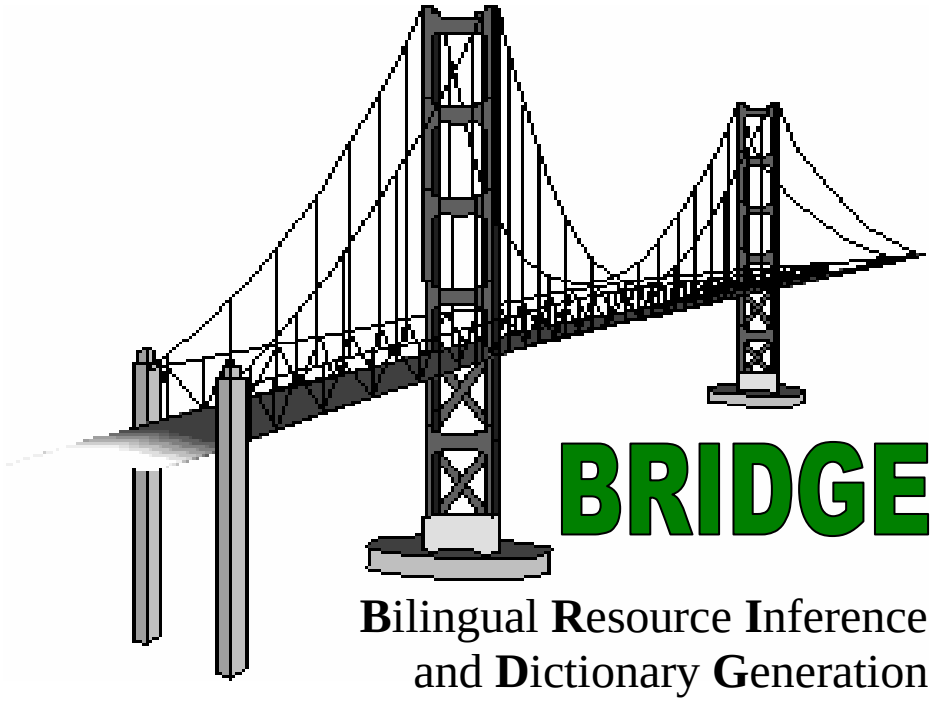
USER MANUAL

Bilingual Resource Inference and Dictionary Generation Environment (BRIDGE)

Version 0.1
March 23, 2004



Bilingual **R**esource **I**nference
and **D**ictionary **G**eneration
Environment

The BRIDGE system was developed in part by DARPA under the TIDES program and contract xxx and by the Center for the Advanced Study of Language (CASL) at the University of Maryland under contract yyy.

TABLE OF CONTENTS

1 Introduction

Bilingual dictionaries hold great potential as a source of lexical resources for training automated systems for optical character recognition, machine translation and cross language information retrieval.

The Bilingual Resource Inference and Dictionary Generation Environment (BRIDGE) system is a tool that investigates methodologies and techniques for automating the process of acquiring lexical knowledge about Less Commonly Taught Languages (LCTL). The main goal of this tool is the rapid production of a high quality OCRed dictionary that can serve both as an enhanced standalone aid, and as a lexical resource for language processing systems. The scope of this tool covers the acquisition of written (rather than spoken) language and lexical resources, incorporating information that can be obtained from a paper dictionary.

In our tool, the user interface is designed in Java and the backbone processing programming is done in C++.

System requirements:
Disk space: At least 60 MB of free disk space.
Display Adapter: ATI FireGL X1-128 Video Accelerator (Current System). The graphics card should have 1 MB of memory.
Monitor: 1280 * 1024 (Current System)

External requirements:
Scanner (Currently Fujitsu fi-4220C)

2 System Overview

BRIDGE provides a Dictionary-to-XML capability where the operator of the system can acquire, segment, tag and generate a formatted electronic version of a hardcopy bilingual dictionary source given minimal knowledge about the secondary language. BRIDGE is a staged approach to dictionary parsing and tagging. Pages are scanned into the system, and stored individually. The process consists of: **Acquisition and OCR, Segmentation, Tagging,** and **Generation,** with **Access** as a follow-on process. The goal of the interface it to provide a seamless integration of the stages, along with the ability to correct errors in the processing at will. Figure 2-1 provides an overview of the general process. Each basic module is highlighted below and described in detail in the following sections.

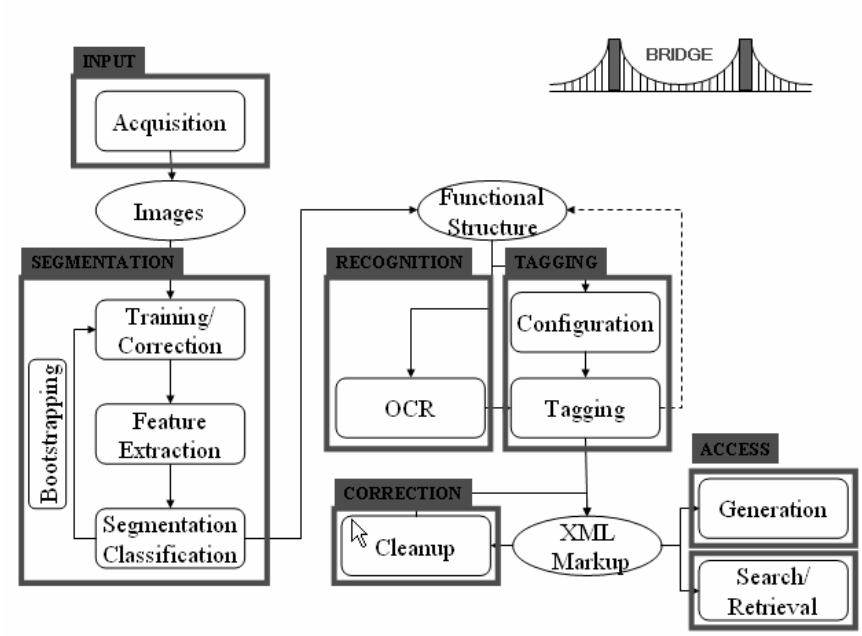


Figure 2-1 System Overview

Setup: The system is administered through a set of configuration panels available from a dropdown menu. It has the ability to set up new dictionaries, install the directory structure and configure default parameter files for each of the modules. At each stage in the process, the operator will set various parameters to customize the interpretation and processing of the dictionary.

Acquisition: We assume that the operator has control of the scanning process. Our scanning specifications are as follows.

- Burst the binding of the dictionary
- Scan dictionary pages at 300-400 dpi, according to a defined naming convention.

The system is fairly sensitive to the scanning process so the operator should make sure that the highest possible quality scans are obtained.

OCR: We are using ScanSoft SDK for OCR, with the goal of providing a consistent character representation of the content. OCRing the images prepares them for further processing such as segmentation, tagging and generation.

Segmentation: The goal of segmentation is to identify individual dictionary entries physically in the image. The operator prepares sample segmentation pages and the system is trained to learn page features. The system segments the pages based on the learned features. The result is feedback into the system to refine the segmentation procedure.

Tagging: The purpose of the tagging process is to label different types of linguistic constructs, such as headwords, translations, parts-of-speech (POS) or pronunciations, for each dictionary entry, so that resources can be generated. Tagging uses the repetitive structure of each entry to learn the labels. The operator provides the features for learning.

Generation: The generation module is responsible for taking the output of tagging and producing various result sets. The operator will be able to select from different formats. Currently the following formats are supported: TEI, Rosetta, Lexicon-Translation Pairs, Example of Usage-Translation Pairs, HTML and HTML with Images.

At various stages in the system the operator can correct word and character attributes. The "correction modules" allow the operators to ultimately build a "perfect" resource.

Correction:

OCR Correction Interface: An OCR correction interface is built on top of the OCR correction module. It consists of an edit window that will show the recognized text formatted in the same way as the original dictionary. The suggestion window provides options for words found in the dictionary or suggested by the OCR correction module. The original image clip is shown in the right window. Operators will be able to make corrections, and if desired, propagate them throughout the remainder of the document.

Labeling Correction: Labels can be corrected by selecting groups of regions and relabeling them on the workflow panel.

After the resources are generated, the results can be used for follow-on applications or used to drive tools such as BRIDGE-View.

Search and Retrieval: The BRIDGE-View application was developed to fill the gap between noisy data and the page image. It allows users to search the results of parsing and then return to the image of the original dictionary.

3 Installation of the System

At present, we have preinstalled the entire system onto a desktop personal computer. The compilation and execution is done using a “batch” file. The BRIDGE system uses BRIDGE-GUI.bat, and Search and Retrieval is run using BRIDGE-View.bat.

4 Overview of the User Interface

The interface (shown in Figure 4-1) consists of various panels arranged in a Java GUI. Overall the interface is broken into 3 main areas: The Menu Operation Area (Menus), the Workflow Panels Area and the Display Area.

Menus: The Menu Operation Area (Menus) consists of menu controls and their sub-menus. These controls are used to control a majority of the high level operations.

Workflow Panels: The Workflow Panels Area holds control panels for the five different working modes. These five different working modes are: OCR, Segmentation, Tagging, Generation and View which can be selected by clicking the panel with the mode name. The first four modes are described in sections Optical Character Recognition (OCR), Segmentation, Tagging and Generation respectively. View is described in section BRIDGE-View. The Workflow Area can be used to classify different parts of the image, to divide the image into zones and to use these zones to train the system. For the first four working modes, the four panels in the workflow have a similar look and feel, allowing the operator to open images, display metadata, and control execution through a series of processing buttons, and to add corrections back to a training set. The View mode will have two buttons namely Create View and Run View. The details will be explained in section BRIDGE-View.

Display Area: The third part of the interface is the Display Area. The Display Area consists of a Toolbar Area and an Image Display Area. The Toolbar Area consists of shortcuts (tools) for the menus and image scaling operations. The Image Display Area displays the scanned dictionary page image. A detailed explanation of these parts is given in the following sections.

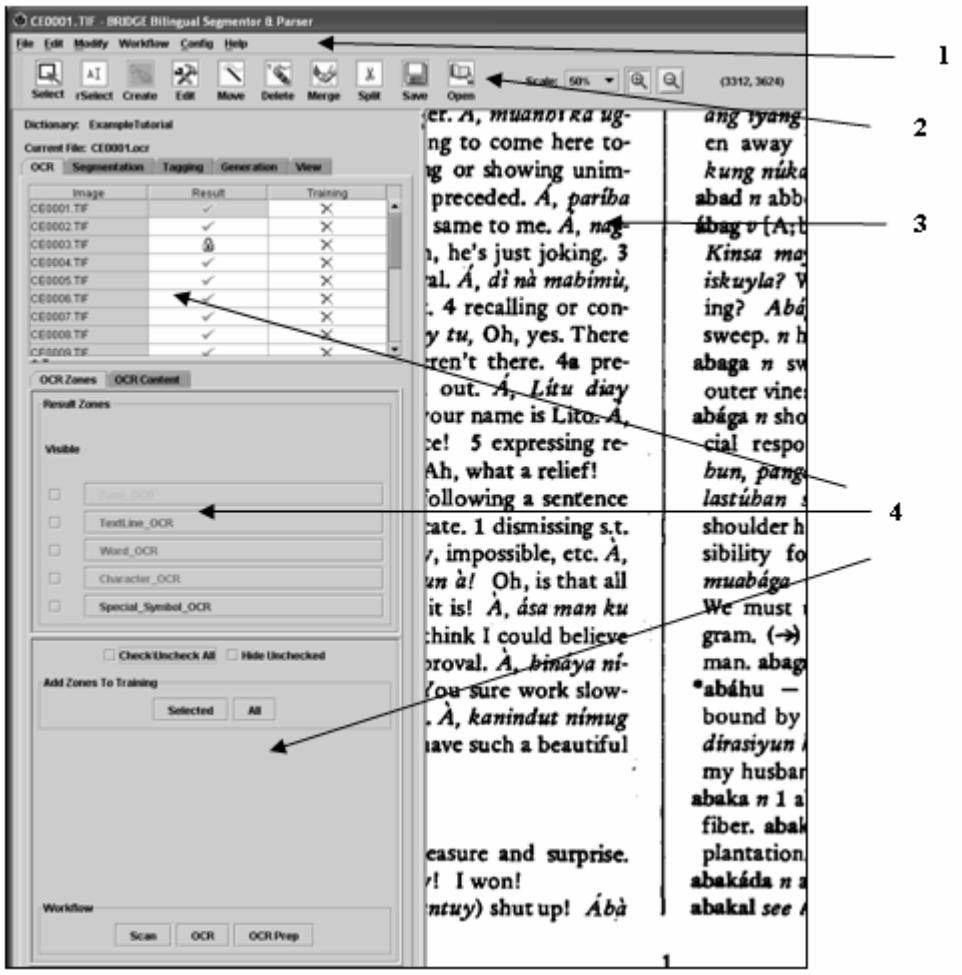


Figure 4-1 Overview of the GUI; 1: Menus, 2: Toolbar Area, 3: Image Display Panel, 4: Workflow Panels

4.1 Menus

The menu bar contains the controls to operate the interface. Although many of the same operations are found in the individual panels, they can typically be found in the pull downs as well.

4.1.1 File

Open a Dictionary: Clicking on this menu will open up a FileChooser dialog. The operator then selects the appropriate dictionary to process. The dictionary to be selected should already exist. Preparation of a dictionary is explained in detail in Section 5.2 Creating and Manipulating Dictionaries.

Refresh Dictionary: After any modification, if the selected dictionary is not updated automatically, the operator should select this menu option to refresh the dictionary and update it.

Save: The operator can alter the data or add new training information. Selecting this menu item will show a "Save Modified Files?" dialog box that lists all changed files, that can be saved as seen in Figure 4-2.

Figure 4-2 "Save Modified Files?" dialog

Delete: Not implemented currently.

Close: Disabled currently.

View Log File: If there is any problem, the operator can view the system log by clicking on this menu and opening the log file. Information such as "Could not find Tagging config file" is printed for debugging purposes.

Exit: The operator selects this option to exit the system. The operator will be prompted and the decision will be confirmed before exiting. The system will also prompt the operator to save the changed data before exiting.

4.1.2 Edit

Undo: The operator can cancel a recent operation and return to the previous status by clicking on this menu option.

4.1.3 Modify

This menu and its submenus provide various editing operations that can be performed on the display area. This helps to create new regions. The functions provided are creating (Create) and selecting boxes (Select and Reading Order Select), moving and deleting the boxes (Move and Delete, respectively), merging the selected boxes (Merge) and splitting them (Split). All the menus have shortcuts on the toolbar. All the sub-menus also have mnemonics or keyboard shortcuts. They are explained in Appendix 7.2.

4.1.4 Workflow

This menu and its sub-menus control the different operations in the processing of the dictionary. The sub-menus are described below.

OCR: It contains three items **Scan**, **Run OCR and OCR Prep**. Clicking the **Scan** option will bring up the scanning software and its configuration interface. The operator should save the scanning output in an appropriate directory. After scanning the pages, the operator can run OCR on them by clicking the **Run OCR** option. Clicking the **OCR Prep** option will bring up the **OCR Organizer** dialog. OCR Prep organizes the raw data into a hierarchical structure. These options have the same function as the buttons **Scan**, **OCR** and **OCR Prep** in the Workflow area of the OCR panel, respectively.

Segmentation: It contains only one item **Segment**. Segmentation helps to mark the image into distinct entries or regions. Based on this classification, the system can be trained using feedback techniques (Bootstrapping). This sub-menu has the same function as the button **Segment** in the Workflow area of the Segmentation panel.

Tagging: It has a sub-menu **Tag**. The function of this sub-menu is to classify different word parts of the image such as headword, part of speech, number, translation, etc. This classification can be used while generating the output. It has the same function as the button **Tag** in the Workflow area of the Tagging panel.

Generation: It contains only one item, **Generate**. This sub-menu enables the operator to generate the data in different formats like HTML, XML, etc. It has the same function as the button **Generate** in the Workflow area of the Generation panel.

4.1.5 Configuration

Config Utility: Under the **Config** (Configuration) menu, there is a sub-menu **Config Utility**. Selecting this item will pop up the **Bridge Preferences** dialog box which allows the operator to configure different parts of the system including System Path, Dictionary, Zone Colors, Scanning, OCR, Segmentation, Tagging and Generation. The configuration for OCR, Segmentation and Generation are reserved for future implementation. Configuration for Dictionary will be described in Section 5.2 Creating and Manipulating Dictionaries. Scanning will be described in Section 5.3 Scanning, and Tagging will be described in Section 5.6 Tagging.

Personal Dictionary: Using this sub-menu, the operator can set up and modify personal wordlists (dictionaries) for spell check operation on the OCRed text. A detailed explanation of this procedure is given in section 5.4.4 Using a personal dictionary (ies) for spell check

4.1.6 Help

Help: Pops up the help manual.
Overview: Opens a PowerPoint file which gives the overview of this project.
Future Work: Pops up a "To do ..." list.
About Bridge: Version and copyright information.

4.2 Workflow Panels

The workflow panels are located on the left side of the interface. Currently, there is one for each of the following; OCR, Segmentation, Tagging, Generation and View. Within each panel (except for View), there are Image Access, Labeling, Workflow and information panel components. View mode has two buttons **Create View** and **Run View** on its workflow panel.

4.2.1 Panel Components

The following figure shows the workflow panel for the segmentation portion. The name of the current dictionary that is in use is shown in the top left corner of the figure.

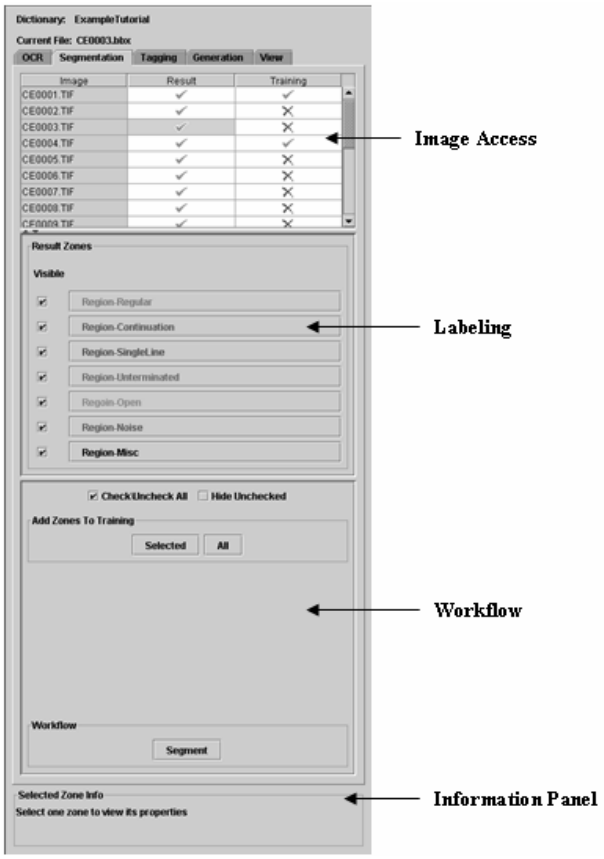


Figure 4-3 Segmentation Panel

Description of different parts of the Workflow Panel is as follows:

Image Access: This panel has three columns. The leftmost column will display the current Image files (pages in the dictionary). The middle column is the Result column and will indicate if the output of this stage has been generated. The Training column indicates if a training data has been prepared or not. For the Result and Training columns a red cross ✕ indicates that the data is not available, while a green check mark ✓ indicates that the data is ready and that this file can be forwarded to the next stage. There is also a provision to lock a file. The operator can Right Click on a file in the Result or Training column to lock it. If a file is locked, it cannot be edited or modified until it is unlocked (in a similar manner).

Labeling: The Labeling sub-panel is used to select labels for various regions in the image such as new Training zones. The operator selects which zones to view in an image by selecting the checkboxes that apply. To create new zones for training, the operator should double click on the zone label. This will activate the Create sub-menu. Then the text in the image can be selected by dragging with the mouse pointer.

Workflow: The workflow sub-panel contains options to control dictionary processing. Depending on the mode of processing (OCR, Segmentation, Tagging or Generation), the operator can select which zones to display and which to add to training using the checkboxes provided. This sub-panel also contains the buttons to perform OCR, Scanning, Segmentation, Tagging etc. Depending on the processing stage, one or more of these controls will be available to the operator.

Information Panel: This panel gives information about the selected zone including the co-ordinates and the type of the selected region.

In the following sections, a brief description of the workflow panel for different stages in processing is provided.

4.2.2 OCR Panel

The OCR Panel shows different textual zones in the dictionary image labeled as Zone_OCR, TextLine_OCR, Word_OCR, Character_OCR and Special_Symbol_OCR. There are two modes - Result mode and Training mode. When in the Result mode, text can be extracted from the image and displayed in the **OCR Content** sub-panel. The operator can perform spell check on the **OCR Content**. This panel allows for the correction of any mistakes that the software may have made. The Labeling sub-panel has another tab that uses special characters when in the Training mode. When in the Result mode, the selected zones can be added to train the system using the corresponding option in the workflow panel. The workflow sub-panel also contains buttons for **Scan,**

OCR and **OCR Prep**. The scanning operation is to be followed by OCR. The result will prepare the file for segmentation. Availability of results is indicated by a green check mark in the Result column.

4.2.3 Segmentation Panel

In segmentation, the operator can divide the image into different regions such as Regular, Continuation, SingleLine etc. These options are provided on the Labeling sub-panel. In the Workflow sub-panel, an option is provided to display only the selected zones. The ghost zones let the word boxes be displayed. The workflow sub-panel also contains the button **Segment**. **Segment** will segment the image into the zones.

4.2.4 Tagging Panel

Tagging allows the operator to classify the parts of text as Headword, Part of Speech, Translation, etc. These options are displayed on the Labeling sub-panel. Accordingly, the zones can be marked as appropriate types and added to train the system. There is also an option to show word boxes (Ghost Zones). The workflow panel contains the **Tag** button. The tagging result is needed for the generation part of processing.

4.2.5 Generation Panel

The Generation panel has a display capability in six different formats, TEI, Rosetta, Termlist, Example of Usage, HTML and HTML with Images. Depending on the user requirements, the output can be generated in a suitable format.

4.2.6 View Panel

The View panel has the buttons Create View and Run View. These options create the underlying data for BRIDGE-View and run BRIDGE-View.

4.3 Display Area

The display area is divided into two main parts (1) Toolbar Area and (2) Image Display Panel. The toolbar area contains shortcuts or tools for the menus. The Image Display Panel is used to display the scanned page of the dictionary and the different bounding boxes with different types. All box operations should be performed in this area. The following tables contain the information about different toolbar shortcuts and their corresponding menus. These tools are used for visual manipulation of the image.

4.3.1 Image Scaling Tools

The toolbar provides three controls to scale the image to the required resolution. Their description is given in the table below.

Control	Toolbar symbol	Description
Image Scaling		Scales the display of the image up or down within the display area.
Zoom in		Increases the image resolution.
Zoom out		Decreases the image resolution.

Table 4-1 Image Scaling Tools

4.3.2 Image Editing Tools

The toolbar also contains shortcuts for the sub-menus in the Menu Operation area. The following table displays these shortcuts with a brief description.

Sub-Menu Item	Toolbar shortcut tool	Description
Select		Most of the time this tool is selected by default. When this tool is selected, clicking a displayed box in the display area will select that box, and the selected box will be highlighted.
Reading Order Select		When this tool is selected, dragging and dropping the mouse on the display area will show the reading order of the selected area. Try this tool and check the result.

Create		This tool is disabled most of the time. Double click on a zone type in the Labeling area to activate it. Select this tool to draw a box in the display area. The type of the box is the same as that of the zone which was double clicked to activate the Create tool.
Edit		This tool is used to change the size of a box. When this tool is selected, dragging a corner of a box will resize that box.
Move		This tool is used to change the location of a box. When it is selected, dragging the box can move it to a desired location.
Delete		When this tool is selected, clicking the box in the display area will remove that box.
Merge		This tool allows the operator to merge two boxes into a single box. When it is selected, clicking one box will select that box as the first box, clicking the second box will merge those two boxes into a single box.
Split		This tool allows the operator to split a single box into two boxes, either in a horizontal or in a vertical direction, depending on the location of the mouse cursor.
Save		Same as menu item File → Save. Saves the changed data. Selecting this menu item will pop a “Save Modified Files?” dialog box that lists all changed files that can be saved. (As shown in Figure 4-2)
Open		Same as menu item File → Open a Dictionary. Opens a new dictionary.

Table 4-2 Image Editing Tools

Note: The box type can easily be changed. The procedure to change the box type is: select the box → click the target box type on the buttons listed in the Result zones area. The change in the box type will be indicated by the change in the corresponding box color. The changes will be saved when the file is saved.

5 Using the Interface

This section will describe how the operator can use the interface to perform various operations. First, we will explain how to configure various elements of the system: Zone Colors, Scanner, etc. Then we will cover preparation of the dictionary in two cases; from scratch and using an existing dictionary. Finally, we will explain the operation of the system in detail from scanning to generation in different file formats.

5.1 Configuring the System

5.1.1 Configuring the Scanner

The operator can select the scanning application by configuring the Scanning panel in the **Bridge Preferences** dialog box (shown in Figure 5-1), which can be opened through the menu item **Config → Config Utility**.

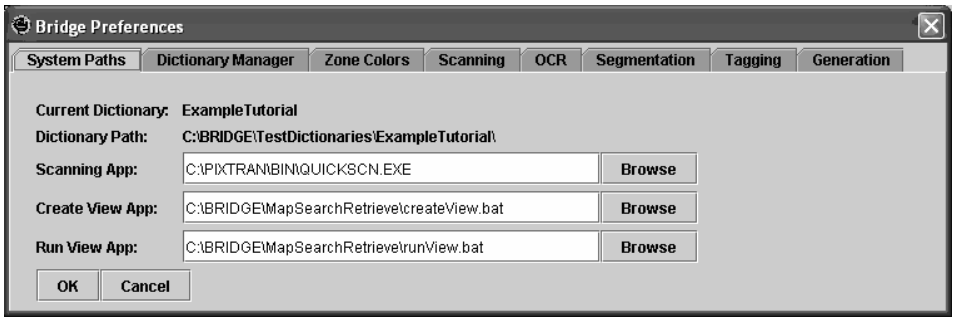


Figure 5-1 Bridge Preferences Dialog Box

5.1.2 Configuring the "Zone Colors"

In the OCR, Segmentation and Tagging modes, different colors are used to differentiate zone types. Zone Colors configuration is the only way to change the color that represents each type of zone. For each mode, click one row of the table shown on top of the panel. This will display another table at the bottom of the current panel (shown in Figure 5-2), which lists the ID, name, color and editable attributes for all the different types.

