITION

To achieve the highest possible OCR accuracy, even on difficult documents, TextBridge provides a **word verifier** (Figure 3–11).

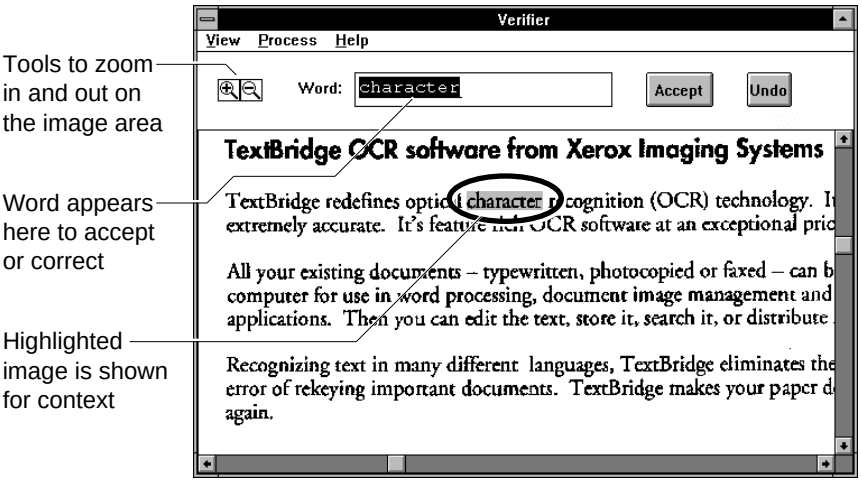

Figure 3–11. Word Verifier window

The word verifier displays **questionable words** for you to correct and/or verify during recognition.

Questionable words are those that fall below a **confidence threshold** built into TextBridge. During OCR, TextBridge assigns a confidence value to each word. If the value falls below the confidence threshold, and you are using the Verifier, TextBridge displays the word as questionable.

By correcting errors and verifying correct words, you help TextBridge improve its recognition accuracy for the rest of the document. TextBridge uses your input to improve recognition decisions as the job progresses.

The Verifier window, which is shown in Figure 3–11, is similar to the Preview window. In part of the Verifier window, TextBridge shows a zoomed image of the page with the questionable word highlighted for context.

Above the image area, zoom tools enable you to zoom further in or out on the page image:

Zoom In

Zoom Out

Next to the toolbox is the Word **edit box** containing the recognized word, which is highlighted for correction or verification. To the right of the edit box are the Accept and Undo buttons.

To verify words during OCR, use the following procedure.

Note The procedure assumes, if you are scanning, that the scanner is properly connected to your PC and is powered on and ready. It also assumes that TextBridge is properly installed and running.

---

1. **If you are scanning, load the page(s) into your scanner, then go to step 2. Otherwise, start at step 2.**

2. **In the Main dialog, define the input source.**

   - Select either File or Scanner in the Input From box (refer to Figures 3–2 and 3–5).
   - Click the Verifier checkbox in the Main dialog.

3. **Optionally, click the Preferences button to further define the scanning and recognition process.**

   Use Table 3–1 as a guide. When you are done specifying preferences, click OK to return to the Main dialog.

4. **Click GO! in the Main dialog.**

   If you are scanning with a TWAIN scanner, the native UI for the scanner appears, and you can execute scanning from this interface. If you are scanning with an ISIS scanner, TextBridge automatically scans a page from the scanner.

   If you are reading the page image(s) from one or more TIFF files, TextBridge first displays the Open dialog (refer to Figure 3–6). In the Open dialog, specify the name(s) of the TIFF file(s), and their directory and drive location, then click OK.

   TextBridge begins the OCR process. When it encounters the first questionable word, it opens the Verifier window, and displays the questionable word highlighted in a Word edit box (refer to Figure 3–11). Below the edit box, the Verifier window displays the image of the word highlighted for context.

5. **Verify and/or correct the questionable word in the Word edit box, then click the Accept button.**

   + If you make a mistake during verification, simply click the **Undo** button. The edit box restores the last word you edited, and you can then correct the mistake.

     You can control the frequency of questionable words to verify with the **Verify** command, which has five possible settings. Pull down the Process menu, select the Verify command, and from its walking menu, select a setting from among Most, More, Normal, Fewer, or Fewest. More and Most will cause more words to be displayed for verification. Less or Least will show fewer words. Normal is the default.

     Some documents have words with characters that are not available on your keyboard (accents, symbols, and so on). To verify such characters, pull down the View menu and select **Show Special Characters**. This displays the special character keypad. To enter special characters in the Word edit box, select them from the keypad by clicking on them with the mouse (Figure 3–12).

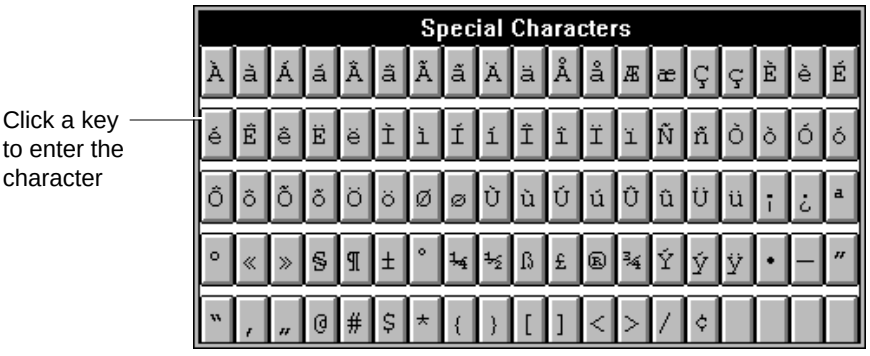

Figure 3–12. Special character keypad

+ Occasionally, TextBridge mistakes **noise** (marks on the page) or a horizontal line as text. In the Verifier window, the Word edit box contains characters, while the image area shows the non-word highlighted. In such cases, delete all the text in the Word edit box, then click Accept. TextBridge ignores the noise, and proceeds to the next questionable word.

  Also, if the image of a particular word is poor in comparison with the rest of the document, you should correct the text without training TextBridge on the image. Simply correct the word in the Word edit box, then hold down the Control key and press the Enter key (or click Accept). The corrected text is output without TextBridge being trained on the image.

  The image area in the Verifier window is zoomed in to approximately the middle of the zoom range. If you want to get more of an idea of the page location of the word being verified, click the Zoom Out icon, then click inside the Verifier image area. The full page image appears in the image area of the Verifier window. Conversely, if you want to further magnify the display, use the Zoom In icon.

6. Repeat step 5 until you verify enough words.

+ Verify at least one page of a multiple-page document to teach the system about the entire document.

7. Close the Verifier.

Pull down the Process menu and select the **End Verification** command.

+ The Close command in the Verifier window's control menu is equivalent to the End Verification command.

The Verifier window closes and OCR continues automatically. When you are finished processing all pages, TextBridge displays the Save As dialog (refer to Figure 3–4).

8. Specify the output file information.

- Specify the file name, destination disk and directory.
- Also specify the output text format in the Save File as Type pop-up menu.
- When done, click OK.

TextBridge converts the recognized text and saves the converted text to a file of the given name, disk and directory location. The Main dialog remains, ready for the next job.