



DICTIONARYMAKER TUTORIAL FOR
VERSION 2.16

M Tempest and B Khambane

12 October 2009

Table of Contents

1 OVERVIEW.....	3
1.1 BEFORE GETTING STARTED.....	3
2 CREATING A NEW PROJECT.....	3
2.1 NAME AND LOCATION.....	3
2.2 GRAPHEMES.....	4
2.3 PHONEMES.....	5
2.4 WORD LIST AND DICTIONARY.....	7
2.4.1 Starting with only a word list.....	7
2.4.2 Starting with only a dictionary.....	8
2.4.3 Starting with both a word list and a dictionary.....	9
3 SAVING A PROJECT.....	10
4 OPENING AN EXISTING PROJECT.....	11
5 USING DICTIONARYMAKER.....	11
5.1 PROVIDING A VERDICT.....	12
5.2 UPDATING THE PRONUNCIATION.....	14
5.3 VIEWING THE WORD LIST.....	14
5.4 DISPLAYING THE STATUS.....	14
6 ADVANCED USE.....	15
6.1 SELECTING DIFFERENT VIEWS OF THE WORD LIST.....	15
6.2 CHANGING THE SYSTEM DEFAULTS.....	15
6.3 CHANGING THE PHONEME PANEL SETTINGS.....	16

1 Overview

The DictionaryMaker Manual describes the system's concepts, installation and use. This tutorial is intended to be a quick-start guide, using a Setswana pronunciation dictionary as an example.

1.1 Before getting started

In the folder created when DictionaryMaker was installed there is a folder named **DMTutorial**. We will be creating the new dictionary from data files stored in that folder. (If you want to create a project from your own data files, please refer to the DictionaryMaker Manual for the required data formats.)

To create a new dictionary we need to create a new project that will contain the word list, graphemes and phonemes used in the new dictionary, as well as an initial bootstrapping dictionary.

A DictionaryMaker project requires you to have the following files available for your target language, and these are provided for the Setswana tutorial in [DMTutorial/Data](#):

Phoneme file – [Setswana.pho](#)

Grapheme – [Setswana.gra](#)

Word list – [Setswana.wdl](#)

Sound files in wav format, all in [sounds/](#)

An initial dictionary is not required, but is supplied in the Data folder, and will be used later.

Dictionary file – [Setswana.initial.dict](#)

2 Creating a new project

2.1 Name and location

Select the New Project option found under the File Menu. This will open the following pane:

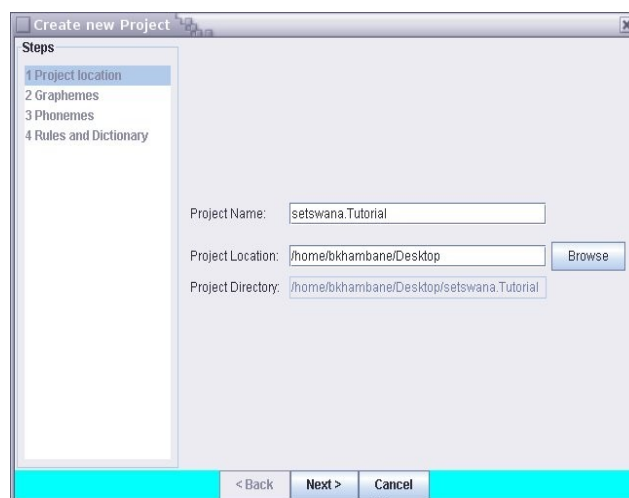


Figure 1: Creating a new project – name and location

To create a new project you need to decide on a name and location.

Fill in the *Project Name*, e.g. *Tutorial*. The name should not have been used previously.

Fill in or browse on your computer for the *Project Location*, e.g. <Where DictionaryMaker is installed>/DMTutorial/.

The *Project Directory* displays the resulting location of the new project. The *Tutorial* folder is created automatically, and the Tutorial project's files will be put in it. The system will automatically create the following files in the Tutorial folder:

Project – *Tutorial.proj*

Log – *Tutorial.log*

The Next> button advances to the next step of project creation – defining the graphemes that the new dictionary will use.