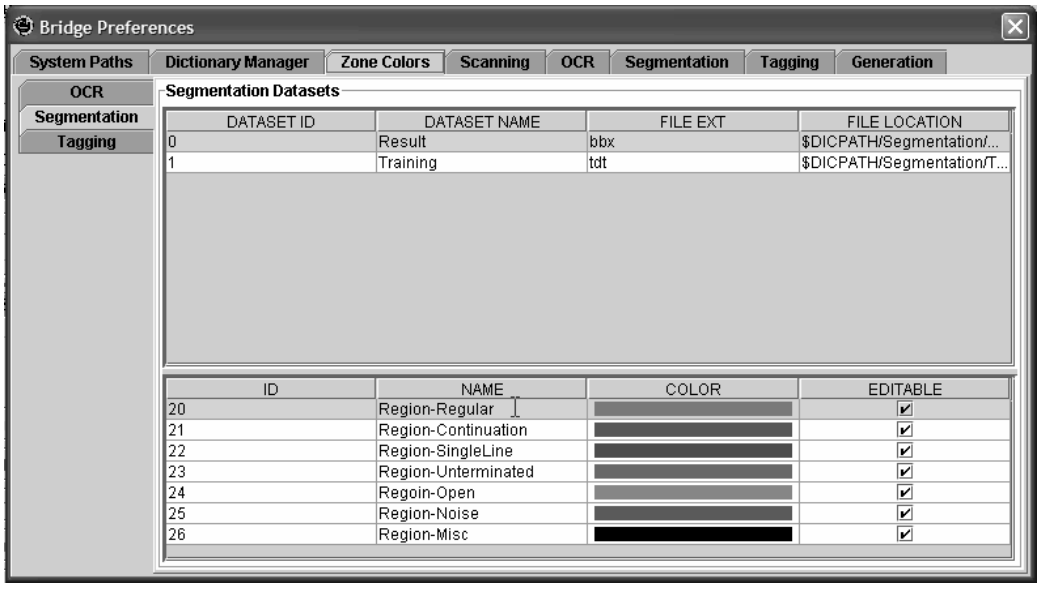


Figure 5-2 Zone Color configuration

To change the color of a zone type, left-click on the color column in that zone type. This will pop up a **Pick a Color** dialog box (shown in Figure 5-3).In this box, the operator can pick a color he desires for that zone type from the options available.

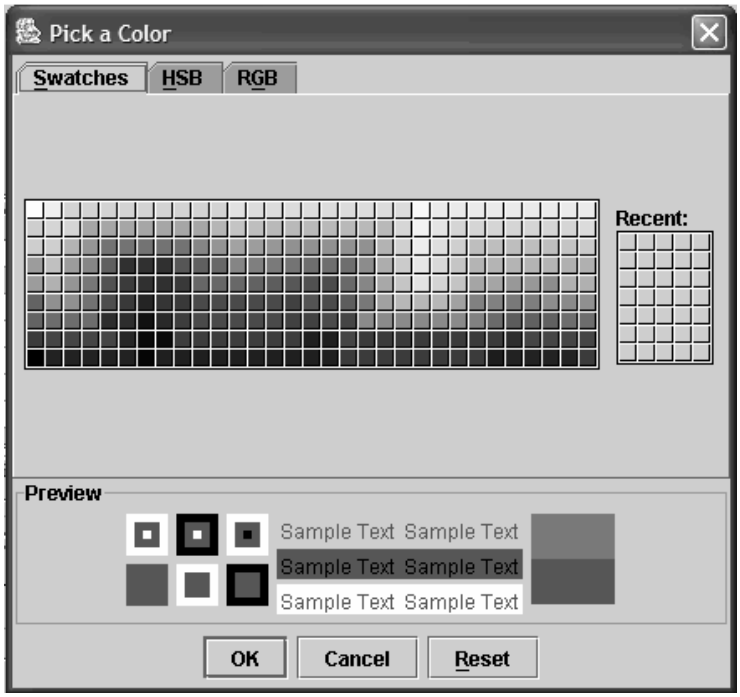


Figure 5-3 Changing color of a zone type

5.1.3 Configuring Personal Dictionaries

To facilitate spelling correction of OCR, the operator can use his/her own Personal Dictionaries. These can be set from **Config → Personal Dictionary**. The details will be described in Section 5.4 Optical Character Recognition (OCR).

5.2 Creating and manipulating dictionaries

There are two ways to create a dictionary, as a new.

5.2.1 Creating new dictionaries from scratch

Step 1: Select the menu item **Config → Config Utility** to pop up the **Bridge Preferences** dialog box.

Step 2: Select the Dictionary Manager panel then the Create New to get the text field shown in Figure 5-4.

Step 3: First select an existing directory by clicking the **Browse...** button, then type the new dictionary's name in the Dictionary Location text field. In the example shown in Figure 5-4, "C:\Bridge" is an existing directory, and "TestDict" is the new dictionary's name.

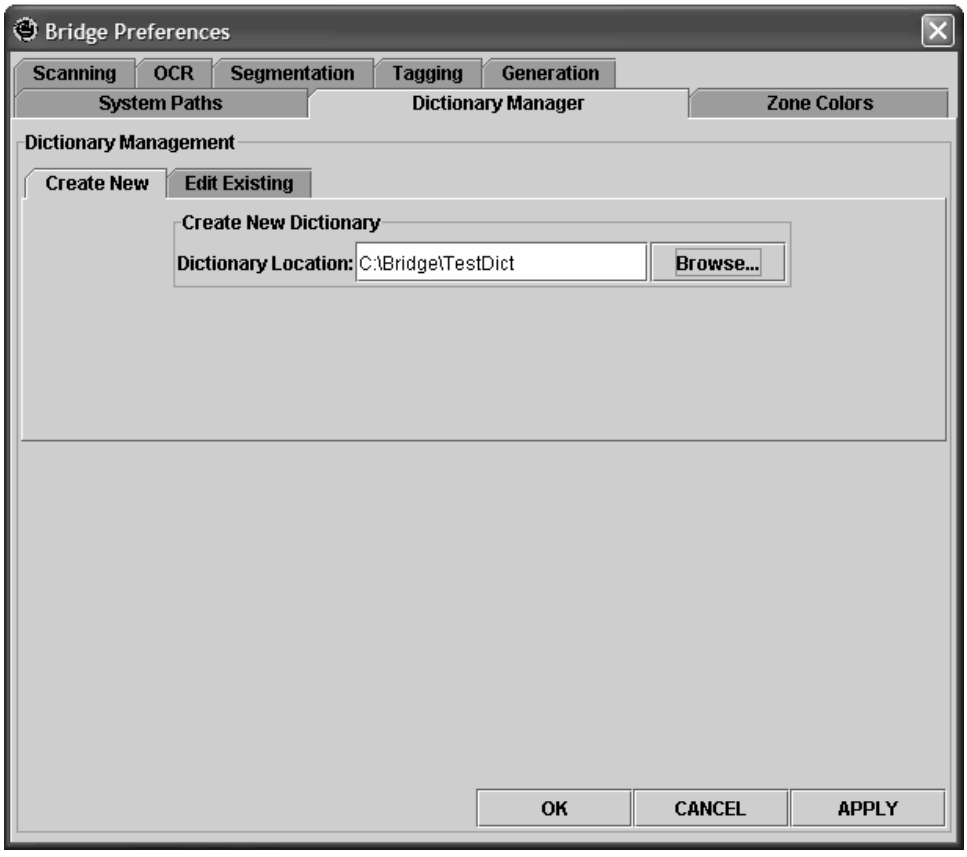


Figure 5-4 Create a dictionary from scratch

5.2.2 Manipulating Existing Dictionaries (Dictionary manager)

If the operator already has a dictionary prepared, they can create a new dictionary from this existing dictionary as described below. In this case we have the pre-existing dictionary as the source dictionary and the new dictionary as the target dictionary. There are multiple ways in which an existing dictionary can be modified. The operator can either copy the whole dictionary as is into the new dictionary, or they can copy the image files or the configuration to the destination. The operator can move the source dictionary or rename it. There is also a provision to delete an existing dictionary. The following steps outline the procedure to make changes to an existing dictionary.

Step 1: Select the menu item **Config → Config Utility** to pop up the **Bridge Preferences** dialog box.
Step 2: Select the Dictionary Manager panel and then the Edit Existing to display the panel shown in Figure 5-5. The Panel will display the operations available, the source dictionary field and the target dictionary field.
Step 3: Set the source dictionary using the Source text field and the target dictionary using the Target text field.
Step 4: Select the desired operation from the Select Operation drop-down list, where:
Copy All: duplicates the whole source dictionary.
Copy Images: copies only images to the target dictionary from the source dictionary.
Copy Configuration: copies only the configuration to the target dictionary from the source dictionary.
Move: moves the source dictionary to the target dictionary.
Rename: changes the name of the source dictionary to the name of the new dictionary.
Delete: removes the source dictionary
Verify: Verifies the structure of the dictionary.

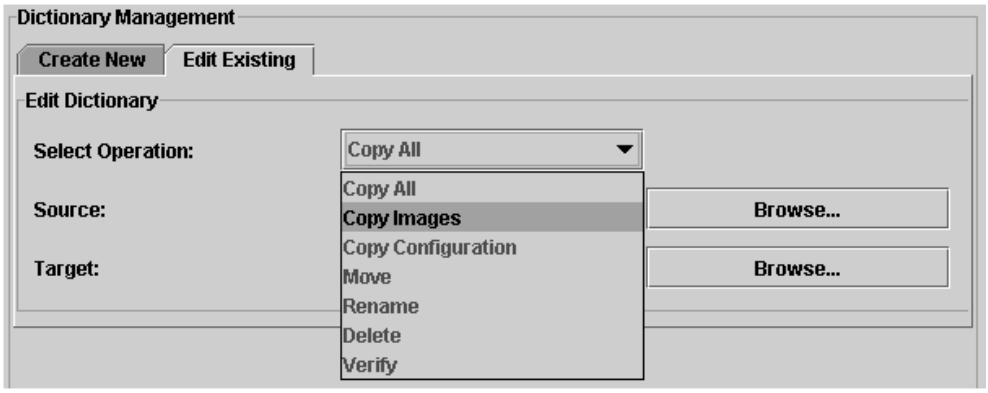


Figure 5-5 Manipulating existing dictionaries

After a dictionary is created, a file with the given dictionary name and extension ".dic" is created. For example, if the new dictionary has name "TestDict", then this dictionary file is named "TestDict.dic" and the contents of this dictionary will be put in directory "TestDict" which has the initial directory structure shown in Figure 5-6. The file TestDict.dic should be created in the same directory that is the parent directory of the dictionary TestDict. For example, if TestDict is put in C:\; then TestDict.dic should also be created in C:\.

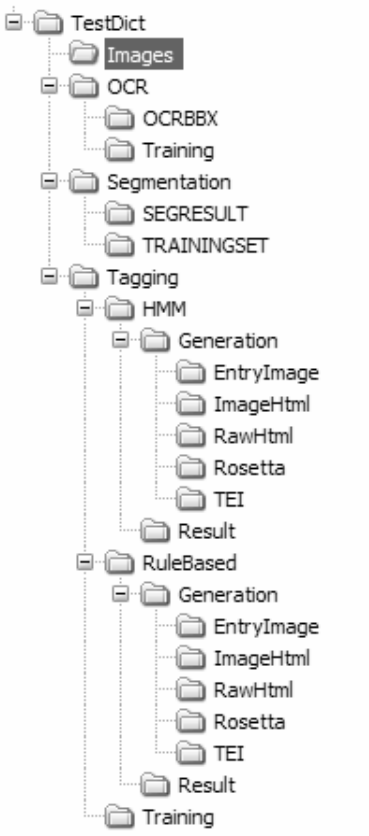


Figure 5-6 Dictionary structure

5.3 Scanning

Click the **Scan** button in the Workflow area to activate the third party scanning software PixTools (shown in Figure 5-7) to perform page scanning. The interface **QuickScan** to control the scanning is shown in Figure 5-8.

Figure 5-7 Starting PixTools

Figure 5-8 QuickScan Interface

In the QuickScan interface, the scanner set up dialog box can be activated through the menu item **Scan → Preview Settings...** (shown in Figure 5-9). In this dialog, the operator can control the scanning mode, resolution, contrast, brightness, paper size and so on (shown in Figure 5-10).

Figure 5-9 Preview settings for scanning

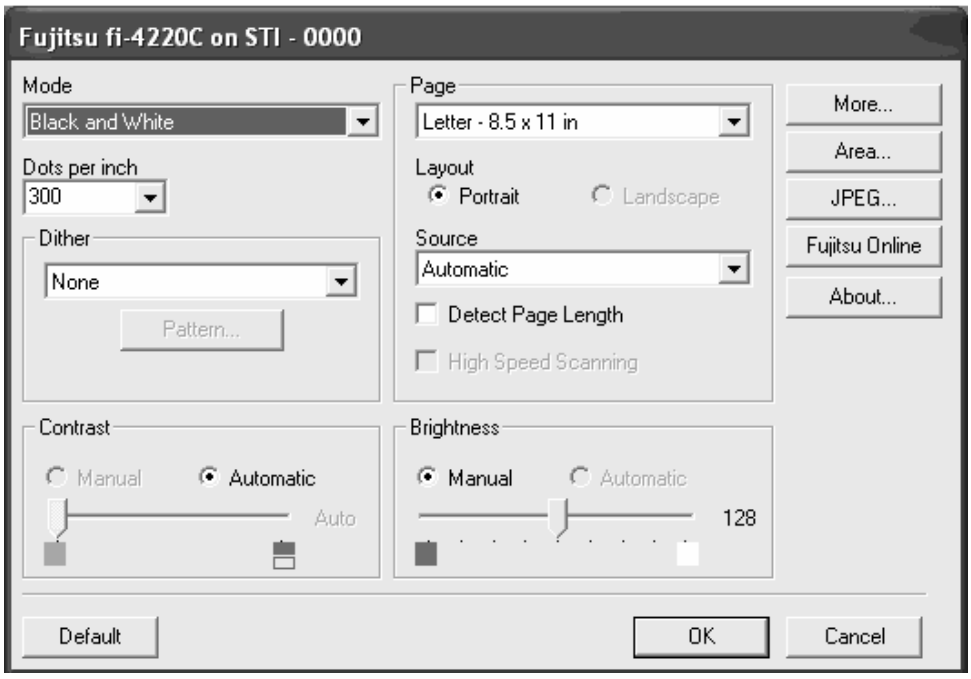


Figure 5-10 Setting scanning parameters

To avoid the inconvenience of configuring the scanning parameters each time the QuickScan interface is open, the operator can create a profile which contains the entire predefined configuration. The **Profile Editor** dialog box is shown in Figure 5-11. In this dialog, the operator can control the scanning type, resolution, image file format and so on.

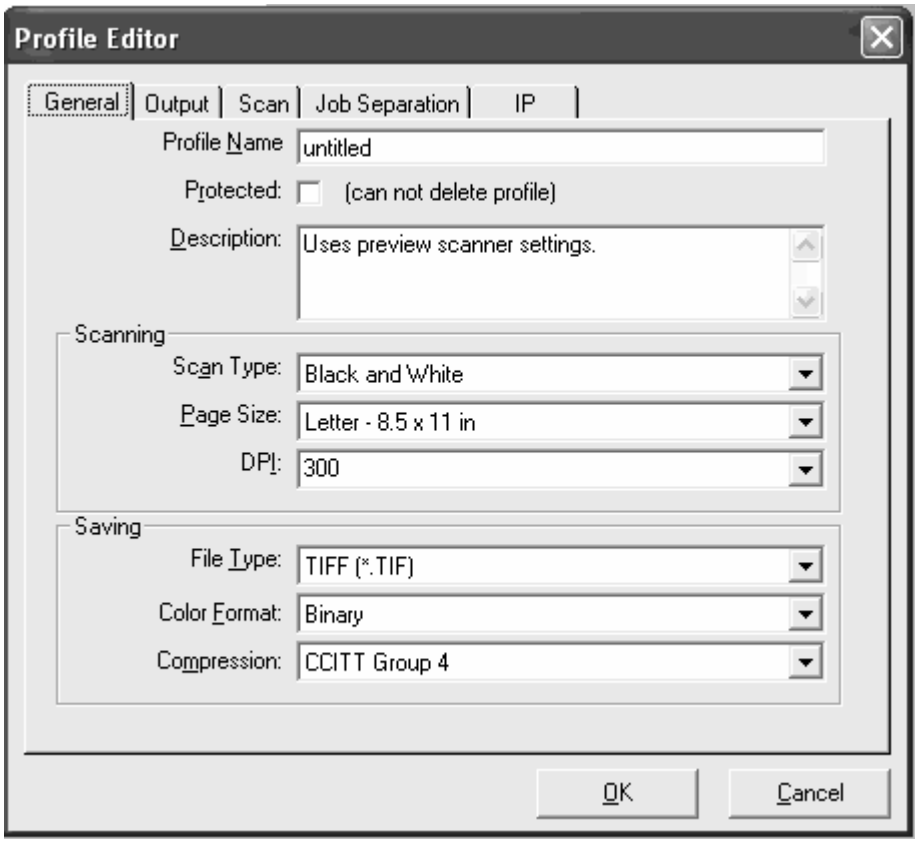


Figure 5-11 Creating user profile for a predefined configuration

5.4 Optical Character Recognition (OCR)

The OCR results are obtained using ScanSoft's OCR software Capture System 12. The OCR panel is shown in Figure 5-12. The panel is divided into four main areas, and the description for each area is as follows.

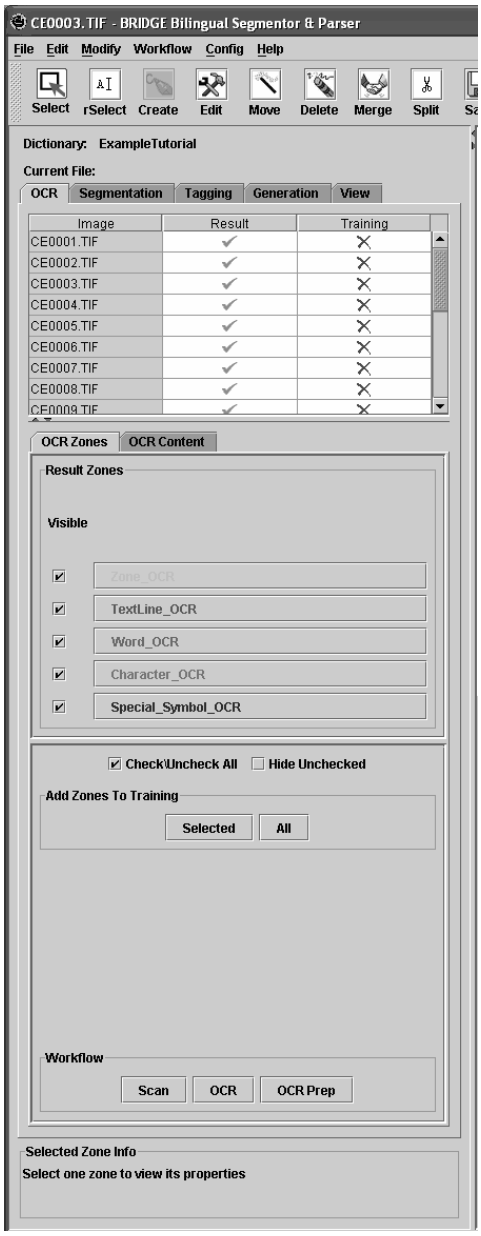


Figure 5-12 OCR Panel and workflow

5.4.1 Description of the Panel

Image Access Area

At the top of this OCR panel is a table, which lists all the pages of the current open dictionary. The first column Image shows the filename of each page. The Result column shows whether or not the OCR result [1] is available. The cross mark ✕ shows that the result is unavailable; while the check mark ✓ shows that the result is available. The Training column shows whether or not the current page contains some training samples; here the cross mark and the check mark have the same meaning as in the Result column. In Figure 5-12, for image CE0001.TIF the result is available and the training output is not available. To display an image, click in either the Result or the Training column for that image. While in the Result mode, the operator can mark various zones in the image using the Labeling area. These zones can be added to train the system. The operator can also select the resolution of the image by using the Image Scaling tool or the zoom in, zoom out tools on the toolbar.

Labeling Area

The sub-panel for Labeling can be switched between two views as per the requirement. When in the OCR Zones view mode, it will display the planar classification of the image for analysis. If the check box before a zone type is checked, this type of zone will be displayed on the image. Check/uncheck the different zone types and see the change in the display of the boxes that divide the image into different sets. Currently there are five types defined Zone, Textline, Word, Character and Special Symbol. The first four types of zones have a hierarchical structure shown in Figure 5-13, while the Special Symbol is at the same level as the Character.

[1] OCR Box Information

A ".ocr" file generated from the original OCR result, which organizes the whole page into a hierarchical structure. From this file, the system can get the type, coordinates, content of one zone. These files are stored in directory ".\OCR\OCRBBX".

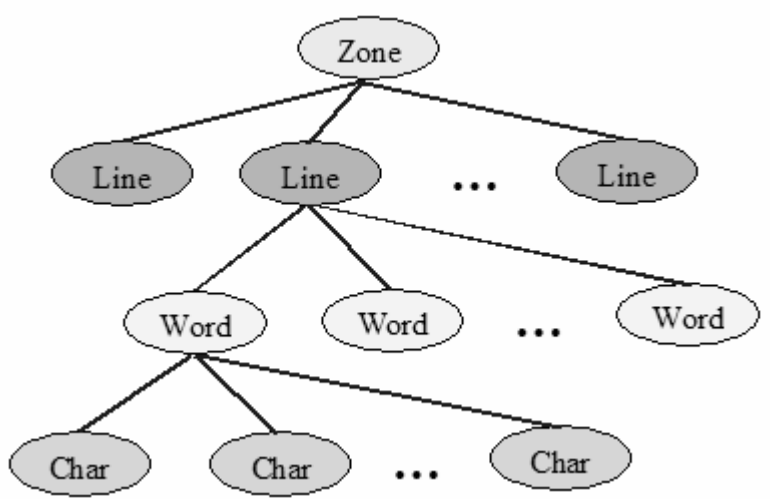


Figure 5-13 Different zone types used for classification

The OCR Content view of the Labeling sub-panel is used for OCR correction which includes a spell check. The interface for the correction operation is shown in Figure 5-14. The interface displays the textual extract of the current page image in the text pane. The correction operation is explained in detail in a later section.

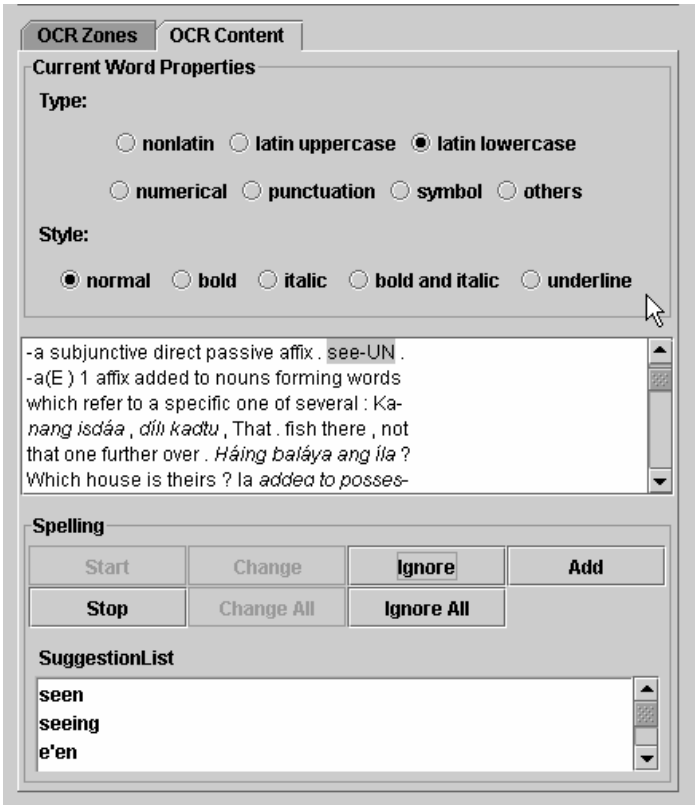


Figure 5-14 OCR Content to be used for OCR Correction

Workflow Area

Figure 5-15 OCR Workflow area

The workflow area sub-panel (Figure 5-15) has the option to use the zones marked on the image to train the system. Depending on which zones to use for training, the operator can choose to either select some zones and click the **Selected** button to add those zones or they can choose to add all zones by clicking the **All** button. The operator can also decide to view all zones or no zones in the image. This option is provided by the Check/Uncheck All checkbox: This will be reflected in the Labeling/Result Zones area. To view only certain types of zones, the operator should check them in the Result Zones area and then check the Hide Unchecked checkbox in Workflow. This will show only the selected zones.

Add Zones to Training: Not useful for current version, this will be implemented in the future.

This sub-panel also has the buttons for Scanning and OCR.

Scan Button: Its operation is explained in the Scanning, Section 5.3.

OCR Button: The operation of the OCR button is described in Performing OCR on the Image, Section 5.4.2.

OCR Prep: The operation of this button is explained in OCR Prep, Section 5.4.3.

Information Panel
This area is used to provide the operator with detailed information of the selected zones such as zone type and content, etc.

5.4.2 Performing OCR on the Image

Clicking the **OCR** button in the Workflow area will pop up the **Perform OCR** dialog box (shown in Figure 5-16). In this dialog box, the operator can select the set of images to perform OCR on by setting the start and end file fields. Another useful operation is the selection of language. Checking the language radio button will provide the operator with the list of languages that ScanSoft's Capture System V12 supports for the OCR operation. Appropriate language selection can improve the OCR performance, so it is very important. Since we always assume that one language is English in a bilingual document, what the operator needs to do is select only the second language (if any) in this list.

Question: What if there is no support for the language I want to perform OCR on, for example, Cebuano?
Answer: If you can not find the desired language in this list, try to select the language that has the similar alphabets with the desired language. For example, since Cebuano contains an alphabet with accents, selecting the language "Catalan" or "Czech" can give more accurate result than just selecting "English".

Figure 5-16 Clicking the OCR button pops up this Perform OCR dialog box

The **Script** radio button in the **Perform OCR** dialog box is reserved for future use.

5.4.3 OCR Prep

There is an **OCR Prep** button in the Workflow area, one click on this button will pop up the **OCR Organizer** dialog (shown in Figure 5-17). In this dialog, the operator can choose the range of pages to perform operations on by choosing the starting file in the start file field and the ending file in the end file field.

Figure 5-17 Clicking the OCR Prep brings up this OCR Organizer dialog

Since the raw OCR results stored in the .let files have only zone and character information, the goal of OCR Organizer is to organize the recognized raw results into a hierarchical structure, i.e. convert a .let file into a .ocr file. The converted .ocr file has a pyramid structure shown in Figure 5-18. The two main operations in this part are (i) Organize characters into words; and (ii) Organize words into text lines.

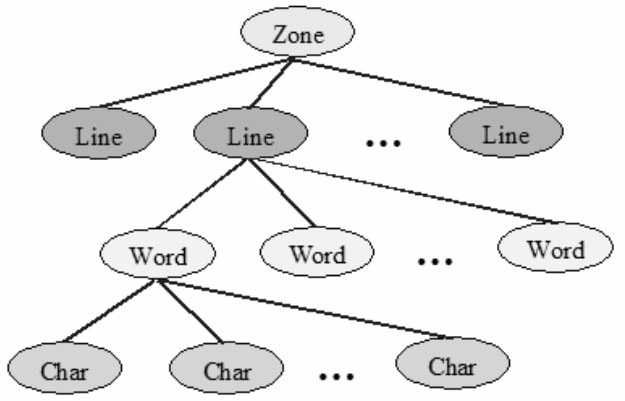


Figure 5-18 Classification of zones

5.4.4 OCR Correction

The main function of the OCR Content panel is the spell check of the OCR result. The extracted electronic text content will be displayed in the text field on the left panel (shown in Figure 5-19).

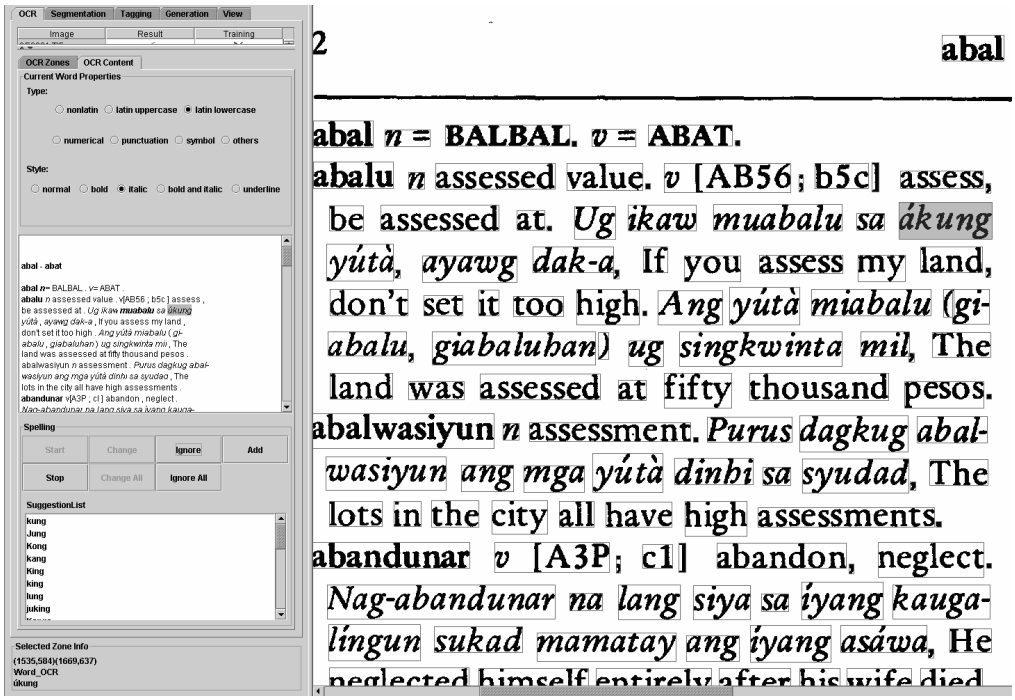


Figure 5-19 Performing OCR correction and spelling check

Check the word content and properties

The operator can select a word either by clicking on the word in the text field on the left panel or clicking the word box displayed on the right hand side display area. The selected word will be highlighted in the displayed image, and the properties (word style and word type) of this word will be shown in the Current Word Properties. For example, the selected word ákung is highlighted, and the properties show that it is a Latin lowercase word with an *italic* style.

Spell check of OCR content

Once in the OCR Content mode, the operator can perform a spell check on the OCR content by clicking the **Start** button under the text field. For a probable incorrect word i.e. if the word is not found in the dictionary, the operator has the options of ignoring the word or of adding the word to the Personal Dictionary for future use similar to other spell check tools. A list of possible correction words is also displayed in the Suggestion List field. The operator can choose to replace the incorrect word with one of the suggestions.

5.4.5 Using a personal dictionary (ies) for spell check

Setting up Personal Dictionary (ies)

Go to the menu item **Config → Personal Dictionary** to pop up the **Personal SpellChecking Dictionaries** dialog box (shown in Figure 5-20).

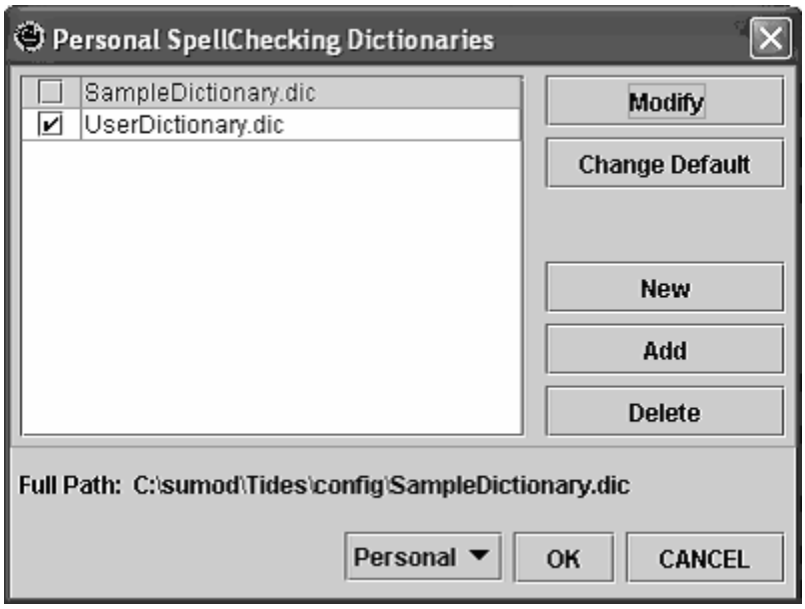


Figure 5-20 Setting up personal dictionaries for spelling check

For better understanding, one SampleDictionary.dic is shown selected. The buttons on the right side panel can be used for modifying the contents of a dictionary (**Modify** Button), changing the default dictionary (**Change Default** Button), creating a new dictionary (**New** Button), adding a dictionary already created (**Add** Button) and for deleting a dictionary from the selection list (**Delete** Button).

The selection is confirmed when the operator presses the **OK** Button. The option Personal and Custom is provided for future use.[2]

Editing a Personal Dictionary

To modify the contents of a personal dictionary, select it from the list on the left hand side and click the **Modify** Button. For better understanding SampleDictionary.dic is selected (shown in Figure 5-21).

[2] **Note:** It is important and is advised to specify the **.dic** extension when creating a new dictionary and to create new dictionaries in the **./config** directory. While adding a dictionary, the current selection option is for **.dic** extension.

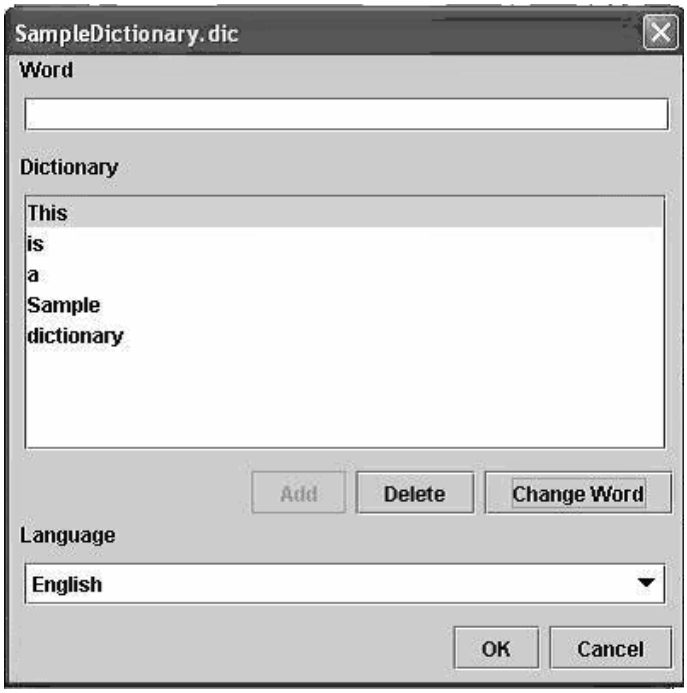


Figure 5-21 Editing a personal dictionary

To edit this dictionary, control buttons **Add**, **Delete** and **Change Word** are provided that can be used for adding a new word, deleting an existing word or modifying a word from the present dictionary.

Presently English is the only language.

No changes are saved to the personal dictionary files until the operator confirms them by clicking the **OK** Button.

5.5 Segmentation

The purpose of segmentation is to break the contents of a page into entries based on the training samples defined by the operator. Segmentation is also based on the OCR result of each page. In order for the segmentation program to run smoothly, the OCR result and the training samples must be available before the segmentation begins.

5.5.1 Description of the Panel

Creating training samples for segmentation and the actual segmentation are both performed through the Segmentation panel which is shown in Figure 5-22.

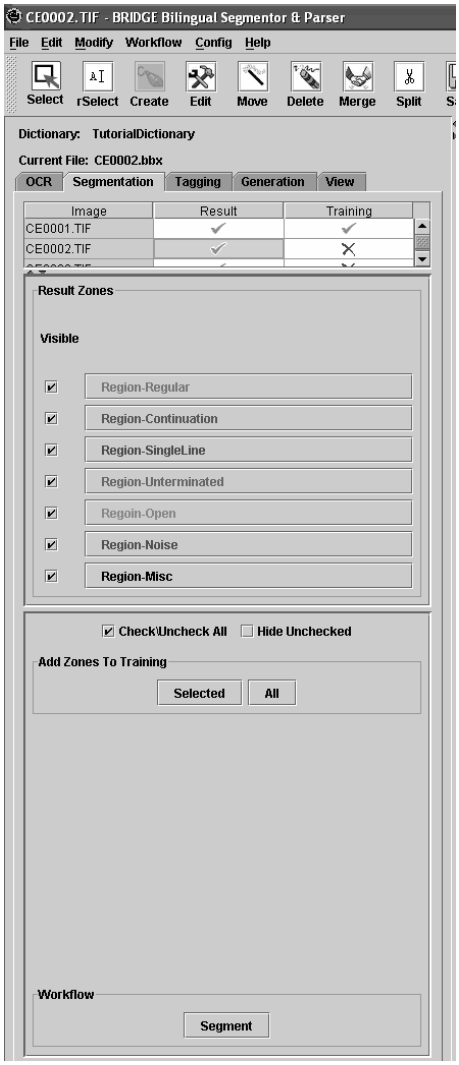


Figure 5-22 Segmentation panel and workflow

Image Access
At the top of the Segmentation panel is a table, which lists all of the pages of the current open dictionary. The first column Image shows the filename of each page. The Result column shows whether or not the segmentation result[3] is available. The cross mark ✗ shows that the result is unavailable; while the check mark ✓ shows that the result is available. The Training column shows whether the current page contains some training samples[4] or not. The cross mark and the check mark have the same meaning as those in the Result column.

Labeling
Below the table is the Labeling sub-panel that contains the Result zones area. It is a group of buttons with a check box in each one of them. These buttons list all the types of entries (regions) that can be recognized by the segmentation program, and the definitions of these entry (region) types are as follows:

Regular: A complete entry that starts and ends in the same column.
Continuation: An entry that is the continuation of an entry from the previous column or page.
SingleLine: A regular entry that contains only one single text line.
Unterminated: An entry that is not ended in one column (or page) and has a continuation part in the next column (or page).
Open: An entry that is the continuation part of an entry in the previous column (or page) and does not end in the current column (or page).
Noise: An entry that should not be taken into account in the parsing of the page, such as the page number, the indexing words that often appear on top of each page.
Misc (Miscellaneous): All entry types that don't belong to any of the above types (not used currently).

[3] Segmentation result
The text file with extension ".bbx" which stores the number of segmented entries, and the type and bounding box of each entry. The first line of this file is always a single number which represents the total number of segmented entries. Each following line is an entry with the format "type left top right bottom".

[4] Training Sample for segmentation
The text file with extension ".tdt" which stores the number of training entries, and the type and bounding box of each entry. The first line of this file is always a single number which represents the total number of training entries. Each following line is an entry with the format "type left top right bottom".

If the result is available, the check box in front of each button can be used to display or not to display a specific type of entry. The other two check boxes under these buttons have the following functions:

Workflow
The workflow sub-panel contains the following controls:

Check/Uncheck All: When this box is checked, all the check boxes above will be checked and vice versa. This control can be used to decide whether to display all the zone types or to display no zones in the image.

Hide Unchecked: If there are unchecked boxes in the above mentioned boxes, checking this box will hide all the unchecked boxes.

Add Zones To Training: The two buttons Selected and All in this area are used for preparing the training samples. Suppose for one page, the segmentation result is available. If the operator clicks the All button, all the available entries for this page will be added to the training sample set. If the operator selects some entries in this page, pressing the button Selected will only add the selected entries to the training sample set. How to select one or multiple entries will be described in the following section.

Workflow: The button **Segment** in this area is used for performing the segmentation which will be described in details in the following section.

5.5.2 Preparing Training Samples for Segmentation

There are three ways to prepare the training samples[4] for segmentation, they are described below.

Working in the result display mode when the result is unavailable
When the segmentation result is unavailable, working in the result display mode to prepare the training samples requires the operator to draw some entries on the displayed page and then add them to the training samples. The procedure is:
1. Click the desired page in the Result column to display the page image in the result display mode.
2. Double click the button with the desired entry type, the disabled **Create** button 

on the tool bar will be activated after the double clicking.

3. Press the **Create** button  on the toolbar, then draw the desired entry box on the displayed image.
4. Each time the operator wants to draw a new entry with a different type, step (2) and (3) should be repeated.

5. After drawing all desired entry boxes, press the **Select** button 

on the tool bar. This will enable the operator to select a drawn entry and the selected entry will be highlighted using the color defined for this type.
6. Add the selected entry to the training set by clicking the **Selected** button in the Add Zones To Training area.
7. With the SHIFT key on the keyboard pressed, the operator can select multiple entries at one time.

Working in the training display mode when the result is unavailable
When the segmentation result is unavailable, working in the training display mode to prepare the training samples takes less time than working in the result display mode, especially when the OCR box information[1] is available.

If the OCR box information is unavailable, the first 4 steps to prepare the training samples are similar to working in the result display mode except the first step which requires the operator to click the desired page in the Training column to display the page image in the training display mode. After drawing all the entries, the operator can click the **Save** button 

on the toolbar and then click the **OK** button on the pop up **Save Modified Files?** dialog.

If the OCR box information is available, it is easier to create the training samples.

The procedure is:
1. Check the show words box in the Ghost Zones area to display the word boxes.
2. Click the **Select** button on the toolbar to get into the Select mode.
3. With the left key down, use the mouse to select all the words in a desired entry. The selected words will be highlighted.
4. Click the button in the Training zones area to assign the selected entry to the type of that button.
5. Press the **Save** button 

to save all the drawn entries into the training set.

Working in the result display mode when the result is available
This is the simplest way to prepare the training samples for segmentation. The operator can select one or multiple entries to add to the training set. The procedure is:
1. Click the desired page in the Result column to display the page image in the result display mode. There should be some entry boxes displayed on the images.

2. Press the **Select** button 

on the toolbar to get into the Select mode.
3. Select the entries and press the **Selected** button in the Add Zones To Training area to add them to the training set.
4. If the operator wants to add all the entries into the training set, they can press the **All** button in the Add Zones To Training area without selecting individual entries.

5.5.3 Configuring Segmentation

After creating the training samples for the segmentation, the system is ready to perform the segmentation. To activate the segmentation the operator clicks on the **Segment** button in the Workflow area. There will be a **Dictionary Segmentor** dialog pop up (Figure 5-23). In this dialog, the operator can select the pages he wants to perform the segmentation on.

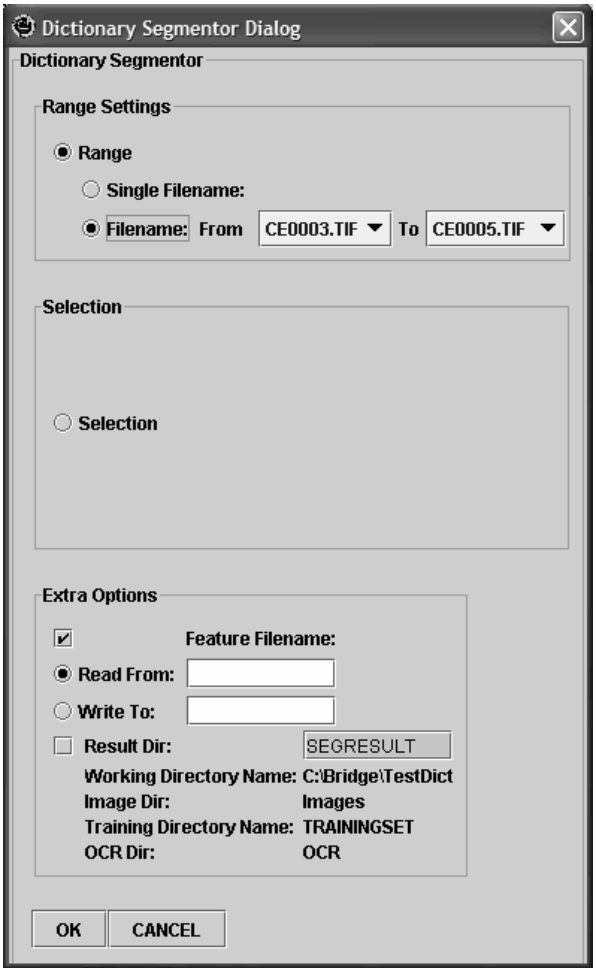


Figure 5-23 Setting parameters for segmentation

Select a single page
1. Select the Range and then the Single Filename radio button (shown in Figure 5-24).
2. Select the page file in the drop down list.

Figure 5-24 Selecting a single page to perform segmentation on

Select pages in a range
1. Select the Range and then the Filename radio button (shown in Figure 5-24).
2. In the From drop down list, select the starting filename.
3. In the To drop down list, select the stop filename, all the files in this range are selected.

Select a specific collection of pages
1. Check the Selection radio button in the Selection area (shown in Figure 5-25).
2. In the left list, click the file you want to select.
3. Press the “>” button to add this file to the selection set (the right list).
4. Repeat step (2) and (3) until all desired files have been selected.

The functions of the other three buttons are:
">>" : adds all the files in the left list to the selection set.
"<" : removes the selected file from the selection set (the right list).
"<<" : removes all files from the selection set (the right list).

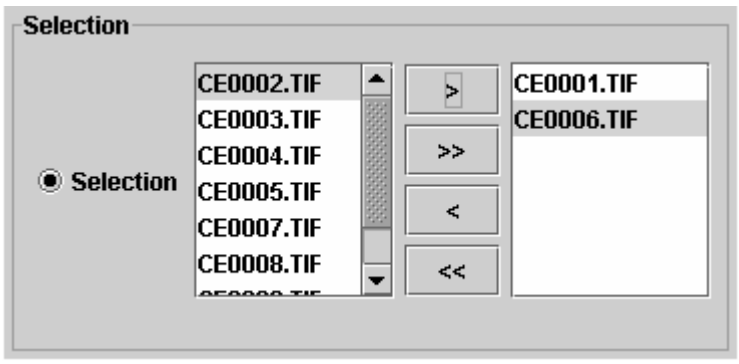


Figure 5-25 Selecting specific files to perform segmentation on

Feature Filename

If the operator does not specify any operation to the feature file, the learned features will be put into a default feature file with the name “Feature.ini". However, if the operator checks the Feature Filename check box in the Extra Options area, there will be two text fields pop ups. One is Read From and the other is Write To. If the operator wants to put the learned features into a specific file, they can select the Write To radio button and input the file name in the text field. If the learned features have already been saved into a file and the operator wants to perform the segmentation based on those features, they can select the Read From radio button and type in the already existing feature file name. In this case, the system will not be retrained and the segmentation is based on already learned features. For example, the configuration shown in Figure 5-26 means to read features from an already existing feature file name "Features.ini".

Figure 5-26 Using segmentation parameters from a feature file

Result Dir

The text field of Result Dir shows where the segmentation results will be put. If the operator wants to change the location to put the segmentation results, they can check the Result Dir check box first then type in the new location (a directory name) in the editing field.

Other Information

There is other uneditable information shown in this dialog, which includes Working Directory Name, Image Dir, Training Directory Name and OCR Dir (see Figure 5-23).

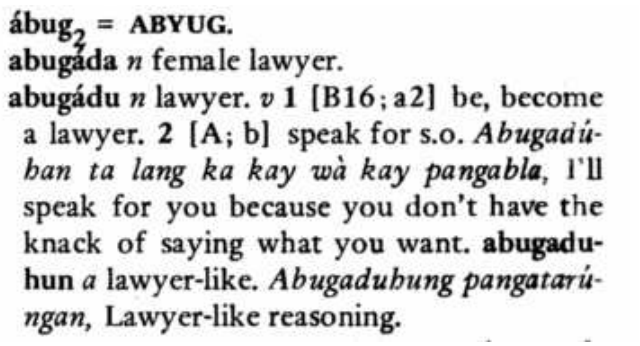
5.5.4 Activating Segmentation

After the entire configuration described above is complete, press the **OK** button to perform the segmentation based on the operator's configuration. When the segmentation is finished, the table in the left panel should show the results for the selected files that are available. If not, choose the menu item **File → Refresh** to update the display.

5.6 Tagging

The goal of the Tagging process is to identify and tag the information types, such as headword, translation, part-of-speech (POS), pronunciation, in a dictionary entry, so that resources can be generated. In order to run tagging, segmentation results must be available.

Each dictionary provides a different set of information types, and to identify these information types, different font styles, various separators are used. The keywords may also provide an explicit interpretation of the information type.



ábug$_2$ = **ABYUG.**
abugáda *n* female lawyer.
abugádu *n* lawyer. *v* **1** [B16; a2] be, become a lawyer. **2** [A; b] speak for s.o. *Abugadúhan ta lang ka kay wà kay pangabla*, I'll speak for you because you don't have the knack of saying what you want. **abugaduhun** *a* lawyer-like. *Abugadubung pangatarúngan*, Lawyer-like reasoning.

Figure 5-27 A sample dictionary section

In the above dictionary sample (Figure 5-27), a **boldface** font indicates headwords and derived words, while a normal font indicates translations, and *italic* font indicates examples of usage. Part-of-speech (POS) information is provided using the keywords *a, n, v,* etc. in *italic* font. Synonyms are preceded by equal to sign, and translations are followed by period. A numbering system is used to identify translations with different POS.

The tagging process uses features font style, the separators, and the keywords. The operator is required to provide these features in the "Configuration File". In the following section, how these features can be provided to the system is explained step by step.

5.6.1 Preparing the configuration file

This is done by the menu item **Config → Config Utility**. Clicking this menu pops up the **Bridge Preferences** dialog. In that window, click the Tagging panel. This panel enables the operator to select which categories are present in the dictionary the operator is working on, and the features of these categories (shown in Figure 5-28).

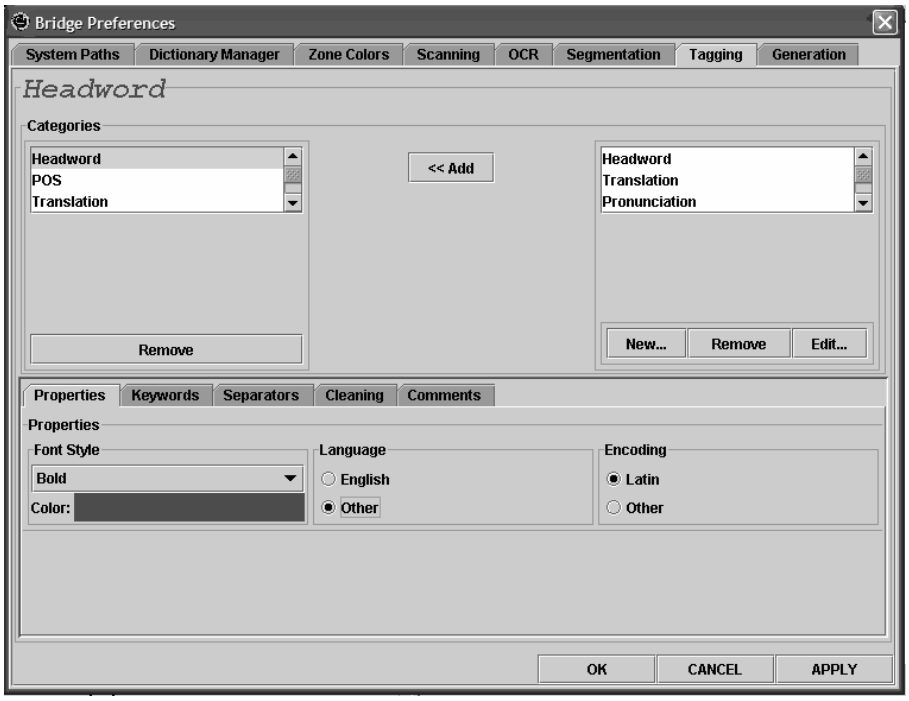


Figure 5-28 Preparing the configuration file for tagging

The operator can use the **Apply** button anytime he wants to save the entered information. The **OK** button saves all the information entered so far and closes the **Bridge Preferences** dialog.

On the top left corner of the panel, the category currently selected in the Categories is displayed in red (Headword in Figure 5-28). There are two lists in the Categories area, the one on the left shows the categories of the current dictionary the operator is working on. The one on the right consists of possible category names. In order to add categories to the left list, select a category from the right list, and click on the **Add** button. The operator can define new categories, edit or remove existing ones from the right list, and remove added categories from the left list.

The categories that appear in the right list are as follows:

Headword: The main word that defines the entry.
Translation: The translation of the headword.
Pronunciation: The representation that shows the way a word is spoken, using phonetic symbols.

POS: Part-of-speech, a classification of words according to their functions in the context.
Domain: The region a word or translation is used in.
Gender: A grammatical category used in the classification of words. Possible values are masculine, feminine, and neutral.
Number: Grammatical number (singular or plural).
Context: The framework or perspective a word or translation is used in.
Language: The original language of a word.
Alternative_Spelling: Another spelling of the word.
Compound/Derived: The word that is lexically related to the headword.
Compound/Derived_Translation: The translation of the compound/derived word.
Example: A phrase or a sentence showing how the word is used.
Example_Translation: The translation of the example.
Idiom: An idiom including the word.
Idiom_Translation: The translation of the idiom.
Cross_Reference: A reference made to another entry in the dictionary.
Antonym: The antonym of the word.
Synonym: The synonym of the word.
Tense: The grammatical tense (past, present, future etc.) associated with a given inflected form.
Person: the "grammatical" person (1st, 2nd, 3rd, etc.) associated with a given inflected form.
Mood: Information about the grammatical mood of verbs, such as indicative, subjunctive, imperative.
Explanation: Additional information provided.
Case: A form of a word which indicates its relation to other words.
Usage: The way in which the word is actually used.
Subcategorization: A further classification of words, such as syntactic patterns for verbs, count/mass distinctions for nouns, etc.
SenseNumber: The numerals that are used to indicate different POS or different translations of a word.
Subject: A branch of knowledge the word is used in or the word indicates.
Collocation: A collocate of the headword.
Abbreviation: A shortened form of a word or phrase used mainly in writing to represent the complete form.
Misc. (Miscellaneous): Reserved for the words that the tagging process could not find a category for.
Separator: Shows the punctuation.

It is worth noting that the two categories **Misc.** and **Separator** are added by default to the list on the left, and these should not be removed from the left list to ensure that the tagging process works correctly.

When a category name is selected from the left list, the operator can define its features. This is done by using the bottom part of this dialog, which contains 5 subpanels Properties, Keywords, Separators, Cleaning, and Comments.

The “Properties" Panel
In the Properties panel (Figure 5-28), the operator can set the following properties:
1. The font of the category used in the dictionary.
2. The language (English or non-English) of the category.
3. The script (Latin or non-Latin) of the category.
4. The color in which this category will be displayed on some of the outputs.

The "Keywords" Panel
In most of the dictionaries some of the information is displayed by the abbreviations so that less space is used. These usually include POS (noun, verb, etc.), gender (masculine, feminine, neutral), domain (physics, medicine, etc.), etc. A list of the abbreviations used in a dictionary is usually given in the preface of the dictionary. The full forms of the most commonly used abbreviations are already included in the right list in the Keywords subpanel (Figure 5-29).