2.2 Graphemes

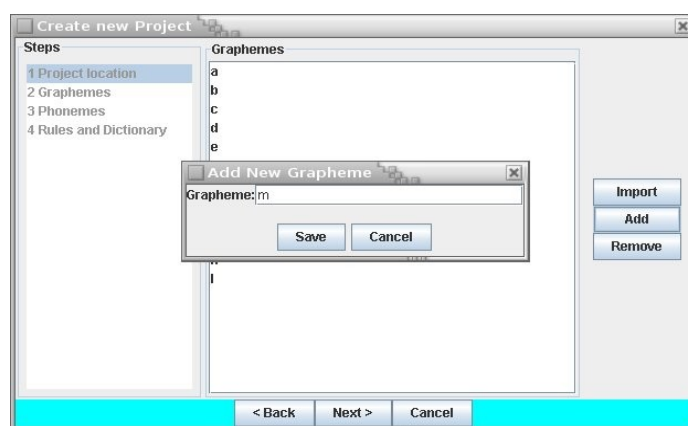


Figure 2: Adding graphemes

You can add and remove graphemes individually, or import a list of graphemes from a predefined grapheme text file. Note that this text file must contain only one entry on each line. *Setswana.gra* can be imported from the *Data/* folder as shown in figure 3.

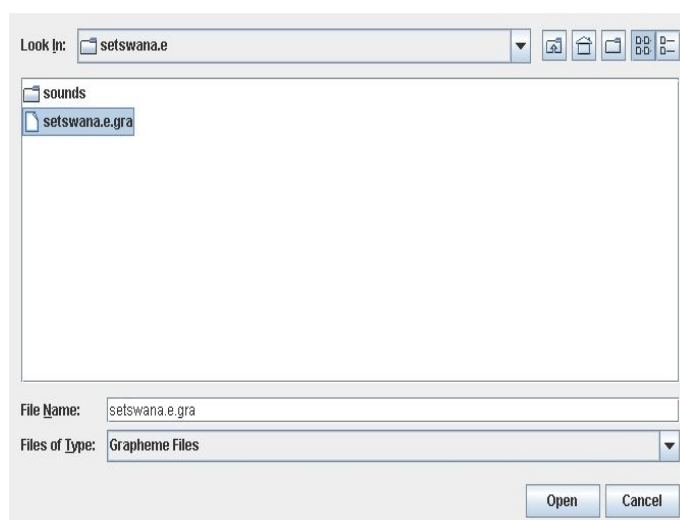


Figure 3: Creating a new project – importing graphemes

The Next> button now advances to the next step of project creation – defining the phonemes that the new dictionary will use.

2.3 Phonemes

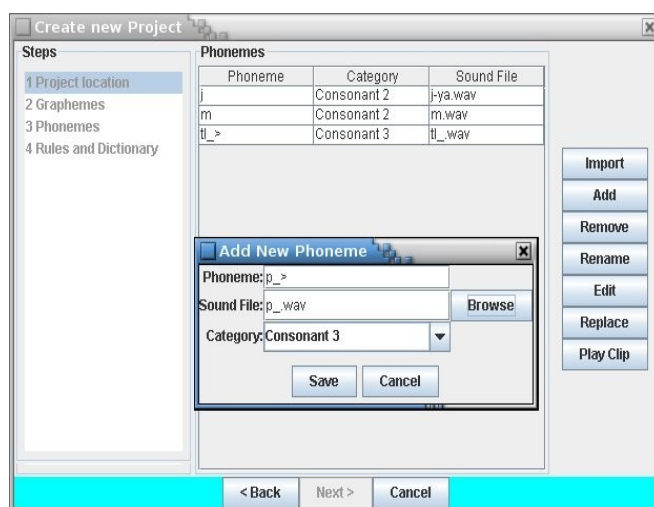


Figure 4: Adding phonemes

You can add, edit and remove phonemes individually – each phoneme is associated with a sound file, and classified into a category.

Alternatively, you can import a list of phonemes from a predefined phoneme text file.

Note that this text file must contain only one phoneme entry on each line. *Setswana.pho* can be imported from the *Data/* folder.

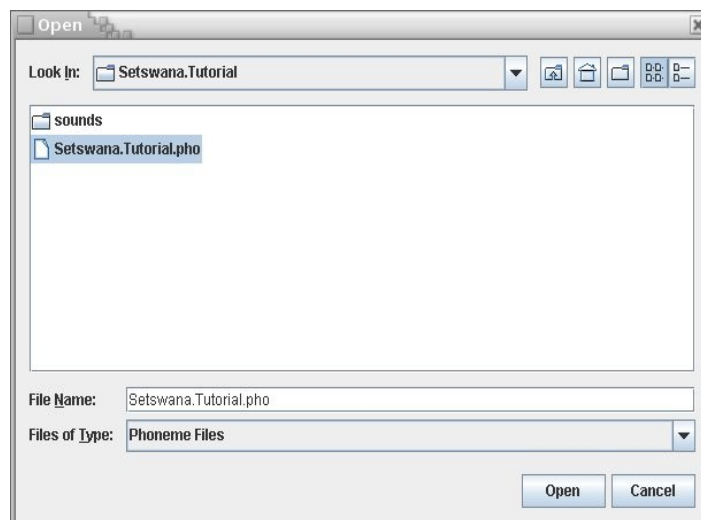


Figure 5: Importing phonemes

The phoneme text file has the following format:

```
<phoneme symbol><tab><sound file><tab><category>
```