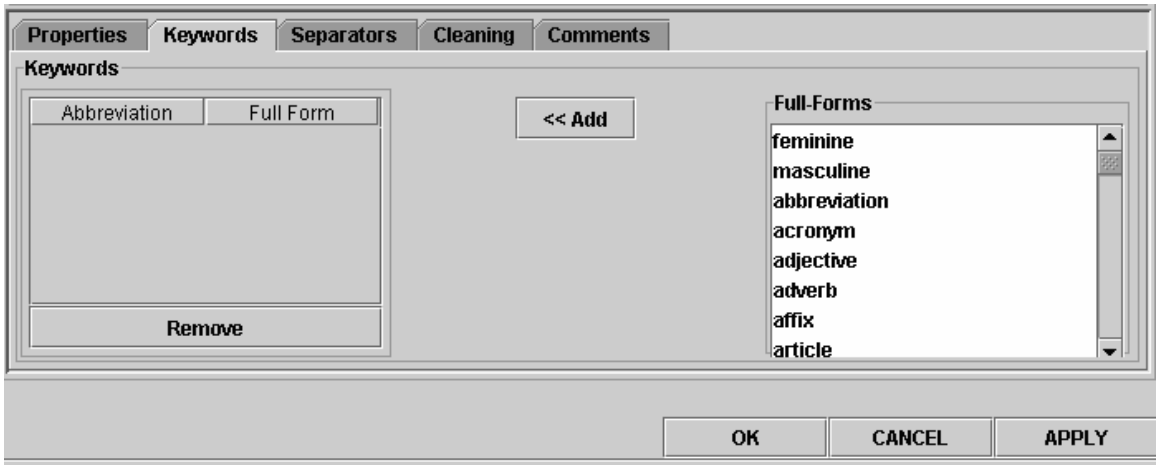


Figure 5-29 The Keywords panel

When the operator selects one of the full forms on the right list, and clicks on the **Add** button, an **Edit Dialog** box (shown in Figure 5-30) appears asking the operator to enter the abbreviation used in the current dictionary for this full form. If a period is always used after the abbreviations in the dictionary, it is useful to add the period to the abbreviated form as well. All POS abbreviations, for instance, should be entered when the current category is POS (i.e. the red text on the top left corner is POS).

Figure 5-30 Entering abbreviation for a keyword

The "Separators" Panel

Separators are usually punctuations used in the dictionaries to separate different categories. The “Separators” subpanel (shown in Figure 5-31) allows the operator to describe different kinds of separators and their functions. Five operators that are used during the tagging process (PreviousEndsWith, StartsWith, EndsWith, InPlaceOf, and Contains) and the most commonly used separators in the dictionaries are given in the right hand side list. The left hand side list shows the information for the current dictionary and current category as usual.

- **PreviousEndsWith** is reserved for the separator used at the end of the previous word. In Figure 5-27, a period is used at the end of the word before the example of usage. This is indicated by “Example of Usage PreviousEndsWith”.
- **StartsWith** operator is for the separators that appear at the beginning of a category. For instance in Figure 5-27, the verb subcategorization category begins with a left square bracket, “[“. This information can be identified by the operator as “Subcategorization StartsWith [“.
- **EndsWith** operator is for the separators that appear at the end of a category. In Figure 5-27, the translation ends with a period. This is given to the tagging process as “Translation EndsWith”.
- **InPlaceOf** is reserved for the shortcuts used primarily for headwords, such as using **"~"**, for the purposes of using less space in a dictionary. If “~” is used as a shortcut for the headword, this is indicated as “Headword InPlaceOf ~”.
- **Contains** should only be used if some identifying separator is used in the middle of a category. Using the other operators gives more specific information.

“number” represents numbers used especially to divide different POS of a word, such "1", "2", etc. “numberletter” is similar, but in the format "1a", "1b", etc. These can be used just like the other separators, such as for the category “Translation PreviousEndsWith number**”**.

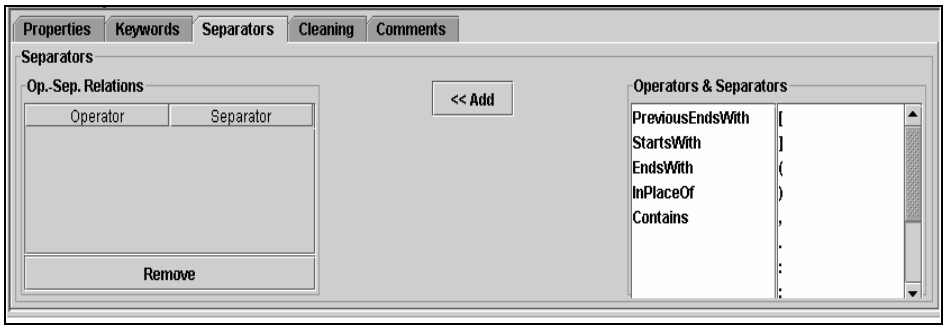


Figure 5-31 The Separators panel

For any category, there is no limit to the number of separators the system operator can identify. The same operator (with different separators) can be used more than once with the same category as well.

In most of the dictionaries, same separators are used for all. In this case, the operator has to add these categories (Translation, Compound/Derived_Translation, Example_Translation) to the list in Figure 5-28, and describe the separators for the Translation, but there is no need to repeat all the separators in the other categories.

In order to give the tagging process these separators, the system operator selects the operator and the separator from the list on the right, and clicks the **Add** button. For instance, to enter the information "Translation EndsWith", the system operator selects the operator EndsWith and the separator period (.) from the list on the right and clicks the **Add** button when the current category is Translation.

The "Cleaning" Panel

This subpanel (shown in Figure 5-32) allows the operator to define the characters that they do not want to be included in the final output. There are two special cases:

- "number" removes all the digits.
- "all punc" removes all punctuations.

Figure 5-32 The Cleaning panel

The "Comments" Panel

This subpanel is reserved for future usage.

5.6.2 Performing Tagging

After the segmentation results are available and the configuration file is ready, the tag process can be called by clicking the menu item **Processing → Tagging → Tag** or by pressing the **Tag** button in the Workflow area of the left panel of the tool (shown in Figure 5-33). This opens a dialog box (Figure 5-34) which asks the operator the range of pages they want to be tagged. The tagging process usually takes around 15-20 minutes for a 1000 paged dictionary.

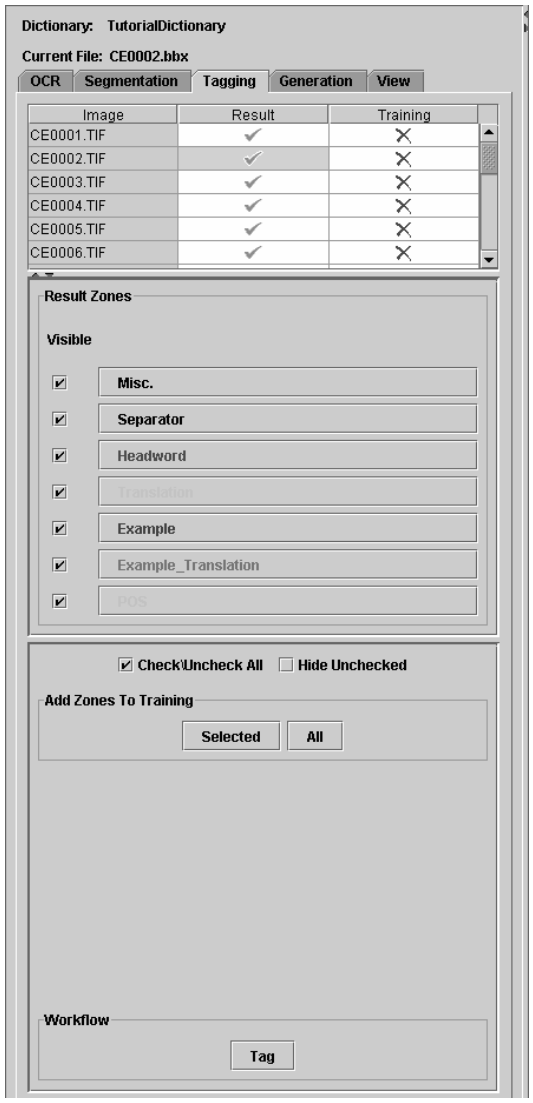


Figure 5-33 Performing tagging

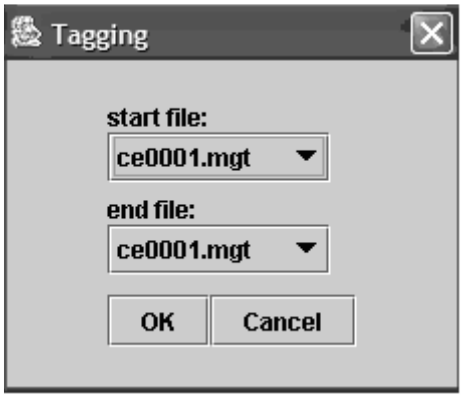


Figure 5-34 Tagging dialog box

5.6.3 Preparing Training Samples for Tagging

The purpose of preparing the training samples for tagging is evaluation, not system training. The operator can prepare the training files just like in the segmentation part. However, **Create, Move, Delete, Merge**, and **Split** operations are useless in the tagging results.

There are two ways to prepare the training samples for tagging, one is in the result display mode and one is in the training display mode, which are similar to preparing the training samples for segmentation. Preparation of the training samples for tagging differs by the presence of the Begin Bit which is used to indicate whether the word is the first word of a phrase or not. When it is on, there is a small point on the top left corner of the box around a word and it indicates that this word is the first word in a phrase. It can be changed by right clicking on the box (Figure 5-35).

In order to prepare the training samples in the result display mode, the result should be available for the desired pages. The procedure is as follows:
1. Click the desired page in the Result column to display the page image in the result display mode.
2. Press the **Select** button 

or **rSelect** button 

on the toolbar to get into the Select mode.
3. Select the entries and press the **Selected** button in the Add Zones To Training area to add them to the training set.
4. If the operator wants to add all the entries into the training set, they can press the **All** button in the Add Zones To Training area without selecting specific entries.

The procedure to prepare training data in the training display mode is:
1. Check the show words box in the Ghost Zones area to display the word boxes.
2. Click the **Select** button or the **rSelect** button on the toolbar to get into the Select mode.
3. Use the mouse to select all words in a desired entry. The selected words will be highlighted.
4. Click one of the buttons in the Training zones area to assign to the selected that button type.
5. Press the **Save** button 

to save all the drawn entries into the training set.

If the operator wants to make some corrections on the Result or Training data, they select the boxes they wants to correct in the appropriate display mode, and changes the category and/or the begin bit of the boxes (Figure 5-35).

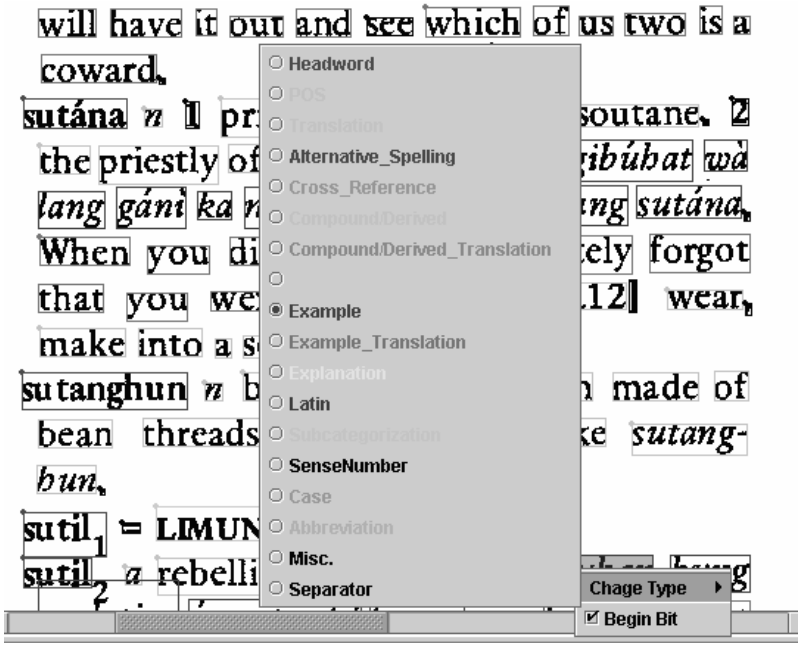


Figure 5-35 Preparing training samples

5.7 Generation

The generation module is responsible for taking the output of tagging and producing usable results. In order for the generation procedure to execute smoothly, tagging results should be available. The generation procedure can be called by pressing the menu item **Processing → Generation → Generate** or by pressing the **Generate** button in the Workflow area of the left panel of the tool (shown in Figure 5-36). This opens a dialog box (Figure 5-37) in which the operator is asked to input a dictionary ID that is required for Rosetta format. The dictionary ID must consist of 7 characters. The first 3 characters must be letters and the remaining 4 characters must be digits (e.g. ceb0001). Then, the operator selects the range of pages they want to generate, and the output formats for those pages.

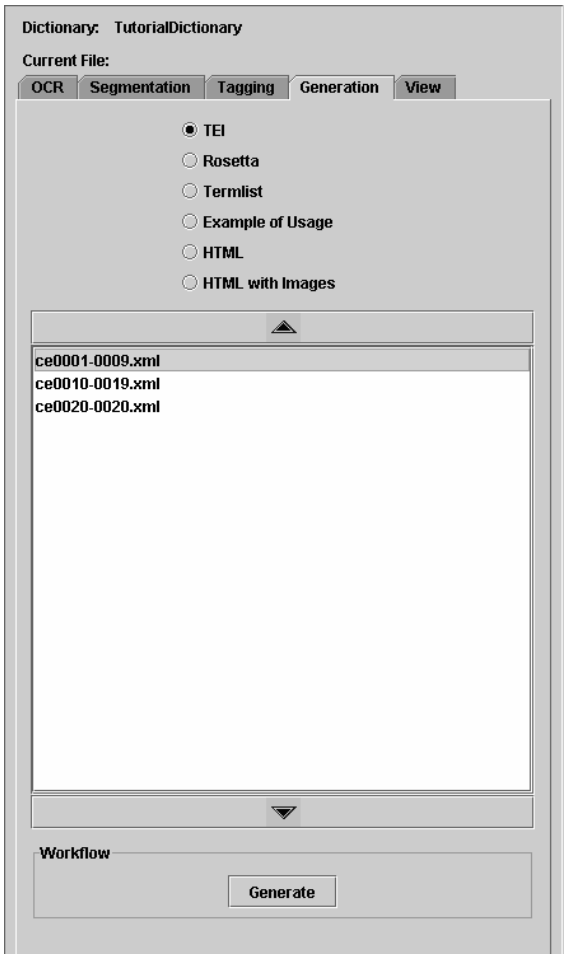


Figure 5-36 Generation panel and workflow

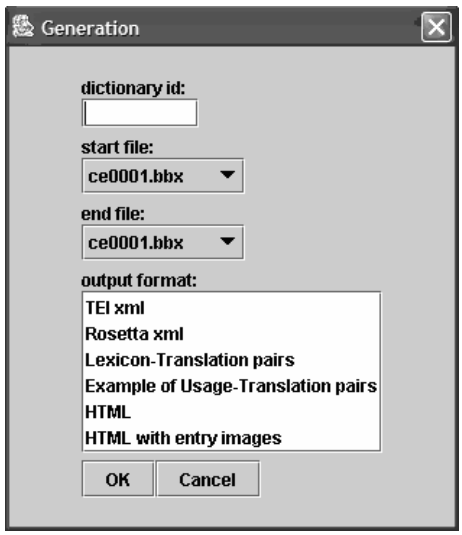


Figure 5-37 Generation dialog box

Right now, the generation procedure produces the output in 6 different formats. These are:

1. **TEI:** An XML based format. More information about this can be found at: http://www.tei-c.org/P4X/DI.html. The dictionary is divided into 10-paged parts, and each is given in one file. The page numbers are in the range x0-x9. If the first page starts at 1, then the range will be x1-x9 for that part and then continue like x10-x19 and so on. If the operator creates a new category, this will not be represented in the TEI result. Table 5-1 shows how categories are represented in the TEI format.
2. **Rosetta:** Another XML based format. The dictionary is divided into 10-paged parts, and each is given in one file. The page numbers are in the range x0-x9. If the first page starts at 1, then the range will be x1-x9 for that part and then continue like x10-x19 and so on. Not all the categories are represented in Rosetta format and if the operator creates a new category, this will not be represented in the Rosetta result. Table 5-2 shows how categories are represented in the Rosetta format.
3. **Lexicon-Translation Pairs:** Gives a list of Lexicon and its Translation separated by a tab. All information is given in one file.
4. **Example of Usage-Translation Pairs:** Gives a list of Example of Usage and its Translation separated by a tab. All information is given in one file.
5. **HTML:** Each dictionary page is displayed with given tags as an HTML file.
6. **HTML with Entry Images:** Each dictionary page is displayed with tags represented as different colors along with the original dictionary entry image.

Figure 5-38, Figure 5-39, Figure 5-40, and Figure 5-41 show outputs for TEI, Rosetta, HTML, and HTML with Entry Images formats, and Table 5-3 and Table 5-4 show outputs Lexicon-Translation pairs and Example of Usage-Translation pairs for the dictionary entries in Figure 5-27.

Note: The results of generation process presented here are from the Ceb-Eng (Cebuano-English) dictionary.

Information	TEI XML
Dictionary entry	<entry> **X** </entry>
Headword	<form><orth> **X** </orth></form>
Alternative_Spelling	<form><orth> **X** </orth></form>
Pronunciation	<form><pron> **X** </pron></form>
POS	<gramGrp><pos> **X** </pos></gramGrp>
Gender	<gramGrp><gen> **X** </gen></gramGrp>
Case	<gramGrp><case> **X** </case></gramGrp>
Subcategorization	<gramGrp><subc> **X** </subc></gramGrp>
Tense	<gramGrp><tns> **X** </tns></gramGrp>
Person	<gramGrp><per> **X** </per></gramGrp>
Mood	<gramGrp><mood> **X** </mood></gramGrp>
Number	<gramGrp><number> **X** </number></gramGrp>
SenseNumber	<sense n="**X**"> </sense>
Translation	<trans><tr> **X** </tr></trans>
Compound/Derived	<re type="**derived**"><form><orth> **X** </orth></form></re>
Compound/Derived_Translation	<re type="**derived**"><trans><tr> **X** </tr></trans></re>
Example	<eg><q> **X** </q></eg>
Example_Translation	<eg><trans><tr> **X** </tr></trans></eg>
Idiom	<eg><q> **X** </q></eg>
Idiom_Translation	<eg><trans><tr> **X** </tr></trans></eg>
Language	<etym><lang> **X** </lang></etym>
Cross_Reference	<xr><ref target="**X**"> **X** </ref></xr>
Synonym	<xr><ref target="**X**"> **X** </ref></xr>
Domain	<usg type="**dom**"> **X** </usg>
Context	<usg type="**dom**"> **X** </usg>
Subject	<usg type="**dom**"> **X** </usg>
Style	<usg type="**style**"> **X** </usg>
Usage	<usg type="**style**"> **X** </usg>
Explanation	<usg type="**hint**"> **X** </usg>

Table 5-1 Representation of categories in TEI XML format. X corresponds to the actual information.

Information	Rosetta XML
Dictionary entry	<entry id=’**entry id**’ cl='**U**'> **X** </entry>
Headword	<keyForm type="**word**"><term scr="**la**" orth="**normal**"> **X** </term></keyForm>
Pronunciation	<pron scr='**la**' orth='**phonetics**'> **X** </pron>
POS	<pos> **X** </pos>
Gender	<note type='**grammar**'> **X** </note>
Case	<pos> **X** </pos>
Translation	<sense><gloss> **X** </gloss></sense>
Compound/Derived	<keyForm type="**derivative**"><term scr="**la**" orth="**normal**"> **X** </term></keyForm>
Compound/Derived_Translation	<sense><gloss> **X** </gloss></sense>
Example	<example><exTerm scr="**la**" orth="**normal**"> **X** </exTerm></example>
Example_Translation	<example><exGloss>**X**</exGloss></example>
Idiom	<keyForm type="**idiom**"><term scr="**la**" orth="**normal**"> **X** </term></keyForm>
Idiom_Translation	<sense><gloss> **X** </gloss></sense>
Language	<note type='**etym**'> **X** </note>
Cross_Reference	<refForm type='**crossref**'><term scr='**la**' orth='**normal**'> **X** </term></reform>
Synonym	<refForm type='**synonym**'><term scr='**la**' orth='**normal**'> **X** </term></refForm>
Antonym	<refForm type='**antonym**'><term scr='**la**' orth='**normal**'> **X** </term></refForm>
Domain	<note type='**domain**'> **X** </note>
Context	<note type='**context**'> **X** </note>
Subject	<subject> **X** </subject>
Style	<note type='**style**'> **X** </note>
Explanation	<note type='**hint**'> **X** </note>

Table 5-2 Representation of categories in Rosetta XML format. X corresponds to the actual information.

```
<entry>
  <form><orth>ábug</orth></form>
  <xr><ref target="ABYUG">ABYUG</ref></xr>
</entry>
<entry>
  <form><orth>abugada</orth></form>
  <gramGrp><pos>noun</pos></gramGrp>
  <trans><tr>female lawyer</tr></trans>
</entry>
<entry>
  <form><orth>abugádu</orth></form>
  <hom>
    <gramGrp><pos>noun</pos></gramGrp>
    <trans><tr>lawyer</tr></trans>
  </hom>
  <hom>
    <gramGrp><pos>verb</pos></gramGrp>
    <sense n="1">
      <trans><tr>be, become a lawyer</tr></trans>
    </sense>
    <sense n="2">
      <trans><tr>speak for someone</tr></trans>
      <eg>
        <q>Abugadúhan ta lang ka kay wá kay pangablt</q>
        <trans><tr>I'll speak for you because you don't have the knack
    of saying what you want</tr></trans>
      <re type="derived">
        <form><orth>abugaduhun</orth></form>
        <gramGrp><pos>adjective</pos></gramGrp>
        <trans><tr>lawyer-like</tr></trans>
        <eg>
          <q>Abugaduhung pangaarúngan</q>
          <trans><tr>Lawyer-like reasoning</tr></trans>
        </eg>
      </re>
    </sense>
  </hom>
</entry>
```

Figure 5-38. Dictionary part in Figure 5-27 in TEI Format

<entry id="**CEB0001000186**" cl="**U**">
<keyForm type="**word**" lang="**ceb**" reg="**modern**">
<term scr="**la**" orth="**normal**">**ábug**</term>
</keyForm>
<refForm type="**synonym**" lang="**ceb**" reg="**modern**">
<term scr="**la**" orth="**normal**">**ABYUG**</term>
</refForm>
</entry>
<entry id="**CEB0001000187**" cl="**U**">
<keyForm type="**word**" lang="**ceb**" reg="**modern**">
<term scr="**la**" orth="**normal**">**abugada**</term>
</keyForm>
<pos>**noun**</pos>
<sense><gloss>**female lawyer**</gloss></sense>
</entry>
<entry id="**CEB0001000188**" cl="**U**">
<keyForm type="**word**" lang="**ceb**" reg="**modern**">
<term scr="**la**" orth="**normal**">**abugádu**</term>
</keyForm>
<pos>**noun**</pos>
<sense><gloss>**lawyer**</gloss></sense>
</entry>
<entry id="**CEB0001000189**" cl="**U**">
<keyForm type="**word**" lang="**ceb**" reg="**modern**">
<term scr="**la**" orth="**normal**">**abugádu**</term>
</keyForm>
<pos>**verb**</pos>
<sense><gloss>**be**, **become a lawyer**</gloss></sense>
</entry>
<entry id="**CEB0001000190**" cl="**U**">
<keyForm type="**word**" lang="**ceb**" reg="**modern**">
<term scr="**la**" orth="**normal**">**abugádu**</term>
</keyForm>
<pos>**verb**</pos>
<sense>
<gloss>**speak for someone**</gloss>
<example>
<exTerm scr="**la**" orth="**normal**">**Abugadúhan ta lang ka kay wá kay pangablt**</exTerm>
<exGloss>**I'll speak for you because you don't have the knack of saying what you want**</exGloss>
</example>
</sense>
</entry>
<entry id="**CEB0001000191**" cl="**U**">
<keyForm type="**derivative**" lang="**ceb**" reg="**modern**">
<term scr="**la**" orth="**normal**"> **abugaduhun**</term>
</keyForm>
<pos>**adjective**</pos>
<sense>

```
    <gloss>lawyer-like</gloss>
    <example>
      <exTerm scr="la" orth="normal">Abugaduhung
    pangaarúngan</exTerm>
      <exGloss>Lawyer-like reasoning</exGloss>
    </example>
  </sense>
</entry>
```

Figure 5-39 Dictionary part in Figure 5-27 in Rosetta Format

abugada	female lawyer
abugádu	lawyer
abugádu	be, become a lawyer
abugádu	speak for someone
abugaduhun	lawyer-like

Table 5-3 Lexicon-Translation Pairs for the Dictionary part in Figure 5-27

Abugadúhan ta lang ka kay wá kay pangabla	I'll speak for you because you don't have the knack of saying what you want
Abugaduhung pangatarúngan	Lawyer-like reasoning

Table 5-4 Example of Usage-Translation Pairs for the Dictionary part in Figure 5-27

ábug

Synonym: ABYUG

abugada

POS: n [noun]
Translation: female lawyer

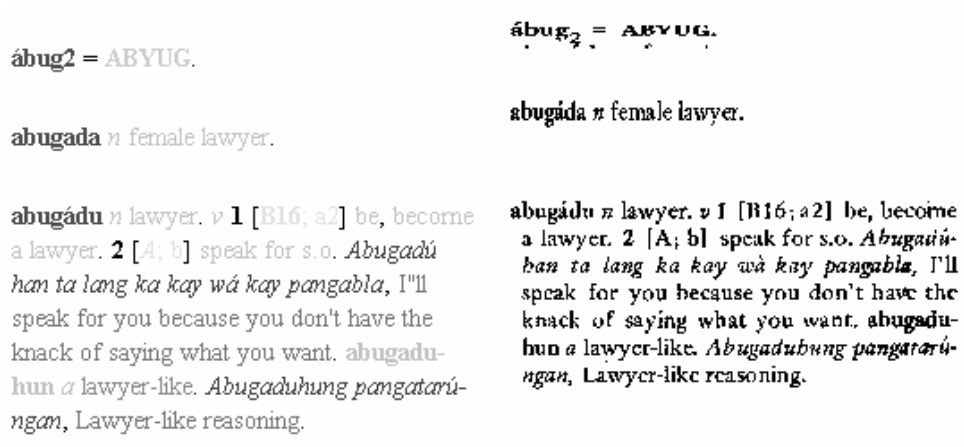
abugádu

POS: n [noun]
Translation: lawyer
POS: v [verb]
SenseNumber: 1
Translation: be become a lawyer
SenseNumber: 2
Translation: speak for someone
Example: [Abugadúhan ta lang ka kay wá kay pangabla] I'll speak for you because you don't have the knack of saying what you want

abugaduhun *(Compound/Derived)*

POS: a
Translation: lawyer-like
Example: [Abugaduhung pangatarúngan] Lawyer-like reasoning

Figure 5-40 Html output for the entries in Figure 5-27



ábug2 = ABYUG.

abugada *n* female lawyer.

abugádu *n* lawyer. *v* **1** [B16; a2] be, become a lawyer. **2** [A; b] speak for s.o. *Abugadú han ta lang ka kay wá kay pangabla*, I'll speak for you because you don't have the knack of saying what you want. **abugaduhun** *a* lawyer-like. *Abugaduhung pangatarúngan*, Lawyer-like reasoning.

ábug2 = ABYUG.

abugáda *n* female lawyer.

abugádu *n* lawyer. *v* **1** [B16; a2] be, become a lawyer. **2** [A; b] speak for s.o. *Abugadúhan ta lang ka kay wà kay pangabla*, I'll speak for you because you don't have the knack of saying what you want. **abugaduhun** *a* lawyer-like. *Abugaduhung pangatarúngan*, Lawyer-like reasoning.

Figure 5-41 Dictionary part in Figure 5-27 in HTML with Entry Images

6 BRIDGE-View

As mentioned earlier, the generation module is responsible for taking the output of tagging and producing usable results. One of the direct applications of the results so obtained is the Search and Retrieval Interface which represents the BRIDGE-View. This is a utility that lets the user search for an actual dictionary entry. In other words, this module emulates the process of looking up for a particular word in a hard copied paper dictionary. The idea behind this tool is to convert the results obtained from the entire process followed till this point into a usable format.

6.1 Introduction to the Search and Retrieval Tool

Figure 5-38 and Figure 5-39 show the XML representation of each of the individual dictionary entries after the generation operation is performed. A closer look at these entries reveals that the information contained within these entries, as it appears in the dictionary, for instance, the headword, its part of speech, translation, example of usage, etc. is enclosed inside a separate tag in the XML. Table 5-1 and Table 5-2 explain how this information is tagged inside the XML using the conventions followed by the TEI and Rosetta standards. The search and retrieval tool makes use of these distinct tags to enable searching of a given text based on this tagged information. This search operation can be performed by parsing the XML tags and storing them as separate fields. The parsing of the XML is done using Digester, an open source utility from Jakarta Commons that facilitates XML file processing and parsing. The search operation is implemented using Jakarta Lucene**,** a high-performance, full-featured text search engine library from The Jakarta Project.

The user must first store the information inside the tags as separate fields so that they can be searched individually. As observed from Figure 5-38 and Figure 5-39, the dictionary information for the entries is stored inside tags or the XML elements in a hierarchy. For example, the headword is specified within the elements as:

<entry> <form> <orth> [headword] </orth> </form> </entry>.

Similarly, the part of speech information is contained as:

<entry> <gramGrp> <pos> [part of speech] </pos> </gramGrp> </entry>

For simplicity reasons, ‘this element hierarchy’ is first removed so that the information contained always appears in the following format:

<entry> [headword] </entry>
<entry> [part of speech] </entry>

Thus, we create a new set of XML files, which is an overlay on the existing XML files created at the end of generation process, and use these new files to get the information inside the fields. After, creating the XML files in the above format, they need to be parsed to extract and store the information within the elements, as fields. A "field" is simply an XML element which is stored as: element name and information enclosed within the element. Thus, the information about, say, part of speech, enclosed within

<pos> [part of speech] </pos>

corresponds to the field name "pos" and the information within it, i.e., [part of speech].

Note:
The content enclosed within the square braces ([ ]) is indicative of the actual information. The XML files used for the search and retrieval operation must be in the TEI format. Currently, XML files in the Rosetta format cannot be used to generate the search and retrieval results.

We are currently developing the search and retrieval tool such that this operation is not required and can be eliminated. That is, the user will not be required to create another extra set of XMLs and parsing and searching can be performed directly on the XMLs obtained from Generation.

When the actual interface is run, any text can be looked up in the stored fields, as mentioned above. Optionally, constraints can be put on the search, such that, the text that is being searched appears in conjunction with specified field values only. The search operation is defined in greater detail in the following section.

6.2 Using the BRIDGE-View

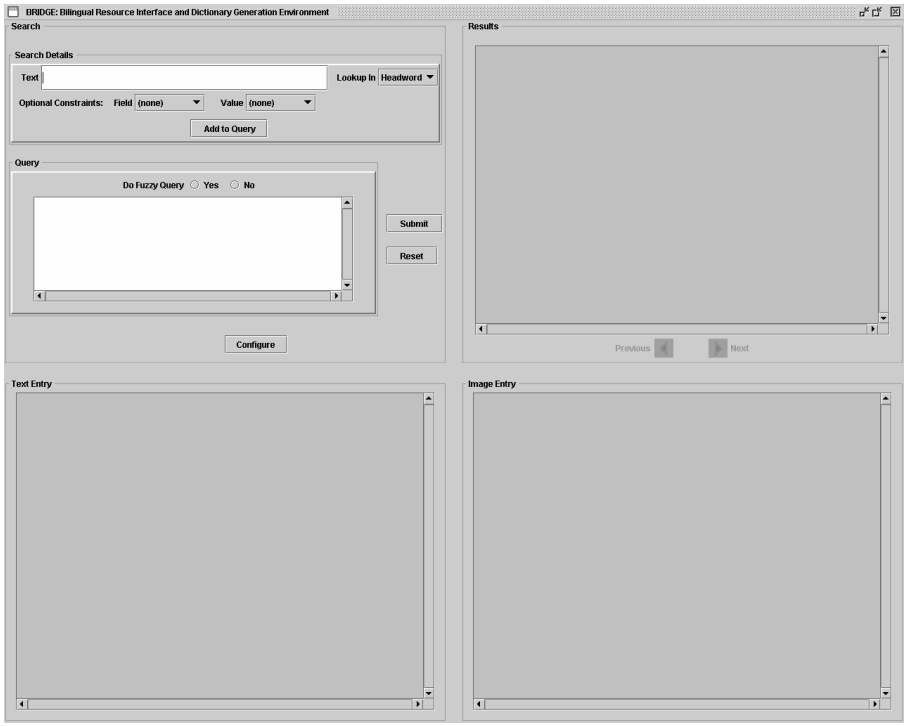


Figure 6-1 BRIDGE-View Interface

In the BRIDGE interface, click on the **Create View** button in the View panel. As explained above, this will create the set of new XML files, parse them, and store all the information contained within the various elements (or tags) in the XML in the form of fields.

Now click on the **Run View** button. This will display the actual GUI to be used for conducting the search (Figure 6-1). The BRIDGE-View interface is divided into four panels: Search, Results, Text Entry and Image Entry. Each of these four panels serves different purposes as explained below:

6.2.1 Search Panel

This panel consists of two sub-panels: Search Details and Query.

6.2.1.1 Search Details

This sub-panel allows the user to specify the query to be searched, the fields in which to search and to set the optional constructs. These components are explained below.

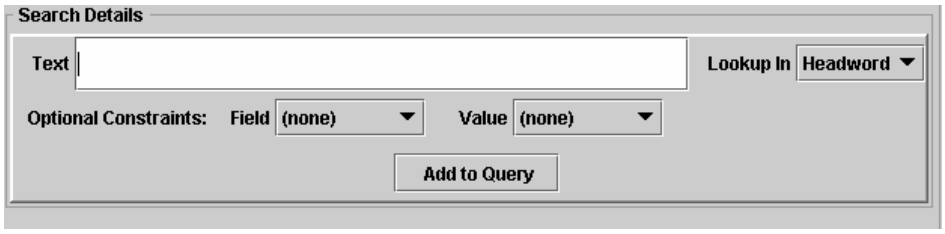


Figure 6-2 Search Details (within the Search panel): Used to specify the search criteria

Text: The user must specify the text that he wants to look up, in the empty area labeled as Text. For instance, if you want to look up for a word panabu, this word must be entered in the Text field.

Lookup in: This menu allows the user to select the fields that must be searched for the desired text. Thus, if the user wants to search for the occurrences of panabu in the headword, select Headword from the selection list. Similarly, selecting Examples of Usage from the selection list had the effect of looking for the specified text in all the examples of usage in the dictionary (stored inside the field name Examples of Usage).

The following two optional constraints can be specified to perform a narrower search. This allows the user to search the desired text within the entire dictionary such that a specific field holds a particular value.

Field: This lets the user to select the field that must be constrained.

Value: The user can select the value, to which the field must be constrained.

Thus, the user might want to specify that the text panabu might be looked up in only those entries in the dictionary such that the part of speech (Field) is noun (Value).

Once the user has specified all the details for the search, clicking on the button **Add to Query** adds the query (search text) to the text area shown inside the Query sub-panel.

6.2.1.2 Query

This sub-panel has a text area that displays the query (search text) the user has requested. It provides following options to the user as explained below.

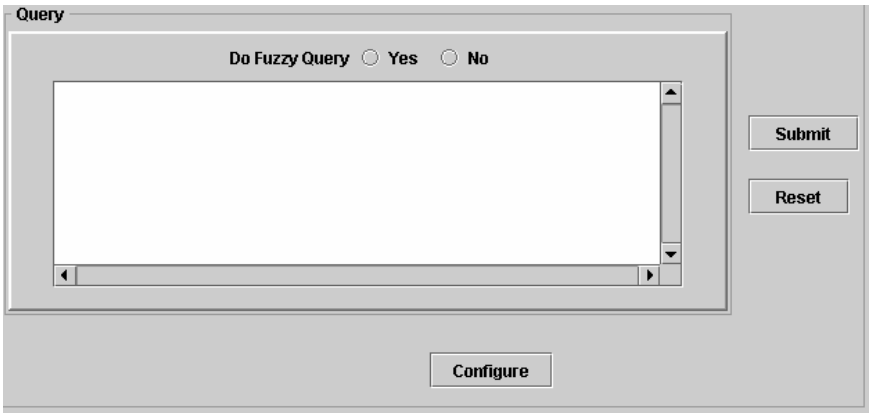


Figure 6-3 Query (within the Search panel): Allows the user to view the query constructed above

Do Fuzzy Query: The user can specify if a Fuzzy Query (based on the Levenshtein algorithm) is to be performed for the specified search.

Note: As an example of a fuzzy search, assume that the user wants to search for the text 'roam'. A fuzzy search will, then, also look up words such as 'roams', 'foams', etc.

Submit: Clicking on the **Submit** button brings up a dialog box asking the user to confirm the query and starting the search. This is shown in Figure 6-4. After the user clicks **Yes** in the dialog box, the search operations is performed and the results are displayed in the Results panel.

Reset: Clicking on this button clears out all the search criteria (and any displays currently seen in other panels), and allows the user to start a new search from the beginning.

Configure: Currently, this button is not functional. Please refer to section 6.3 Additional Capabilities for further details.

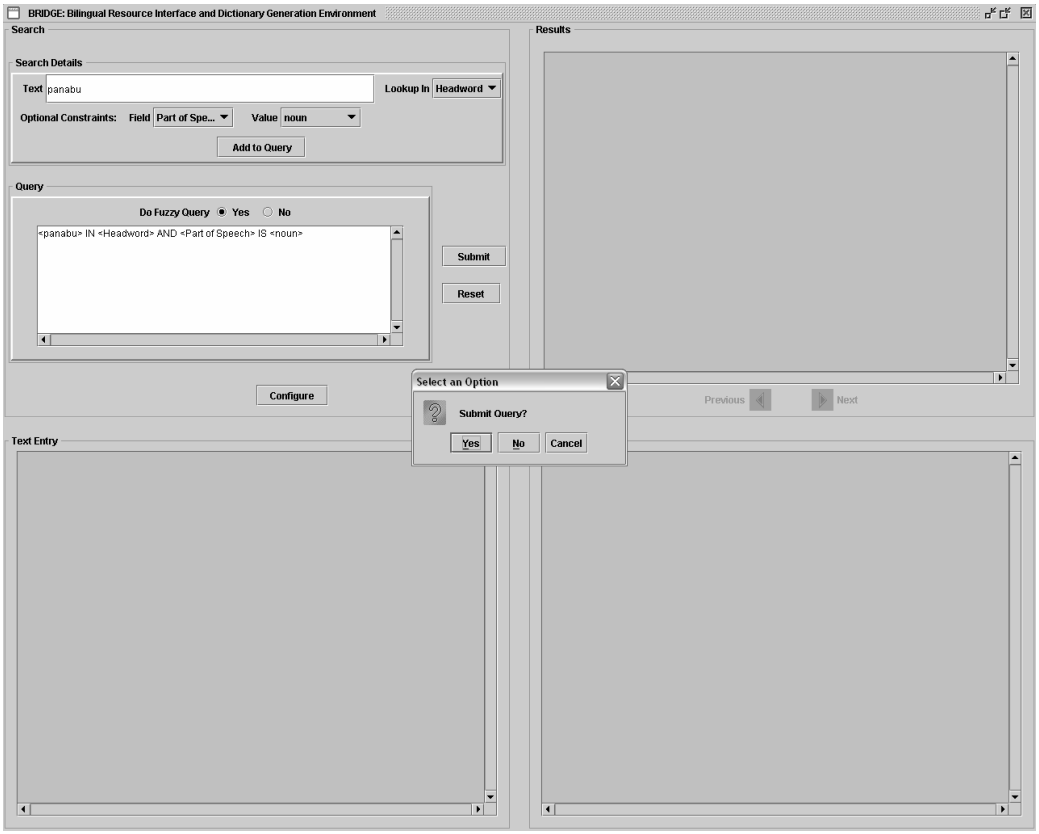


Figure 6-4 Submitting a query

6.2.2 Results Panel

The Results panel is used to display the results of the search performed. It indicates the number of results, or hits, obtained for the search and displays each of them in the decreasing order of the score. The score of a particular hit, here, is calculated based on the frequency of the desired text in the given entry against the other hits, and is computed as a decimal value between 0 and 1.

Each of the results appears in the panel with the headword (displayed as a link) for the matched entry, some context obtained from the dictionary entry, the score and a View Page link.

The **Previous** and **Next** buttons are currently not active. Refer to section 6.3 Additional Capabilities for details.

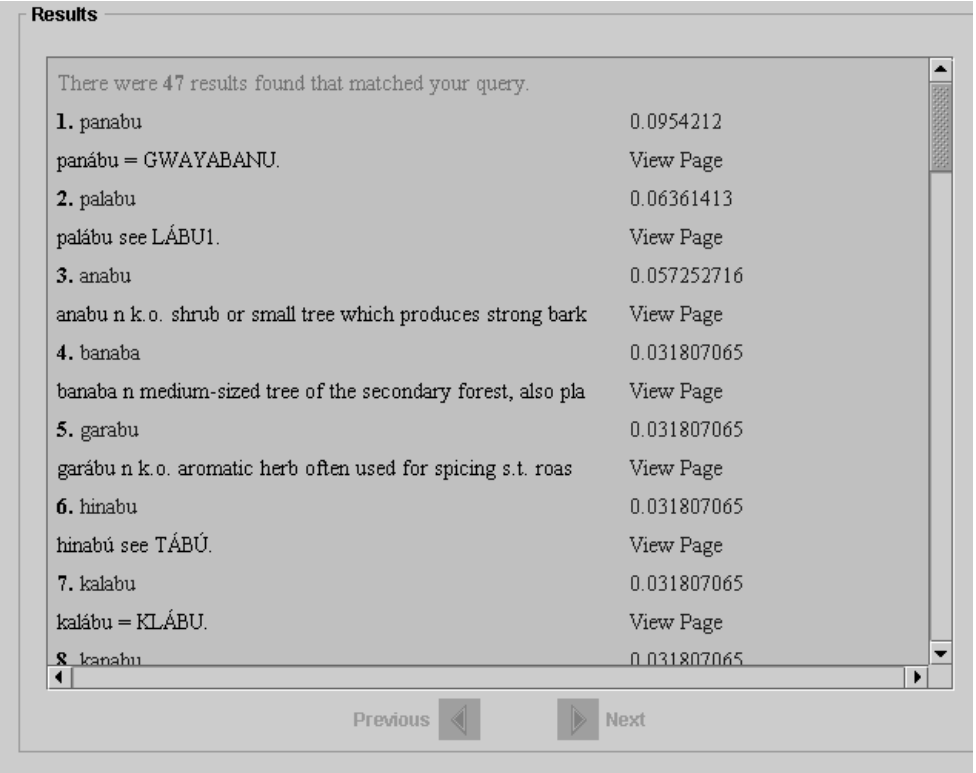


Figure 6-5 Results panel: Displays the results, or hits, obtained for the given search

6.2.3 Text Entry and Image Entry Panels

These panels are used to view the actual dictionary entry corresponding to a given result displayed in the Results panel. Whenever the user clicks on the headword link in the results panel, the actual entry from the dictionary (scanned from the hard copied dictionary and stored as images) gets displayed in the Image Entry panel. Also, the html file generated from the generation operation is displayed in the Text Entry panel (Figure 6-6).

If the user clicks on the View Page button displayed in the Image Entry panel, a new window appears and the actual dictionary page (which was scanned at the beginning of the entire procedure) containing this result can be viewed.

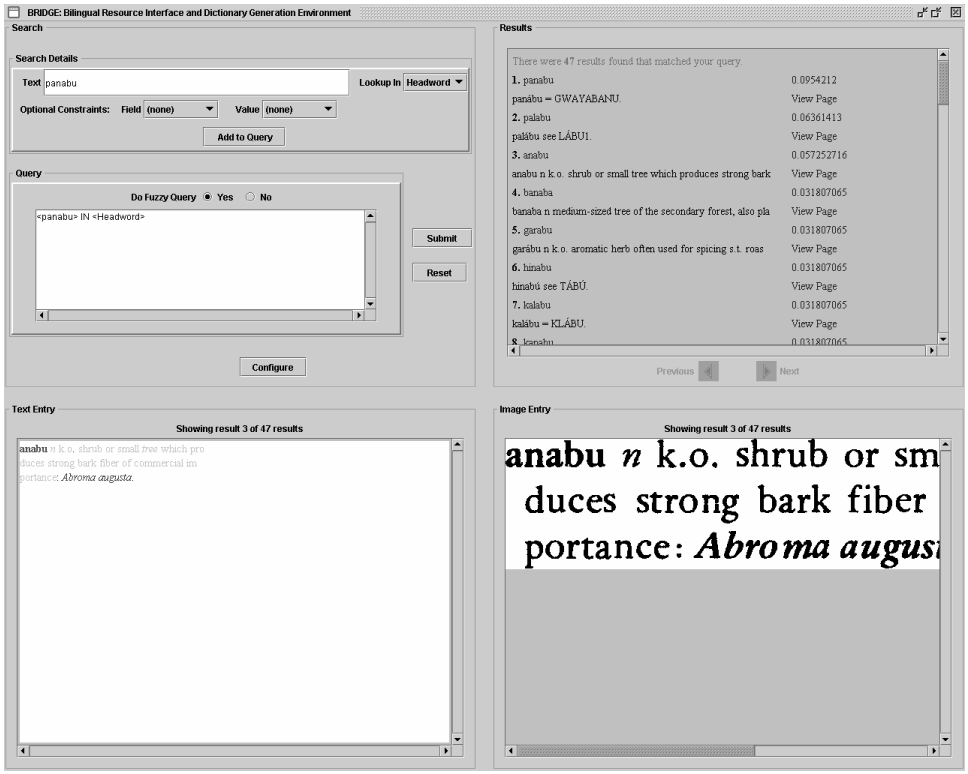


Figure 6-6 User can view the html entry and the actual dictionary entry by clicking on the Headword link in the results panel

Note: The user is advised to take care so that the BRIDGE-View application is quit properly and multiple windows for the application are not open at the same time, since this can lead to “File Not Found Errors” and can cause the application to hang.

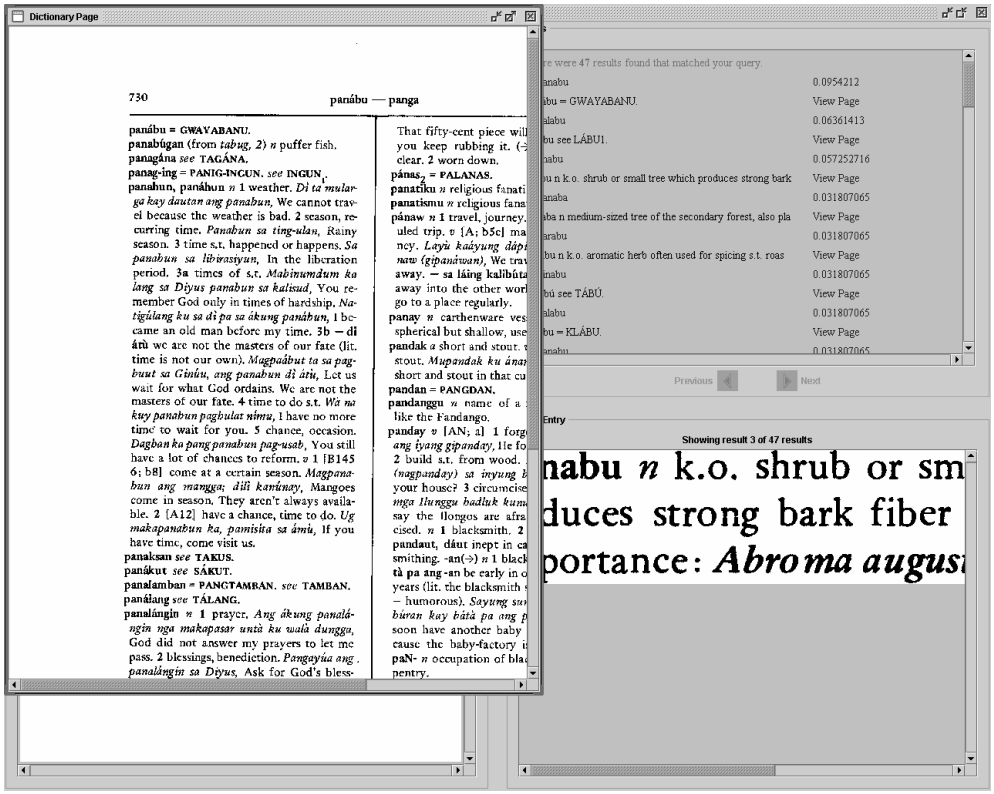


Figure 6-7 Clicking on a View Page link enables the user to view the dictionary page where the headword actually appears

6.3 Additional Capabilities (to be implemented)

The search and retrieval tool has been developed to demonstrate how the results obtained from generation (and in effect, the entire process) can be readily used for an application. Thus, this tool in its current form provides limited capabilities, and we are currently developing this tool to perform additional functionalities, some of which are described below.

As mentioned earlier, this interface is currently available for the Ceubano dictionary only. While it may be used with other dictionaries also, the results for the search and retrieval are not guaranteed to be accurate. Work is in progress to make this utility generic so that it can be used for any bilingual dictionary.

This tool can be made generic so that it can be used for any dictionary that has been previously segmented and tagged by creating the corresponding indices. This includes, for example, obtaining the values for the Lookup in menu (i.e. the fields available for

searching a given text), directly from the tagging results with the help of config files generated at the time of tagging.

The Configure button in the interface shall enable the user to do the following:
1. The user will have an option of specifying the number of results to be displayed per page in the Results panel. The user can then browse through these pages by using the **Previous** and **Next** buttons in the Results panel.
2. The user can also select another dictionary and run the interface for that dictionary by choosing from the Choose Another Dictionary option.

The API for the search tool used for performing the search operation (Lucene 1.3) does not have the capability to display the context for the results obtained. We are trying to include this functionality into the application so that the user has a better idea of the particular result displayed in the Results panel.

Another enhancement planned for this application is the creation of a Full Page Electronic Dictionary Viewer where the user can browse through the dictionary in the same way as he/she would in an actual physical dictionary.

7 Example Tutorial

In the following tutorial we will demonstrate the operation of our system on selected dictionary pages. We will start from system configuration and display the results of the generation procedure. After that we will demonstrate how to use the Search and Retrieval Tool for locating a dictionary entry.

It should be noted that the dictionary that the operator will be using may be a different one and hence the results presented here are applicable to our dictionary pages only.

7.1 Background Preparation

Dictionary selection: We are using twenty pages of the Cebuano dictionary for this example.
Scanner: We are using Fujitsu fi-4220C scanner with 400 DPI setting.

7.1.1 System Configuration

System Configuration involves setting up the scanner, configuring zone colors and personal dictionaries for spell check. We will go over each of these topics one by one.

Configuring the system paths for different applications
In the BRIDGE interface, click on the **Config → Config Utility** sub-menu. This will bring up the Bridge Preferences dialog box as shown in Figure 7-1

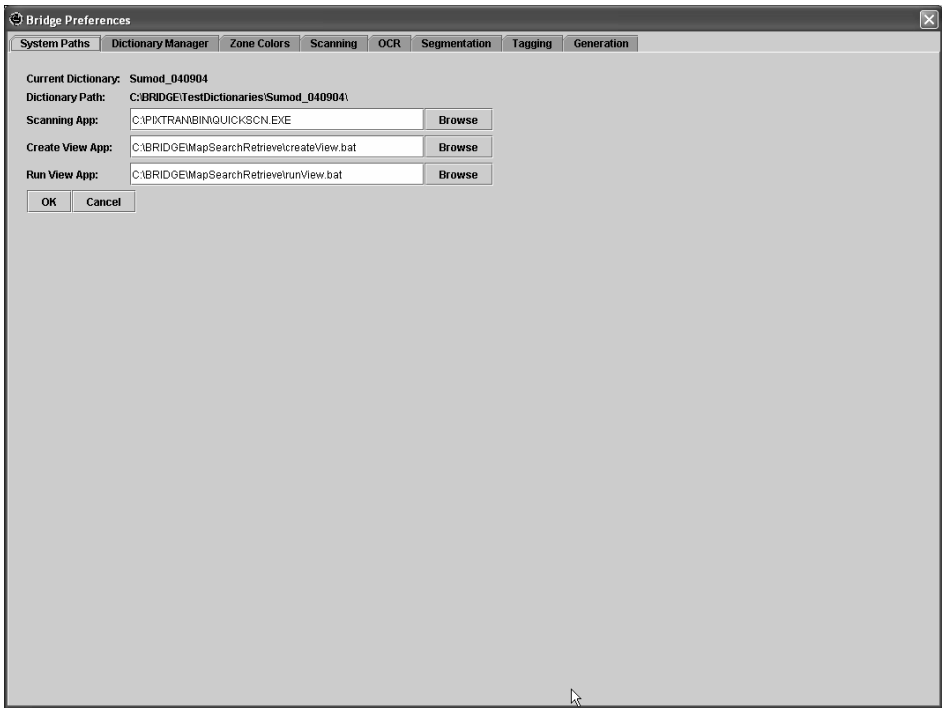


Figure 7-1 Bridge Preferences Dialog Box

Click on the System Paths tab. This will display the text fields containing the information about Current dictionary, Dictionary Path, Scanning Application Path, Create View Application Path and Run View Application Path. The operator can change the paths for the different applications depending on their settings. For the purpose of this example, we will use the predefined paths. Click the **OK** button to save these settings.

Configuring the Zone Colors

In OCR, segmentation, tagging and generation different zone types can be distinguished by assigning to them different colors. To set up the zone colors, bring up the **Bridge Preferences** dialog box by clicking on **Config → Config Utility** (Figure 7-1). Click on the Zone Colors tab to display the following dialog box.

Figure 7-2 Setting up Zone Colors

On the left side of the Zone Colors panel, we can see three tabs arranged vertically. They are OCR, Segmentation and Tagging. As shown, currently the OCR tab is clicked and the properties that apply to OCR are shown in the table on right side.

Click on the Segmentation and Tagging tabs and observe the changes in different parameters. You will notice the changes in file location. This is due to the dictionary structure to be explained later.

In the OCR panel, click on any row of the table located on the right side. This will bring up another table (Figure 7-3) that shows the color representation for different zones.