For example:

a a.way vowels

ch ch.wav consonants

In each phoneme file, one of following phoneme sets can be used:

- IPA
- SAMPA
- HTK

The supported subset of the IPA phoneset is:

tɸ, tʰ, tᵏ, t͡ʃ, t͡ʃʰ, t͡ʃᵏ, tsʰ, tsᵏ, kxʰ, kxᵏ, kt̪, psʰ, psᵏ, p͡fᵏ, p͡fʰ, d͡ʒ, dz, dβ, d͡z, d͡ʒᵏ, d͡ʒᵏ, f, v, θ, ð, s, z, ʃ, ʒ, x, h, ᶱ, ʈ, ʈ͡ʂ, d͡t̪, d͡t̪͡ʂ, ɣ, β, ɸ, ɕ, sw, zw, ɸs, ɸS, β͡ʒ, ʃ͡, , , l͡g_i, lʰ, , g_i, ᵏ, !, !g_i, !ᵏ, p, pᵏ, pʰ, pʲ, p͡jᵏ, b, ḃ, by, t, tᵏ, tʰ, t͡ʃ, t͡ʃᵏ, d, dj, dᵏ, cʰ, cᵏ, j, k, kᵏ, kʰ, kʷ, g, gᵏ, t͡lʰ, t͡lᵏ, m, n, ŋ, ŋ, ŋ, mᵏ, nᵏ, ŋ, ŋᵏ, iː, y, u, uː, e, eː, øː, ɛ, ɜː, ɔ, ɔː, a, aː, ə, æ, o, oː, œ, ɒ, əi, œy, ai, ɔi, əu, au, iə, eə, uə, r, rᵏ, l, j, w,
--

The supported subset of the SAMPA phoneset is:

tp\, tK_>, tK_h, tS, tS_>, tS_h, ts_>, ts_h, kx_>, kx_h, kK_>, ps_>, ps_h, pS_h, pS_>, d_0Z, dz, dB, dz', dz'h\, dZh\, v, T, D, s, z, S, Z, x, h, h\, , K, K\, dK, dK\, G, B, p\, s', sw, zw, p\,s, p\,S, BZ, fS, \, , \,g_0, \,h , \,\, , \,\,g_0, \,\,h, \, , \,g_0, \,h, p, p_h, p_>, pj_>, pj_h, b, b_<, bj, t, t_h, t_>, tj, tj_h, d, dj, dh\, c_>, c_h, J\, k, k_h, k_>, g, gh\, tl_>, tl_h, m, n, J, N, n', m_h, n_h, m_j, J_h, , i:, y, u, u:, e, e:, 2:, E, 3:, O, O:, a, A:, @, {, o, o:, 9, Q, @i, 9y, ai, Oi, @u, au, i@, e@, u@, r, rh\, l, j, w, l',

The supported subset of the HTK phoneset is:

tp_b, tK_>, tK_h, tS, tS_>, tS_h, ts_>, ts_h, kx_>, kx_h, kK_>, ps_>, ps_h, pS_h, pS_>, d_0Z, dz, dB, dz_a, dz_ah_b, dZh_b, v, T, D, s, z, S, Z, x, h, h_b, K, K_b, dK, dK_b, G, B, p_b, s_a, sw, zw, p_bs, p_bS, BZ, fS, |_b, |_bg, |_bh, |_b|_b, |_b|_bg, |_b|_bh, !_b, !_bg, !_bh, p, p_h, p_>, pj_>, pj_h, b, b_<, bj, t, t_h, t_>, tj, tj_h, d, dj, dh_b, c_>, c_h, J_b, k, k_h, k_>, g, gh_b, tl_>, tl_h, m, n, J, N, n_a, m_h, n_h, m_j, J_h, , i:, y, u, u:, e, e:, eu_:, E, E_:, O, O:, a, A:, @, {, o, o:, u_, Q, @i, u_y, ai, Oi, @u, au, i@, e@, u@, r, rh_b, l, j, w, l_a

All sound files referenced in the phoneme file, should be placed in a sounds/ directory. A maximum of four categories are used in each phoneme file. The available categories are:

- Vowels
- Diphtongs
- Consonants 1 – Affricates and fricatives
- Consonants 2 – Nasals, trills and flaps, approximants
- Consonants 3 – Clicks and stops

The Next> button now advances to the next step of project creation – importing the word list and dictionary which are used to generate the g-to-p (grapheme to phoneme) rules.