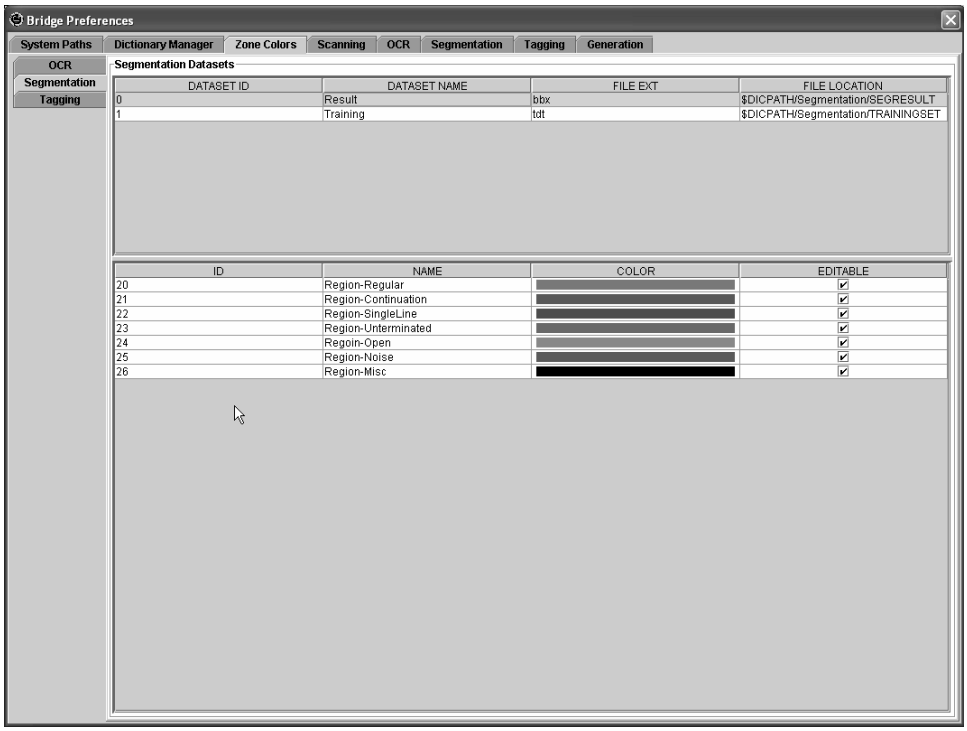


Figure 7-3 Color Configuration for different zones

The above figure shows the zone color selection for the segmentation procedure. There are four attributes for a section in each of the three operations (OCR, Segmentation and Tagging). These are ID, NAME of the zone, COLOR and a boolean attribute EDITABLE. For every procedure (OCR, segmentation etc.) the image will be marked into different regions. For example, in the segmentation operation the image will be divided into different regions such as regular, continuous etc. For each such region, there is an ID, a NAME, a COLOR and a checkbox that tells if the region is editable or not.

Click on the Region-Continuation row in the COLOR column. This will bring up a color selection palette **Pick a Color** as shown in the following figure.

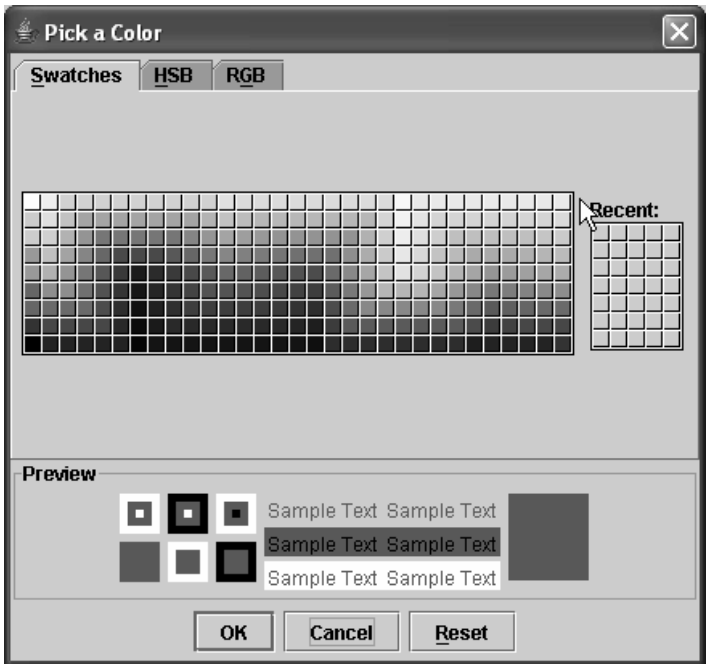


Figure 7-4 Pick a Color dialog

The color for a zone can be changed using this palette. For this example, we will use the default color settings.

Configuring Personal Dictionaries for spell check
This operation will be explained in a later section as we progress along the execution of our example.

7.2 Dictionary Creation

There are two methods of creating a dictionary; to create a new dictionary from the scratch or to transform an existing dictionary into a new dictionary. We will explain both.

7.2.1 Creating a new dictionary from scratch

This procedure is explained for the purpose of understanding. In our example, we will be using the configuration of an already existing dictionary to create a new dictionary. It is explained in section 7.2.2.

Click on the sub-menu **Config → Config Utility,** on the BRIDGE interface. This will bring up the **Bridge Preferences** dialog box. Open the tab Dictionary Manager to bring up following dialog.

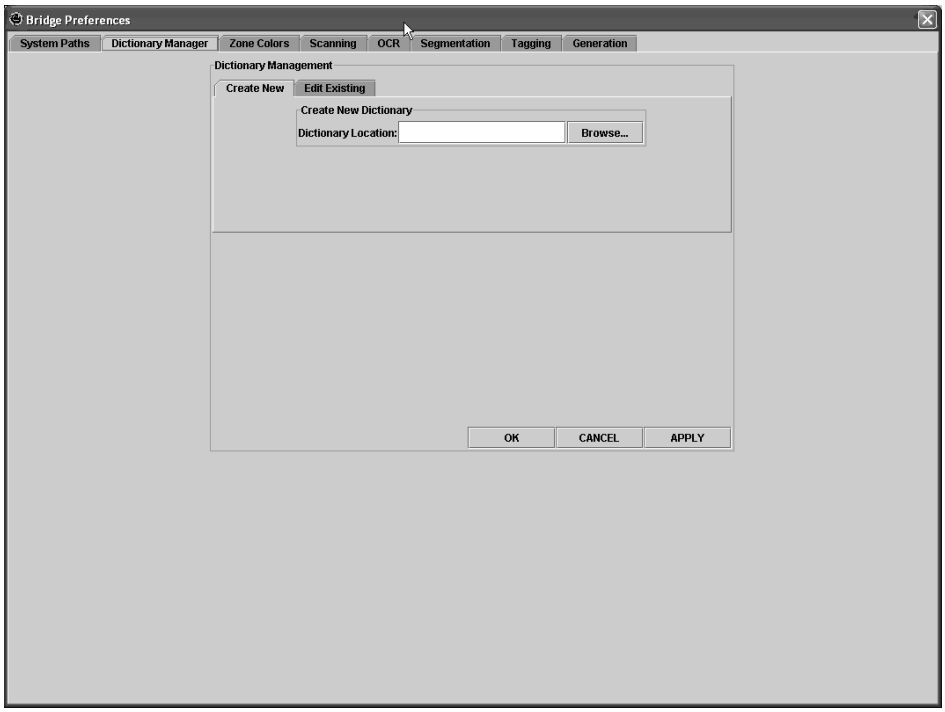


Figure 7-5 Creating a new dictionary from scratch

Click on the Create New tab if it is not already clicked. For the purpose of our example, we are creating a dictionary in C:\BRIDGE\TestDictionaries folder. We will name this dictionary as ExampleDictionary. Type the destination folder name in the Dictionary Location text field as C:\BRIDGE\TestDictionaries\ExampleDictionary and click the **OK** button.

Creating a new dictionary produces a directory structure that holds different modules in the BRIDGE i.e. segmentation, tagging etc. Navigate to the folder we have just created and you can observe the directory structure as shown in the following figure.

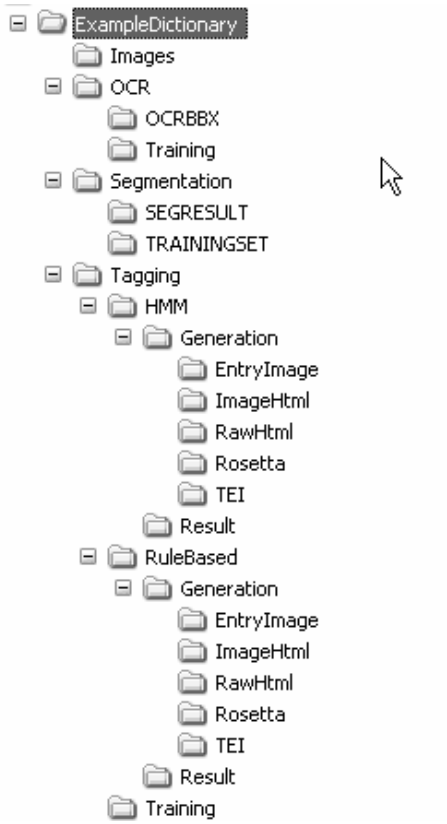


Figure 7-6 Directory tree for a newly created dictionary

7.2.2 Creating a dictionary by editing an already existing dictionary

We will use the configuration of a dictionary already created. We will copy this configuration to our new dictionary. The name of the new dictionary in our example is ExampleTutorial.

Click on the **Config → Config Utility** menu to open the **Bridge Preferences** dialog box as shown in Figure 7-7.

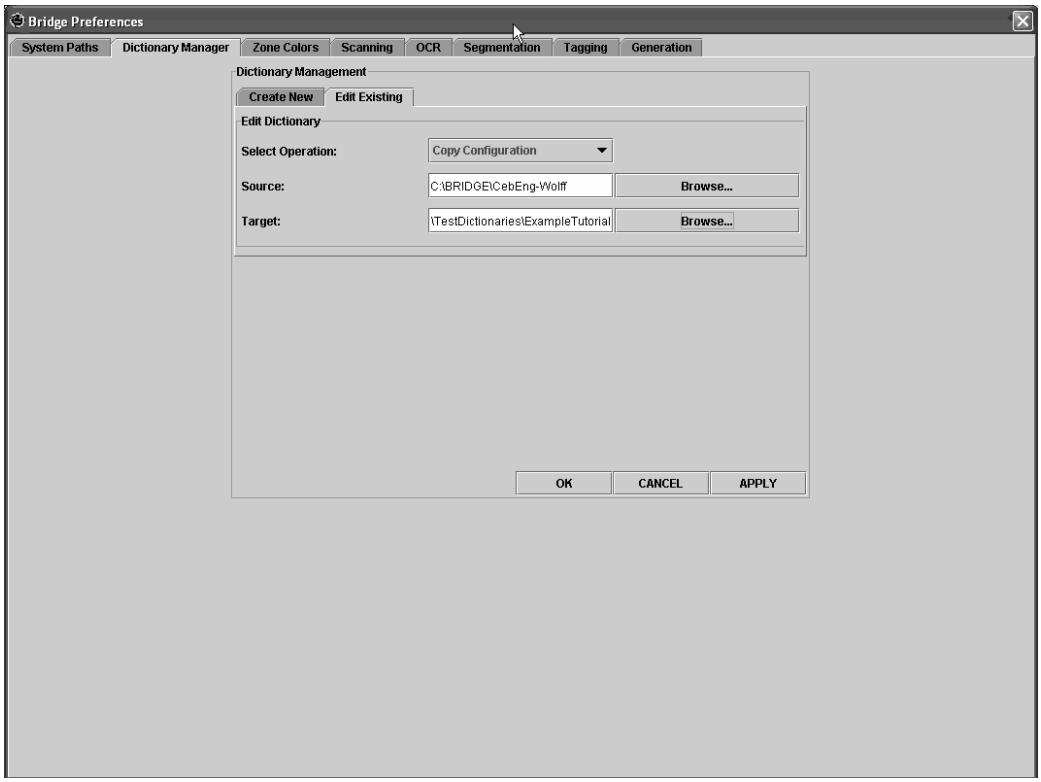


Figure 7-7 Copying the configuration of an existing dictionary to create a new dictionary

Select the operation Copy Configuration from the drop down menu. Specify the paths for the source and the target dictionaries for this operation in the Source and the Target text fields respectively. For our example, the source dictionary is C:\BRIDGE\CebEng-Wolff. We will create the new target dictionary (ExampleTutorial) in C:\BRIDGE\TestDictionaries. Click on the **OK** button to start copying the configuration. After the operation is successful, as indicated by a message box, navigate the directory structure and observe the directory structure of the dictionary folder just created. For our example, the dictionary name is ExampleTutorial and the directory tree structure will be seen as in the following figure.

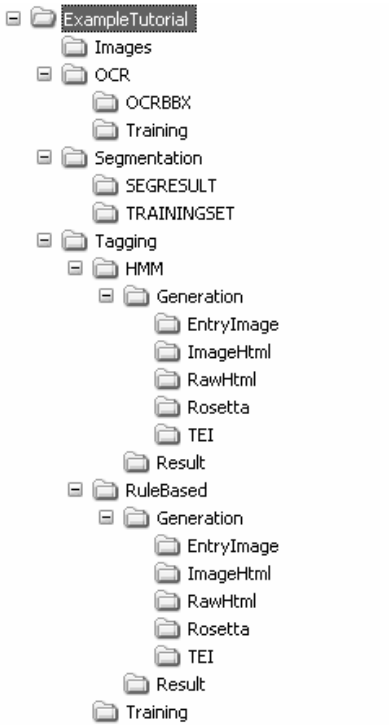


Figure 7-8 Directory tree of a ExampleTutorial

In the folder C:\BRIDGE\TestDictionaries, you will notice the presence of an "ExampleTutorial.dic" file. A file with the name of the created dictionary and ".dic" extension is created when a new dictionary is created. This file is placed in the same parent folder as that of the created dictionary.

After the dictionary is created we will now proceed to the actual operations on the paper dictionary.

Open the BRIDGE interface and select the OCR tab if it is not already selected.

7.3 Scanning

7.3.1 Setting up the scanning configuration

In the workflow area of OCR, click on the **Scan** button to activate the third party Pix Tools software and bring up the QuickScan interface (Figure 7-9).

Figure 7-9 The QuickScan interface

In the QuickScan interface, go to the **Scan → New Batch** sub-menu and you will see the following dialog box.

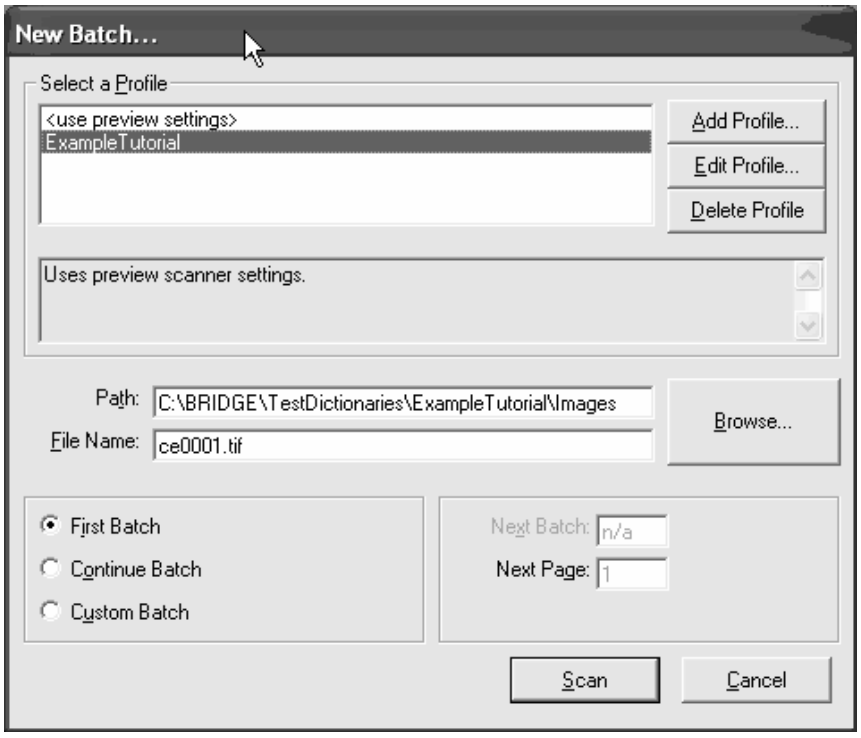


Figure 7-10 Using profile to set scanning parameters

We have created a scanning profile for the purpose of this example. Select the ExampleTutorial profile by highlighting it. We are using 400 DPI for our scanning operation. Depending on the scanner specifications, the operator may need to use different settings and create their own profile.

For the actual scanning operation, the operator needs to specify the destination folder to store the scanned images. To specify the destination folder click on the **Browse** button in the **New Batch** dialog box and navigate to the folder C:\BRIDGE\TestDictionaries\ExampleTutorial and select the Images folder so as to display the path as shown in the Path text field in Figure 7-10 above.

You can edit the scanning profile or add a new profile and base your scanning parameters according to that.

7.3.2 Performing the scanning operation

After you have completed the settings, click the **Scan** button to start scanning. While the scanning is in progress, you can see the images of the pages being scanned in the **QuickScan** interface dialog box.

Note: Every scanner will have different controls and parameters. You need to refer to the manual of your particular scanner and perform this operation accordingly for optimum results. It is advised to select the Images folder in your dictionary structure as the destination folder for the scanned images.

To make sure that the scanning operation has been completed, see the contents of the destination folder where you wanted the scanned images to be stored. This folder should contain the page images scanned in the ".tif" format. For our example, we are using 20 pages of the Cebuano dictionary. Every page (or image) has a naming convention like "ce####.tif"; where "ce" stands for Cebuano and the four #s stand for the image number. So our images range from ce0001 to ce0020.

After scanning we will now move to OCR.

7.4 OCR (Optical Character Recognition)

7.4.1 Performing OCR

In the BRIDGE interface, click on the **File → Open a Dictionary** sub-menu. This will open a file chooser dialog box. Go to C:\BRIDGE\TestDictionaries and select ExampleTutorial.dic. Now the OCR panel will appear as shown in Figure 7-11.

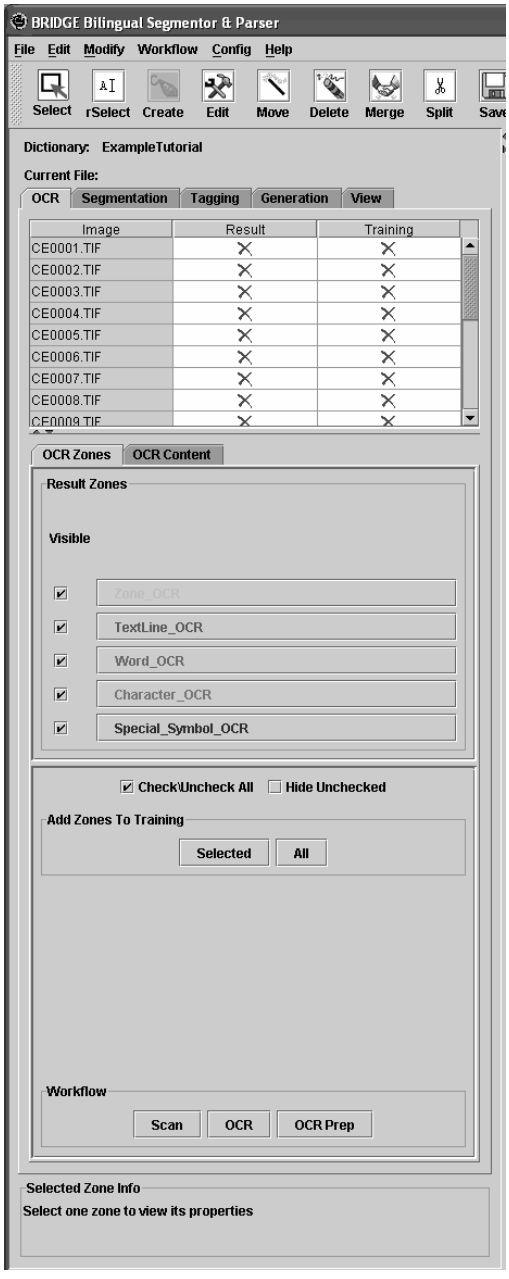


Figure 7-11 OCR Panel

The Image column displays the names of the image files that we have scanned. The absence of the Result files and the Training data sets are indicated by the red cross marks in those columns.

In the Workflow subpanel, click on the **OCR** button. This will bring up the following dialog box.

Figure 7-12 Perform OCR dialog box

Select the “end file” as ce0020.tif, so we can perform OCR on the complete set. In the language option, for this dictionary set, select “Catalan. Spelling supported”. The reason is that Cebuano contains alphabets with accents and Catalan is more suitable than English. Click the **OK** button to start OCR.

After the OCR operation is completed, check the contents of the directory C:\BRIDGE\TestDictionaries\ExampleTutorial\OCR. You will notice ce####.let files where the character ‘#’ stand for a number.

7.4.2 OCR Preparation (OCR Organizer)

To perform OCR correction we need to generate “.ocr” files. This can be achieved by the OCR preparation operation. OCR preparation operation converts the “.let” files into “.ocr” files.

In the OCR tab of the BRIDGE interface, click on the **OCR Prep** button. This will bring up an **OCR Organizer** dialog box as shown in the following figure.

Figure 7-13 OCR Organizer

Select the end file as ce0020.let and click **OK** to perform tagging preparation on the entire batch. After tagging preparation is completed, you can see ".ocr" files in the folder C:\BRIDGE\TestDictionaries\Tutorial\OCR\OCRBBX.

7.4.3 OCR Correction

OCR correction involves performing spell check operation on the electronic text extracted from the image. The operator can setup personal dictionaries for the spell check.

Configuring Personal Dictionaries

Click on the **Config → Personal Dictionary** sub-menu in the BRIDGE interface to bring up the following dialog box.

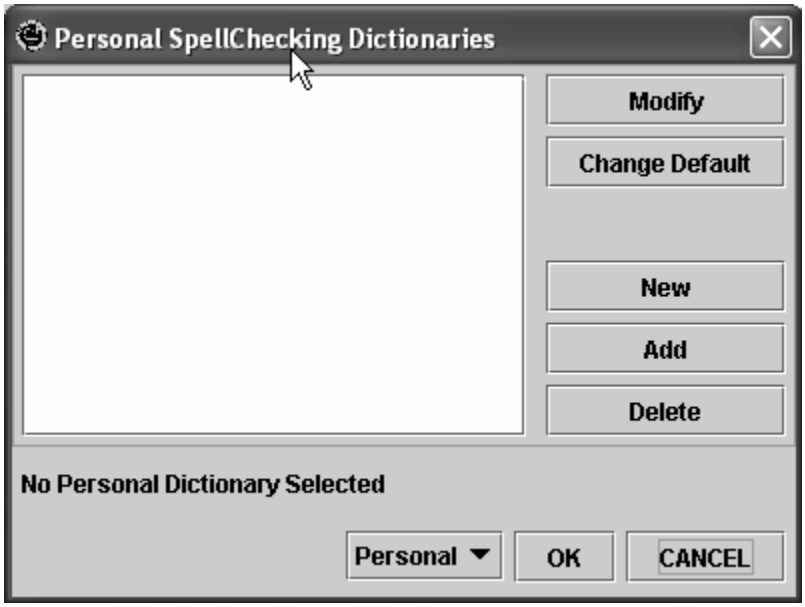


Figure 7-14 Personal Dictionaries for spell check

Click on the **New** button to create a personal dictionary. At present the personal dictionaries are created in the "./config" directory. Also, you need to specify the ".dic" extension with the name of the personal dictionary.

Create a new personal dictionary named TutorialPersonal.dic. Mark the checkbox for this dictionary and click the **OK** button to set this as a personal dictionary.

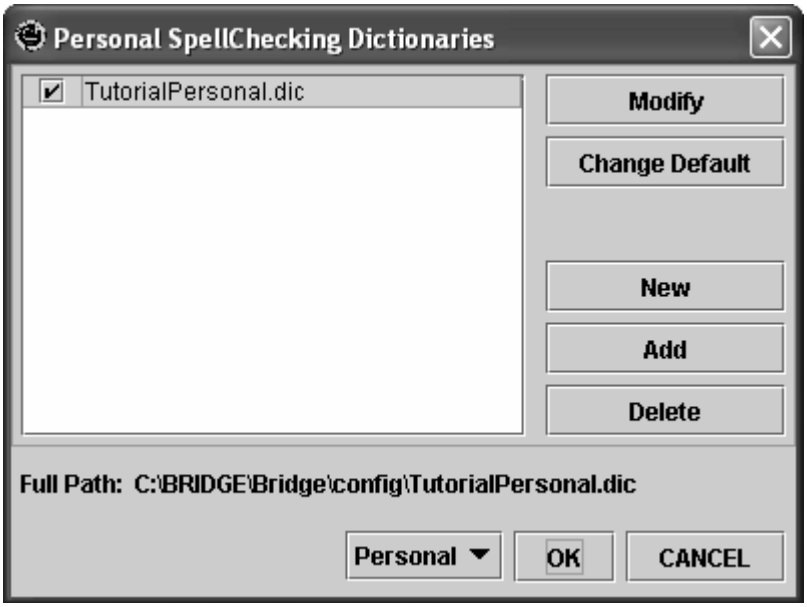


Figure 7-15 Setting up personal dictionaries

Note: When the operator adds a word to the personal list, it gets added to the system. After the process, the added words will be stored in PersonalList.dic, which is created by default in "./config" directory.

Running Spell Check

In the OCR panel, you will notice that the results are available by the presence of the green check marks in the Result column. Open the first file by clicking in that cell. Now click on the OCR Content tab to view the electronically extracted text (Figure 7-16).

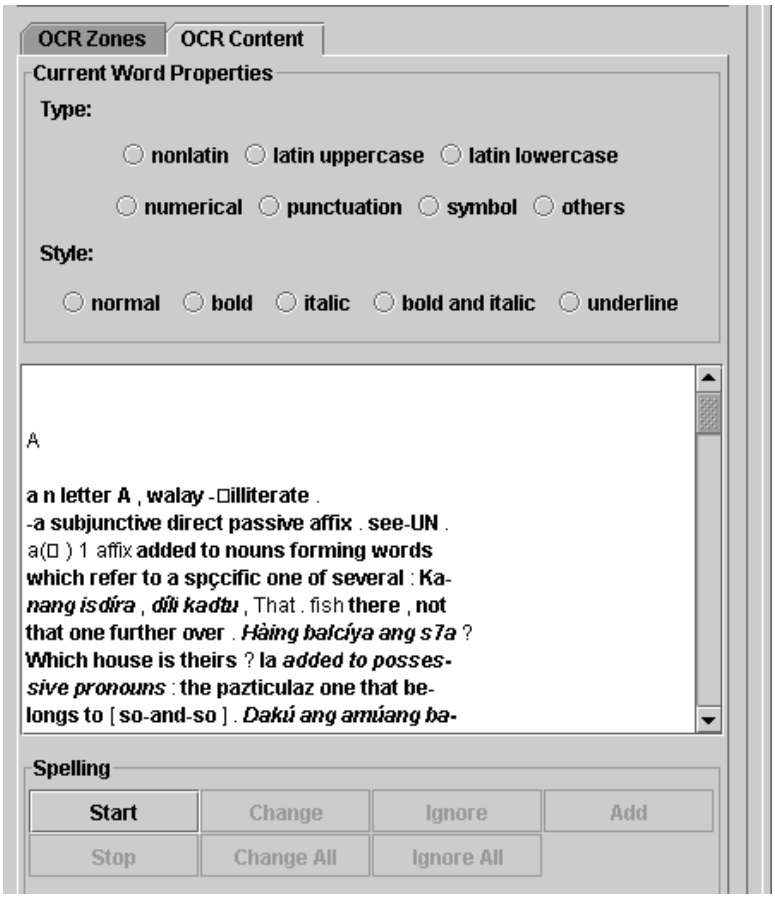


Figure 7-16 Extracted text from the image

To start the spell check, click on the **Start** button. For every word checked, that word will be highlighted and a list of possible suggestions will be displayed as shown in the following figure.

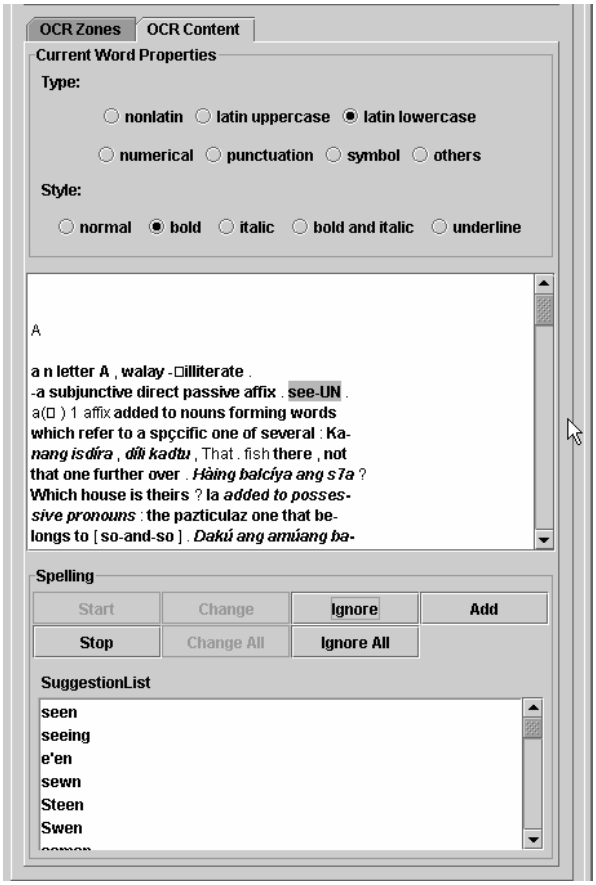


Figure 7-17 Performing spell check

The operator can add this word to the personal list, ignore it or substitute it with a word in the suggestions list. The type and style of the current word are displayed in the panel Current Word Properties. The operator can also change these attributes by using the radio buttons provided.

Example of spelling correction

Observe Figure 7-18; you can see that a word with incorrect spelling is highlighted. Compare the contents of the extracted text with that of the image as shown in Figure 7-19. Select the word “specific” from the Suggestion List to correct the error and click the **Change** button (Figure 7-20). The word gets replaced as shown in Figure 7-21.

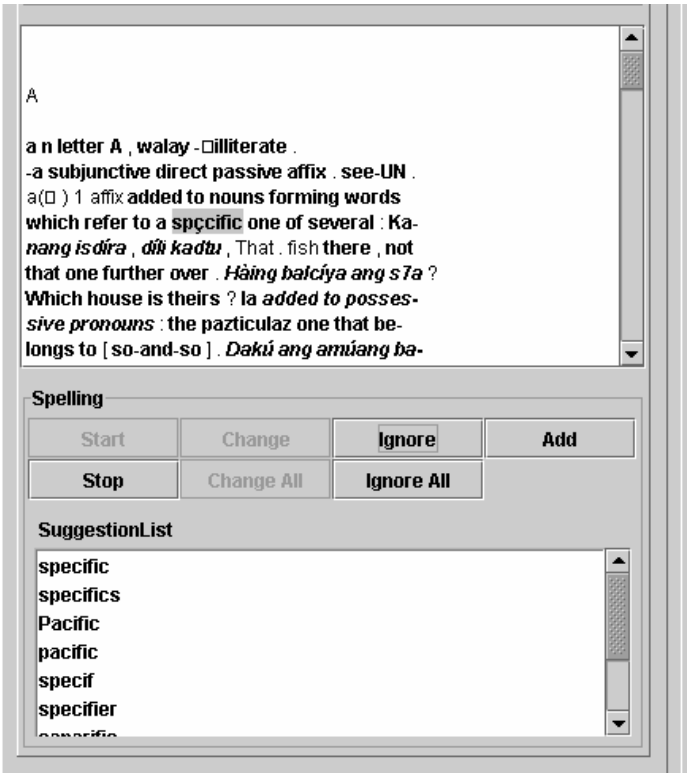


Figure 7-18 An incorrect word is highlighted

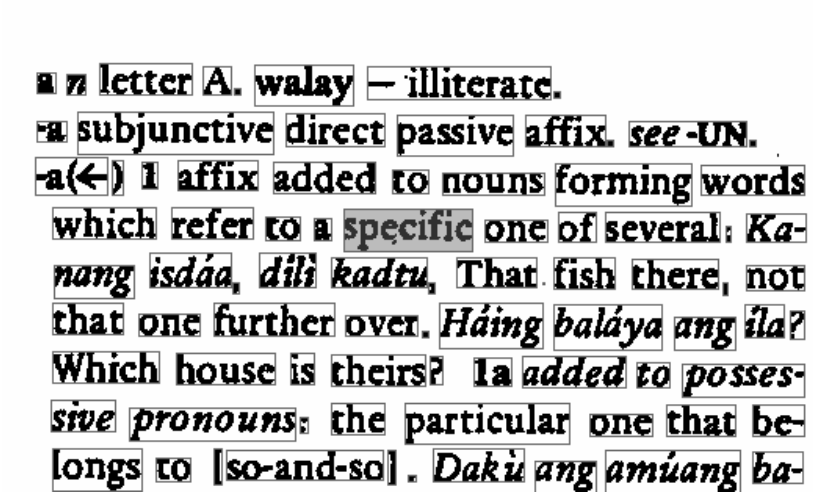


Figure 7-19 The correct word as displayed in the image

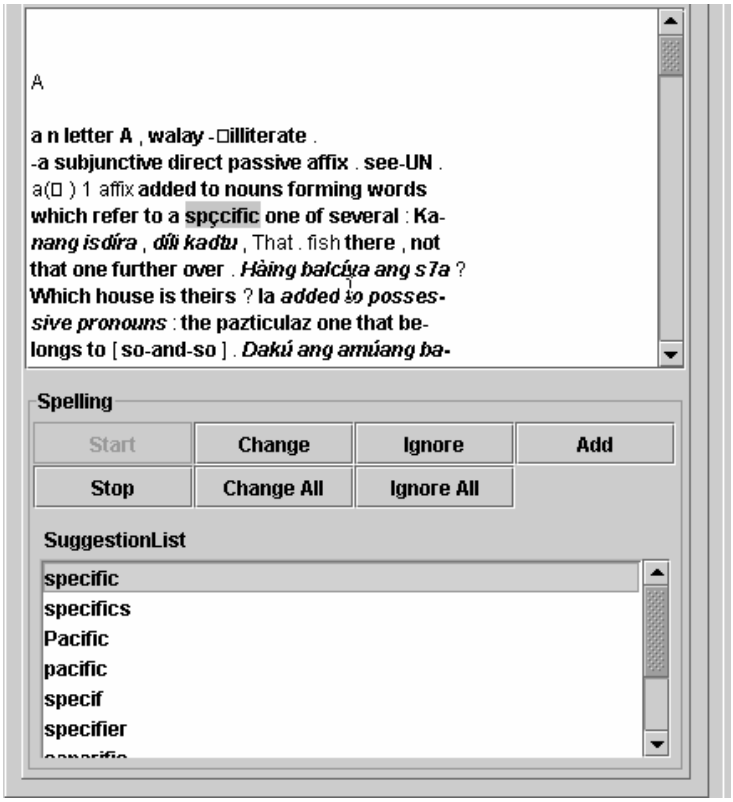


Figure 7-20 Selecting a correct substitution for the incorrectly extracted word

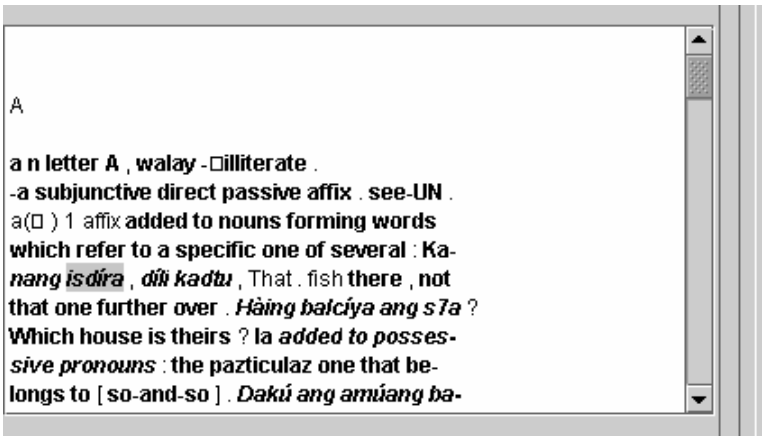


Figure 7-21 The correction is reflected in the extracted text

7.5 Segmentation

After performing OCR, the system is now ready for segmentation. Presence of OCR results is a prerequisite for running the segmentation process. First we will prepare the training samples for segmentation.

7.5.1 Preparing the training samples for segmentation

We will consider the case where we are in the Result display mode and we do not have the results available. As can be seen from the labeling panel, there are different types of zones and the entries can be classified accordingly into these zones.

To mark an entry in the image for training, first find out the zone type of that entry i.e. Regular, Unterminated etc. (as defined in the manual). Then for each such section you want to add to training, double click on the corresponding zone. This will activate the **Create** tool. Click on the **Create** tool and draw a box around the desired entry. Repeat this procedure until you have sufficient entries marked.

To selectively add these entries to training, click on the **Select** tool on the toolbar area. Then click on each entry you want to add to the training. To select multiple entries, press and hold down the SHIFT key while clicking on different entries. After the selection is made, click the **Selected** button in the Add Zones To Training area. You can also press the **All** button to add all the entries marked to training.

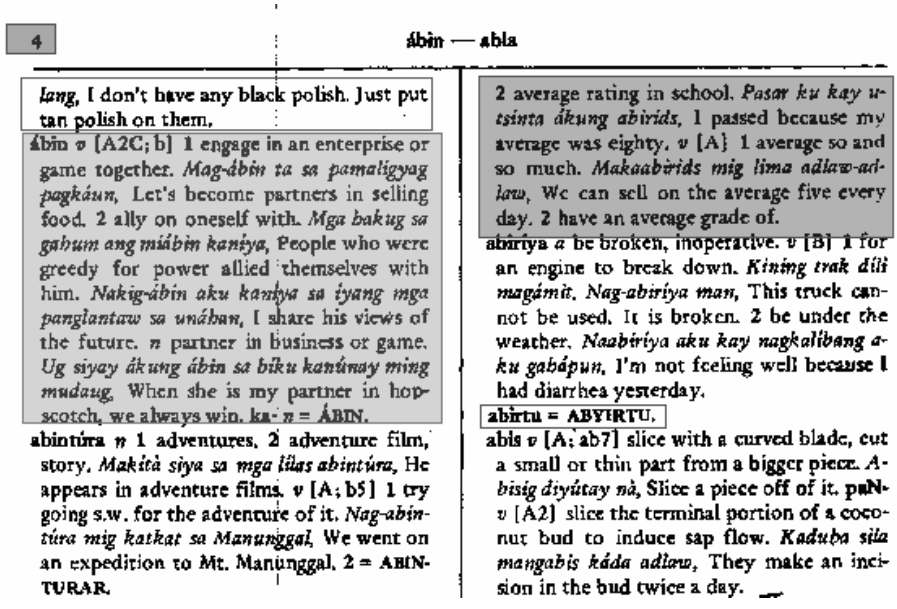


Figure 7-22 Selecting zones to add to training

Thus the operator can prepare training samples for their system. There are two other cases in which the operator can prepare the training samples 1. Working in the training display mode when the result is unavailable 2. Working in the result display mode when the result is available. They are described in detail in the manual.

7.5.2 Performing segmentation

To perform segmentation, click on the **Segment** button in the Workflow area of the Segmentation panel. This will bring up following dialog box.

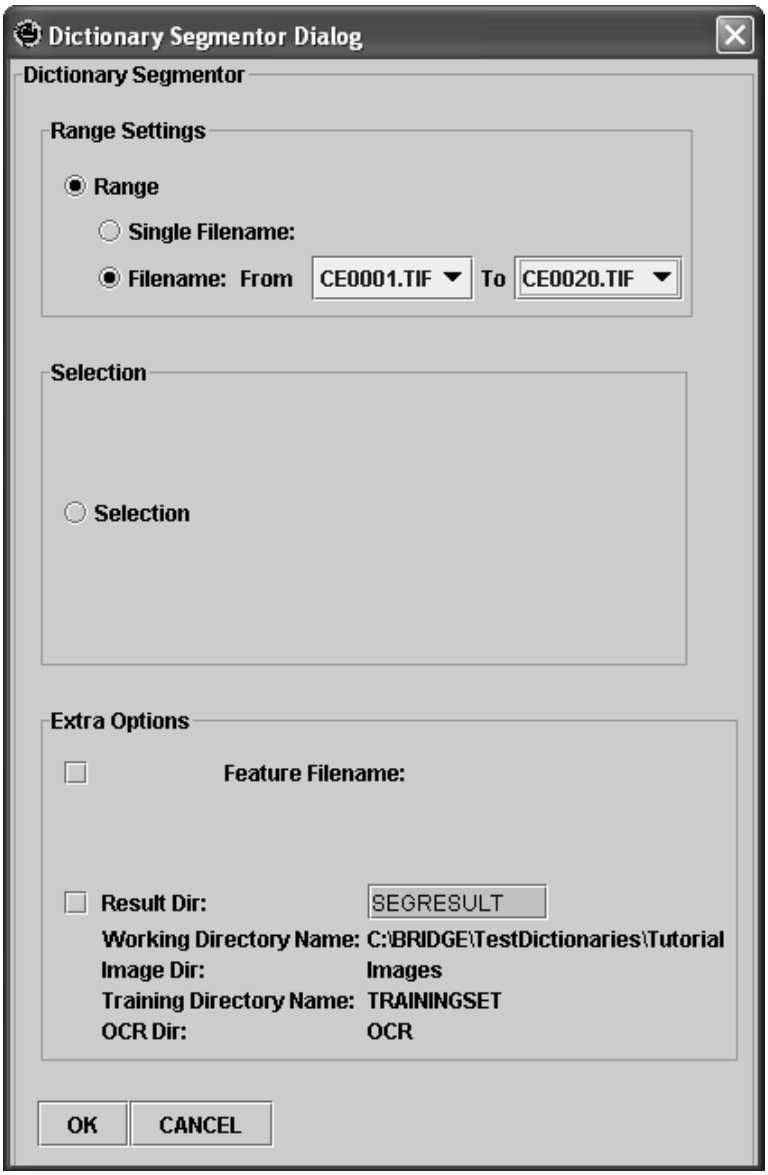


Figure 7-23 Performing Segmentation

As can be seen from the above figure the operator has multiple options with respect to the selection of files to perform segmentation on. For our example, we will select the pages in a range from ce0001.tif to ce0020.tif. The extra options are explained in the manual. For this example, we will accept the default settings.

After setting the configuration for segmentation, click the **OK** button to activate the segmentation. After the segmentation operation is complete, the presence of the result files will be indicated by a green check mark in the cells of the Result column (Figure 7-24).

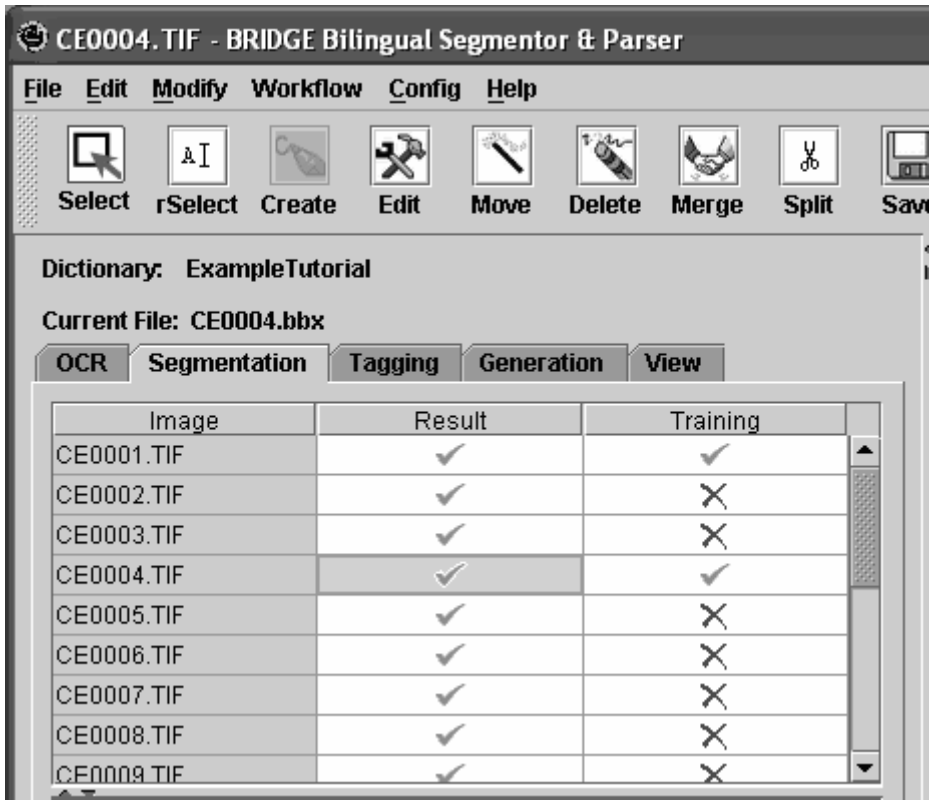


Figure 7-24 Completion of segmentation as indicated by the result files

If the green check marks do not appear, go to **File → Refresh Dictionary** and update the display.

After the segmentation is done successfully, we can now proceed to the tagging part. For the tagging procedure to execute smoothly, segmentation results must be obtained beforehand.

7.6 Tagging

Every dictionary will have different specifications with respect to the representation of the entries. The operator needs to study their structure in order to set up the correct configuration. For our example, we are using the configuration of an already created dictionary and copying the configuration from the source to the target dictionary for the tutorial example. For the purpose of understanding, we will explain how to prepare tagging configuration for an "ExampleDictionary" that we will create. Creation of this dictionary has been explained in section 7.2.1.

7.6.1 Preparing the configuration file from scratch

Click on the **Config → Config Utility** sub-menu to bring up the **Bridge Preferences** dialog box. Click on the Tagging tab to display the configuration dialog. In the Tagging tab, add the categories Headword, Pronunciation and POS from the list of right side to that of left side as shown in the following figure.

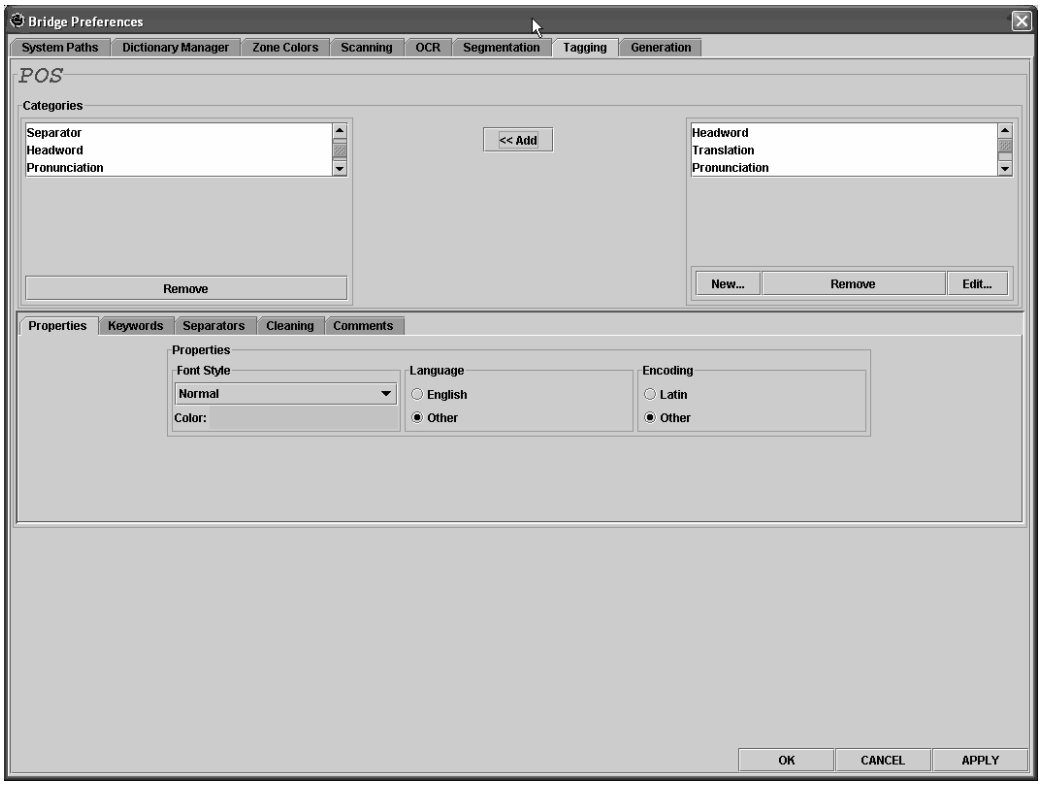


Figure 7-25 Preparing the configuration for tagging

For each of this category, we will set the attributes. Note that these settings differ from dictionary to dictionary and the operator needs to observe the dictionary specifications before setting these attributes.

For Headword, in the Properties panel, set font style as **bold** and color as “red”. Set the language as “other” and the encoding as “latin”. Click the APPLY button to save these changes and move to the Keywords panel. There are no keywords for headword. In the separators panel, we need two parameters. Select InPlaceOf as the operator and for this operator, select two separators “-“ and “~”. This is shown in the following figures. (Figure 7-26 and Figure 7-27)

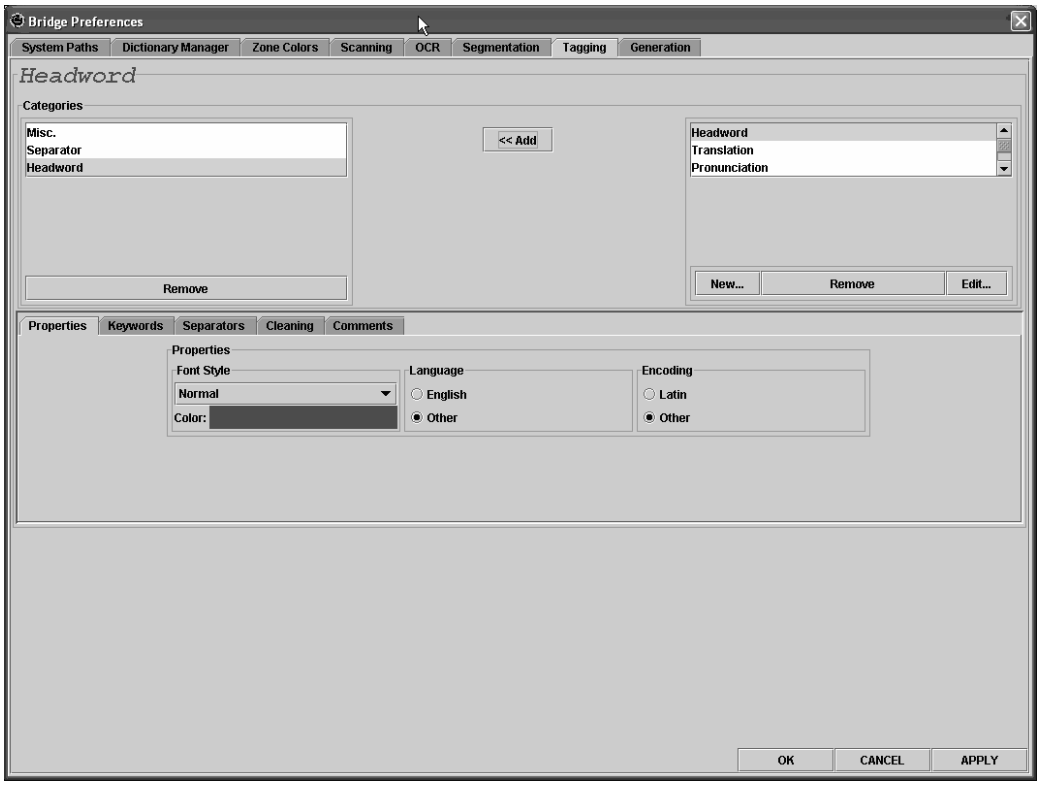


Figure 7-26 Headword attributes

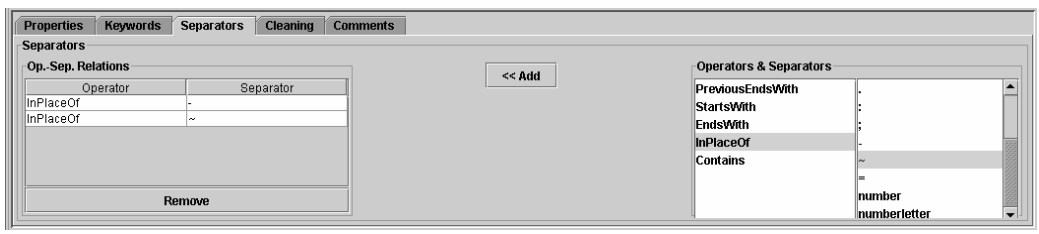


Figure 7-27 Specifying separator attribute for headword

In the generation process, we would like to remove numbers, punctuations and other special characters (e.g. "*") from the headwords. So in the Cleaning panel, select these options and add them to the list on the left side. The special character (*) that we will like to remove is not originally present in the list on the right side. So click on the **New** button and enter the symbol in the **Edit Dialog** and click **OK.** Now add the "*" symbol to the list on the left side. This is shown in the following figures.

Similarly for POS, we have added keywords for noun, verb and adjective. We have also specified properties for pronunciation.

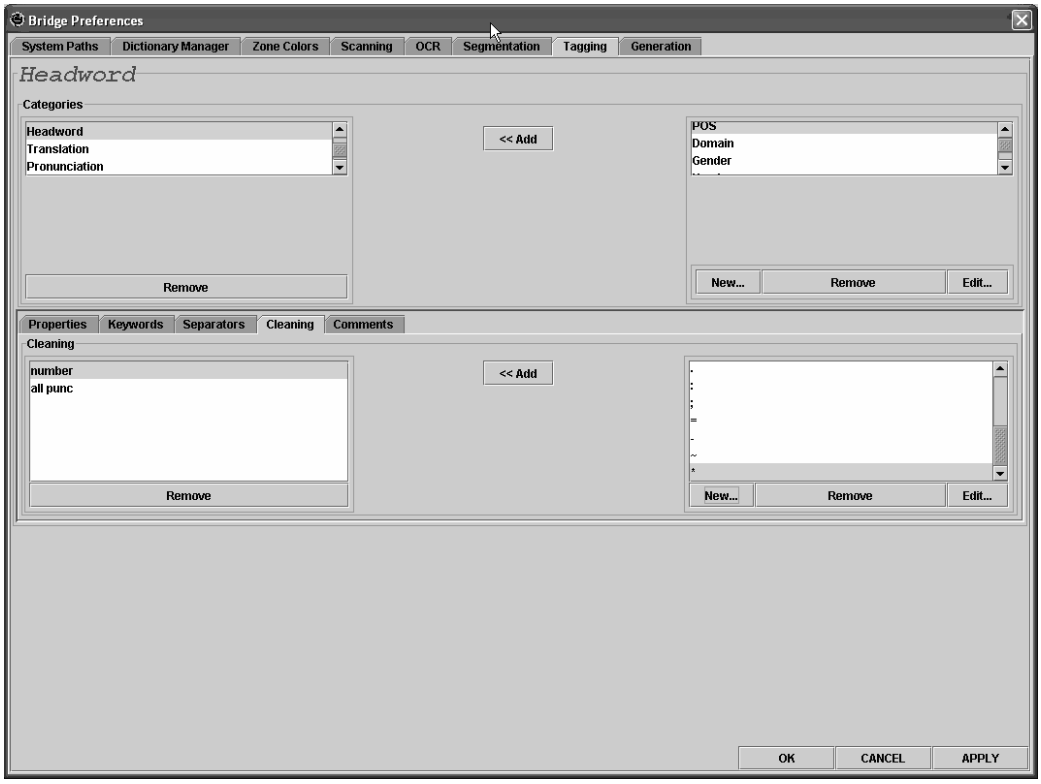


Figure 7-28 Cleaning panel for Headword

Figure 7-29 Adding a new character to the cleaning panel

7.6.2 Copying the configuration from another dictionary

For the purpose of this tutorial, we are using a dictionary that was created by copying the configuration of another dictionary. So this will also copy the configuration for tagging. To examine the tagging configuration parameters for the test dictionary, make sure that the test dictionary is open. Then open the Tagging panel in the Bridge Preferences dialog box and check different attributes.

7.6.3 Performing Tagging

In the Workflow sub-panel of the Tagging panel, click on the **Tag** button. It will pop up the following dialog box.

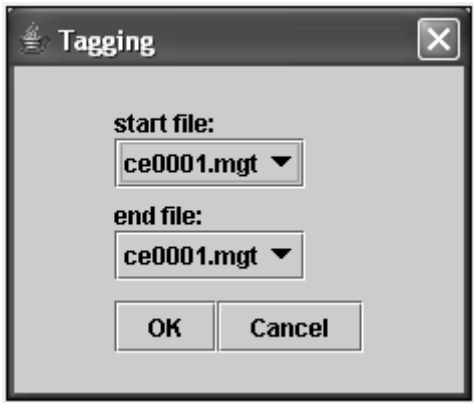


Figure 7-30 Setting files for tagging

In the end file field, select the file name as ce0020.mgt to perform tagging on the entire batch and click the **OK** button.

This will start the tagging procedure. After the tagging process is completed successfully, you can see the green check marks in the Result column of the Tagging panel as shown in the following figure.

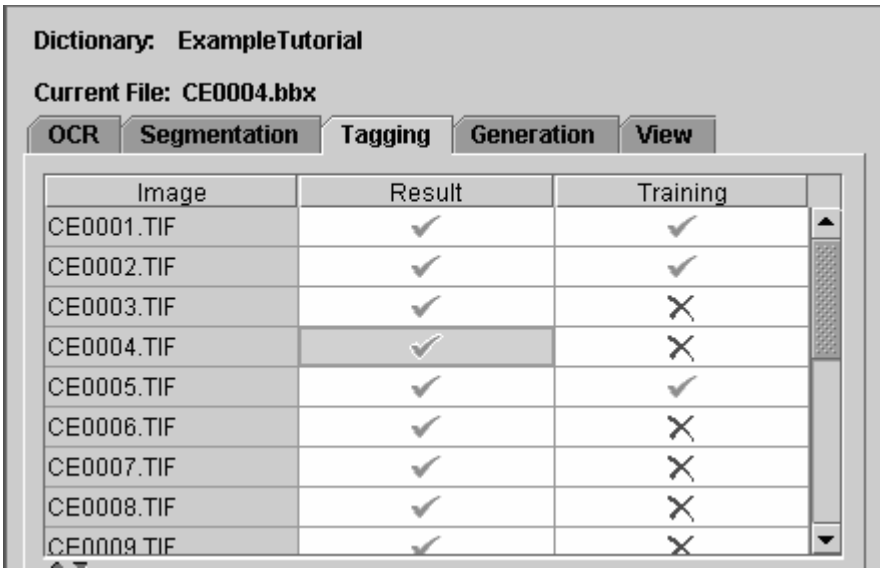


Figure 7-31 Completion of tagging process as indicated by the presence of result files

After tagging, the entries in the image files will have bounding boxes around them indicating their tags as shown in the following figure.

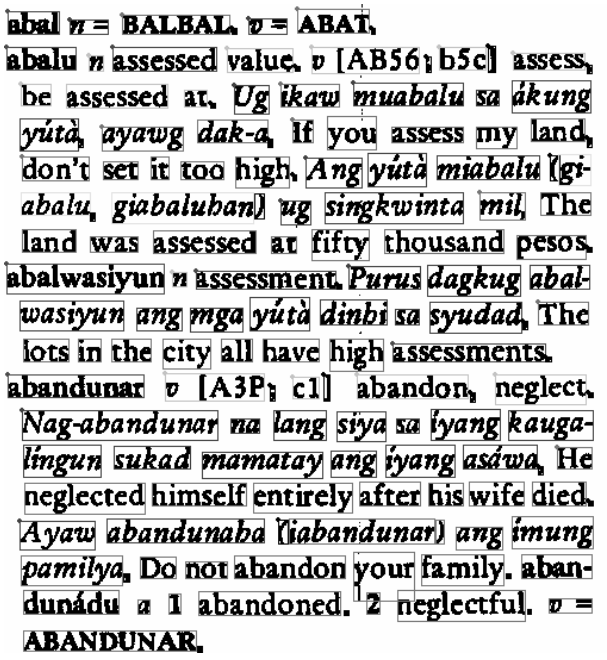


Figure 7-32 Tagged image entries

7.6.4 Preparation of training samples for tagging

As explained in the manual, the purpose of preparing the training samples for tagging is not to train the system but to evaluate the data. The evaluation module will be used in future work. We will now explain how to prepare the training samples for tagging when in result display mode. The prerequisite for this procedure is that tagging results for this page should be available.

Open the page that you want to prepare training samples for. This page will be displayed in the right side panel of the interface.

Click on the **rSelect** tool to activate it. Then with the left button of the mouse pressed, drag the mouse over the region of whose zones you want to add to training. The selected zones will be highlighted. In the Add Zones to Training area, click on the button **Selected** to add the highlighted zones to training. Save all the selected entries by clicking the **Save** tool on the toolbar.

Since we have the tagging results available, we can now proceed to the generation part.

7.7 Generation

In the Generation panel, click on the **Generate** button to bring up the **Generation** dialog box. For this example purpose, we will generate the results in the Rosetta XML format. Enter the dictionary ID as 3 letters followed by 4 numbers format e.g. ceb0001. Now the generation dialog box will appear as shown in the following figure.

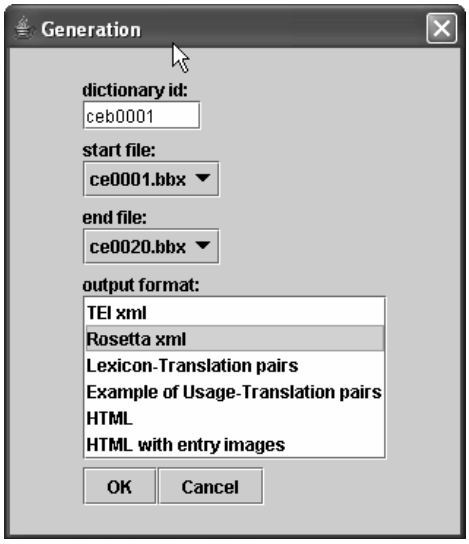


Figure 7-33 Setting parameters for generation

Select the page ranges you want to generate from ce0001.bbx to ce0020.bbx. Select the Rosetta xml option from the list and click the **OK** button. Now in the BRIDGE interface, you can see the output of the generation process as shown in the following figure.

We also have generation results in TEI XML, HTML and HTML with images format.

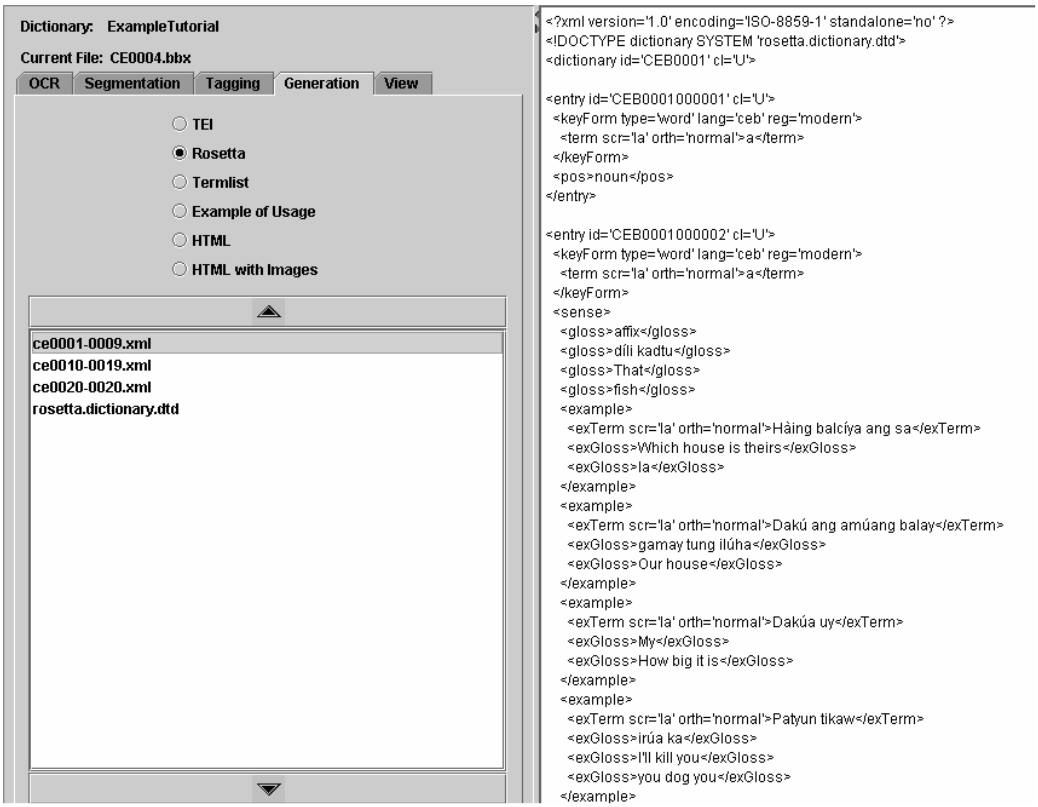


Figure 7-34 Output of Generation in Rosetta format

7.8 The BRIDGE-View (Search and Retrieval)

As explained in the manual, this module helps the user locate an actual term or an entry in a dictionary. XML files in TEI format are used as input for the search and retrieval operation.

7.8.1 Searching for a Query

In the View panel of the Bridge interface, click on the **Create View** button which is placed in the Workflow area. This produces a new set of XML files from the XML files created in the generation process. After this operation is run, click on

the **Run View** button to display the BRIDGE-View interface as shown in the following figure.

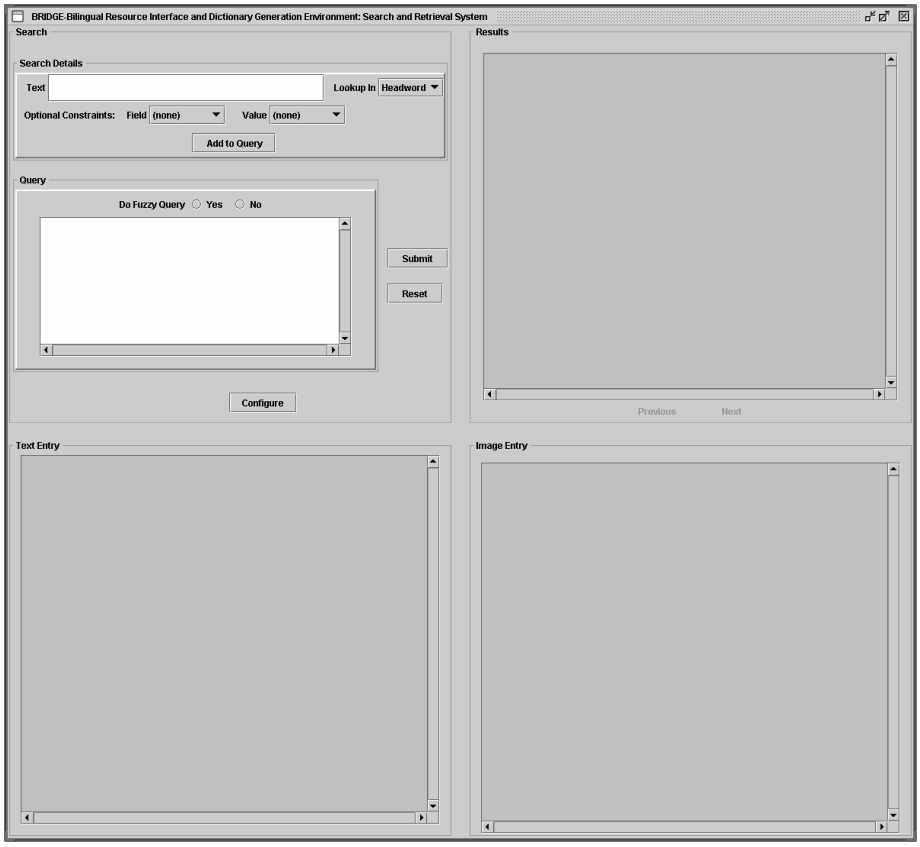


Figure 7-35 The BRIDGE-View interface

The Search panel has the input field for entering the query. It also provides various options to set the parameters to perform the search. Suppose we want to search for the word abirtu. Enter the query in the text field. Set the option Lookup In as Headword. Click the button **Add to Query**. We will show the results of both fuzzy and non-fuzzy query. Select the radio button Yes for the option "Do Fuzzy Query". You can also select the optional constraints Field and Value to restrict the search. The corresponding interface is shown in the following figure.

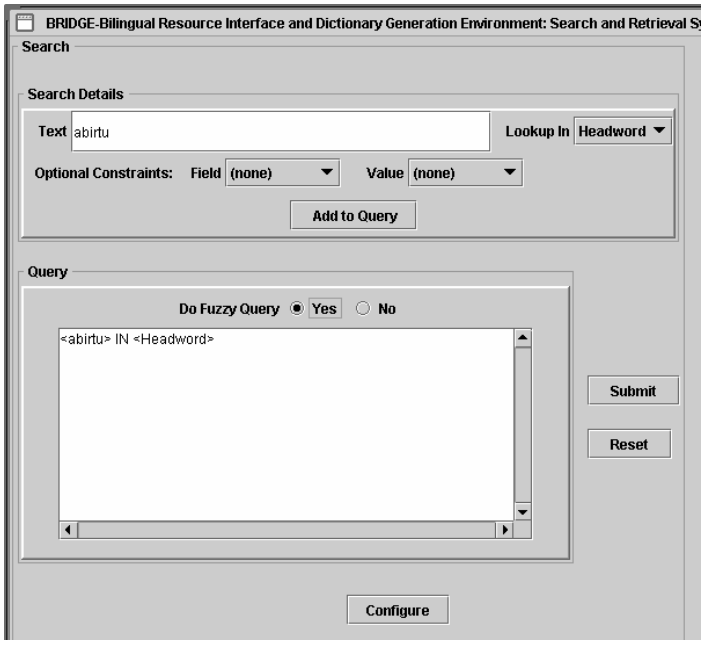


Figure 7-36 Performing search using different options

7.8.2 Retrieving results for the query

Click on the **Submit** button. In the dialog box that will pop up, select Yes to submit the query. The results will be displayed in the Results panel as shown in the following figure.

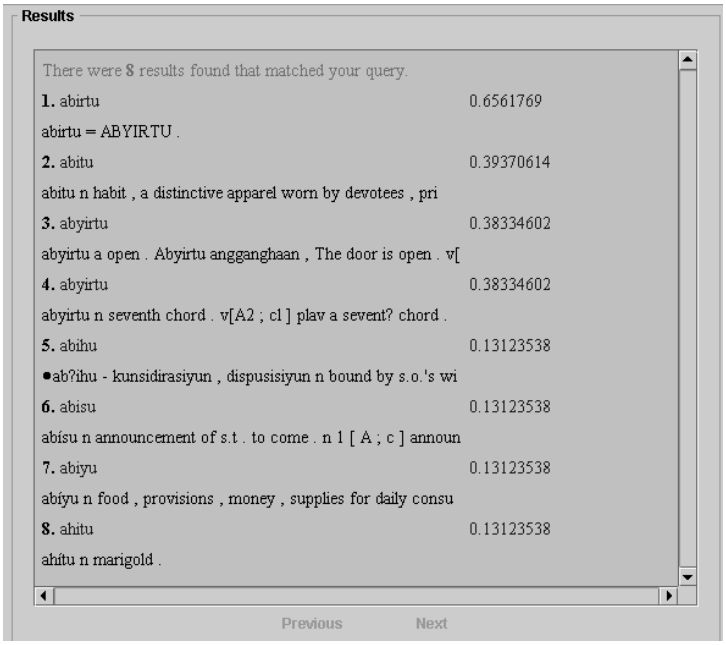


Figure 7-37 Results for the test query using Fuzzy Query

Now for the option Do Fuzzy Query, select the option No and submit the query again. The result of this operation is shown in the following figure. We can clearly see the difference between the number of results generated using fuzzy query and without using fuzzy query.

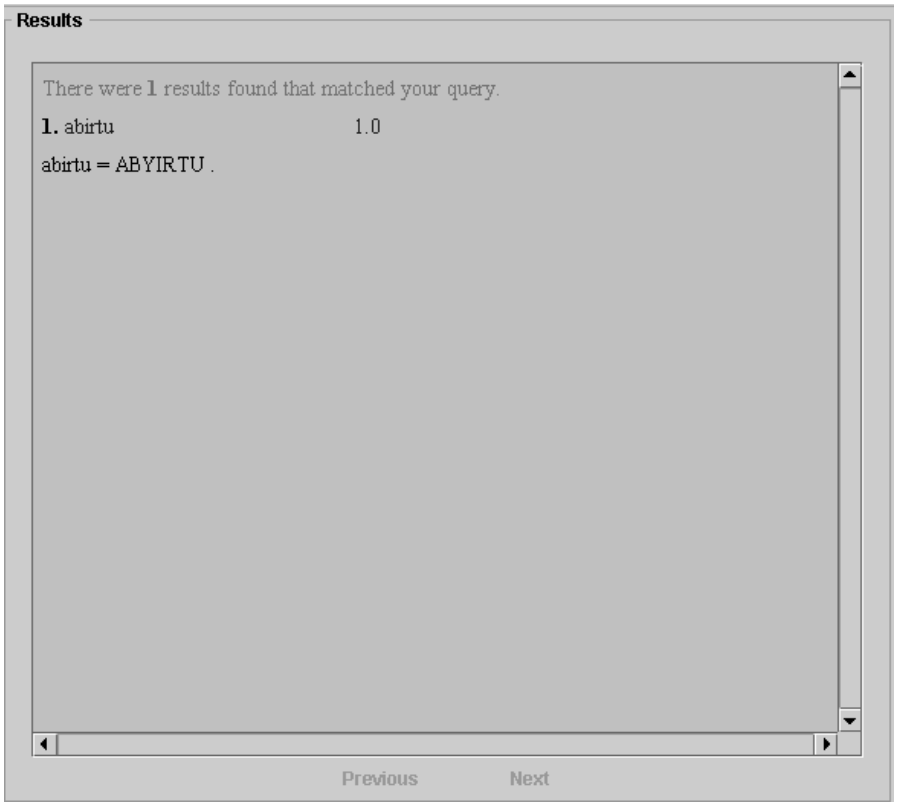


Figure 7-38 Results for the test query without Fuzzy Query

The headword portion of the result serves as the hyperlink for the text and image representations of the result in the dictionary. Click on the test query “abirtu” and observe the output in the Text Entry and Image Entry panels.

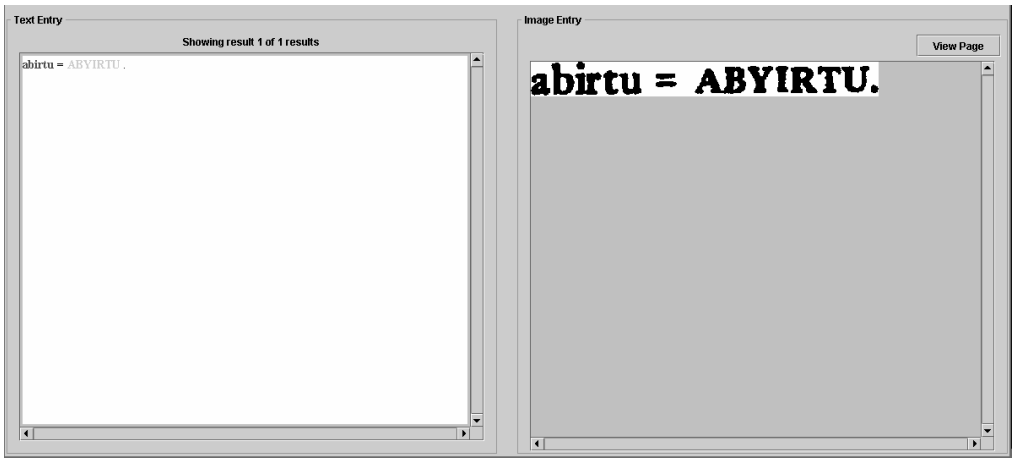


Figure 7-39 Text Entry and Image Entry results for the query

The interface also has the functionality to display the original dictionary page that contains the query. Click on the **View Page** button in the Image Entry panel. This will display following dictionary page.

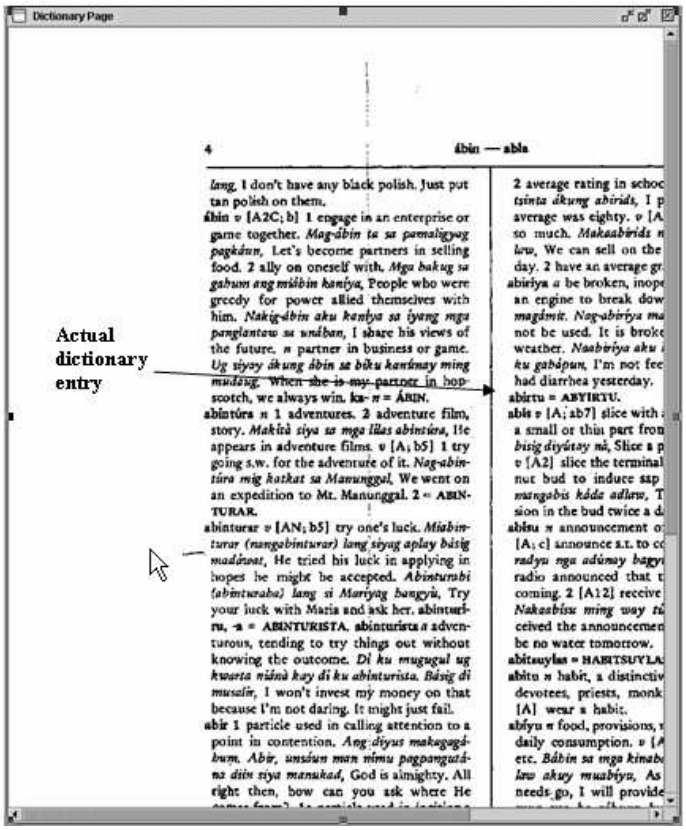


Figure 7-40 Actual dicitonary page containing the query

For better understanding of the interface, we have provided below the results for the query “agiik” as represented in the Text and the Image Entry panels. As shown in the figure, the Text entry panel shows different linguistic parts of the entry in different colors.

Figure 7-41 Text and Image Entry results for another query.

7.9 Summary

In this section we have demonstrated the working of our Bilingual Resource Inference Dictionary Generation Environment (BRIDGE) System using the Cebuano dictionary. We started with scanning and OCR and then we covered segmentation, tagging and generation. We also demonstrated the Search and Retrieval procedure on generation results. It is to be noted that these procedures are interdependent for optimum results. So at each step, the results should be checked for accuracy. It is important to prepare proper training sets for segmentation and to configure the system for tagging. Every dictionary will have different structure. So it should be studied before the tagging configuration.

8 Appendices

8.1 Troubleshooting

8.2 Shortcuts and Key Bindings

For easy access to the interface via keyboard, menus are accompanied by keyboard shortcuts or key bindings also called as mnemonics. A list of the key bindings is as follows.

Menu	Sub-menu	Keybinding
File		Alt + F
	Open a dictionary	Alt + O
	Save	Alt + S
	Delete	Alt + X
	Close	Alt + C
	View log file	F5
	Exit	Alt + F4
Edit		Alt + E
	Undo	Ctrl + Z
Modify		Alt + M
	Create	Ctrl + C
	Edit	Ctrl + E
	Move	Ctrl + M
	Delete	Ctrl + D
	Merge	Ctrl + R
	Split	Ctrl + S
Processing		Alt + P
Config		Alt + C
Help		Alt + H
	Help	F1
	Overview	F2
	Future Work	F3

Table 8-1 Keyboard shortcuts

8.3 File Formats

8.4 Glossary

8.5 Publications

Journal

- H. Ma, B. Karagol-Ayan, D. Doermann, D. Oard, and J. Wang. Parsing and Tagging of Bilingual Dictionaries. *Traitement Automatique Des Langues*, 44(2):125–150, 2003.
- Huanfeng Ma, and David Doermann. Adaptive Hindi OCR Using Generalized Hausdorff Image Comparison. *ACM Transactions on Asian Language Information Processing*, 2(3):193–218, 2003.

Conference

- Huanfeng Ma, and David Doermann. Word Level Script Identification on Scanned Document Images. *SPIE Conference on Document Recognition and Retrieval*, pages 178-191, 2004
- H. Ma and D. Doermann. Adaptive word style classification using a Gaussian mixture model. In 17th International Conference on Pattern Recognition (ICPR), Cambridge, United Kingdom, 2004. to appear.
- H. Ma, and D. Doermann. Gabor Filter Based Multi-class Classifier for Scanned Document Images. *7th International Conference on Document Analysis and Recognition (ICDAR)*, pages 968-972, 2003 .
- H. Ma, B. Karagol-Ayan, and D. Doermann. Segmenting and Tagging Structured Content. *Symposium on Document Image Understanding Technology*, pages 53-64, APR 2003
- B. Karagol-Ayan, D. Doermann, and B. Dorr. Acquisition of Bilingual MT Lexicons from OCRed Dictionaries. Proceedings of the MT Summit, New Orleans, LA, pp. 208--215, 2003.
- H. Ma, and D. Doermann. Bootstrapping Structured Page Segmentation. *SPIE Conference Document Recognition and Retrieval*, pages 179-188, JAN 2003
- D. Doermann, H. Ma, B. Karagol-Ayan, and D. Oard. Lexicon Acquisition from Bilingual Dictionaries. *SPIE Photonic West Electronic Imaging Conference*, pages 37-48, 2002 .