2.4 Word list and dictionary

A project requires either a word list, or an initial dictionary, or both. The word list defines the words that will be used. Importing an initial dictionary will provide rules that can be used to predict pronunciations for words in the word list. Since the initial dictionary is typically small, the predictions will not be very accurate, but will improve through the bootstrapping verification process.

If no dictionary is imported, there are initially no g-to-p prediction rules available, and therefore no initial pronunciation predictions.

2.4.1 Starting with only a word list

To import only a word list, check the Import Word List box only, and click on Next> to move to the import screen.

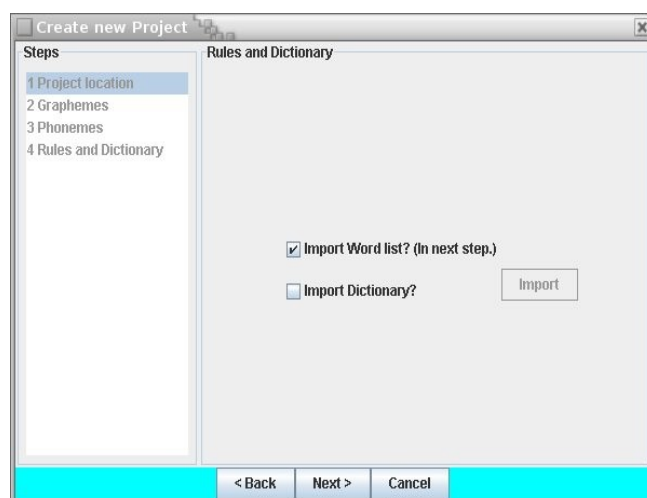


Figure 6: Creating a new project – choosing to import a word list only

You can add and remove words individually, or import a word list from a predefined word list text file. Note that this text file must contain only one entry on each line. `Setswana.wdl` can be imported from the `Data/` folder as shown in figure 3.

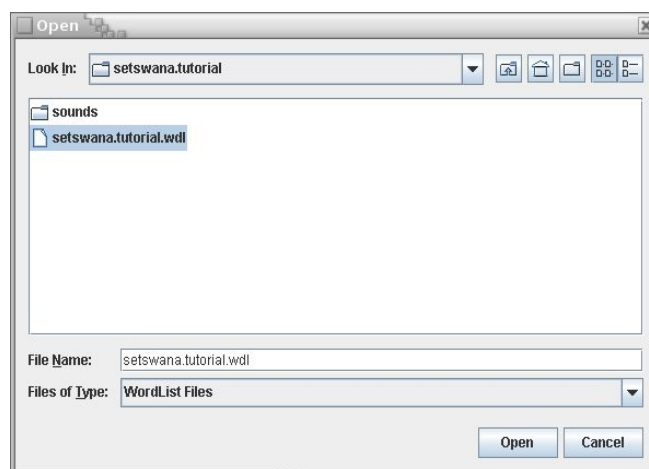


Figure 7: Creating a new project – importing a word list

2.4.2 Starting with only a dictionary

To import only a dictionary, check the Import Dictionary box, and click the Import button.

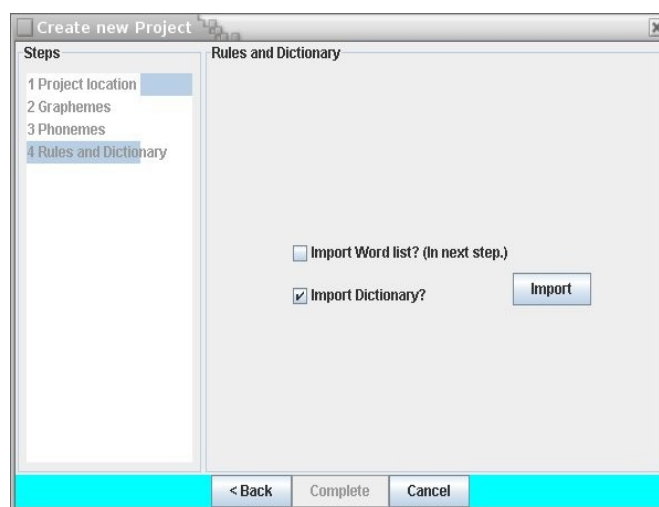


Figure 8: Creating a new project – choosing to import a dictionary only

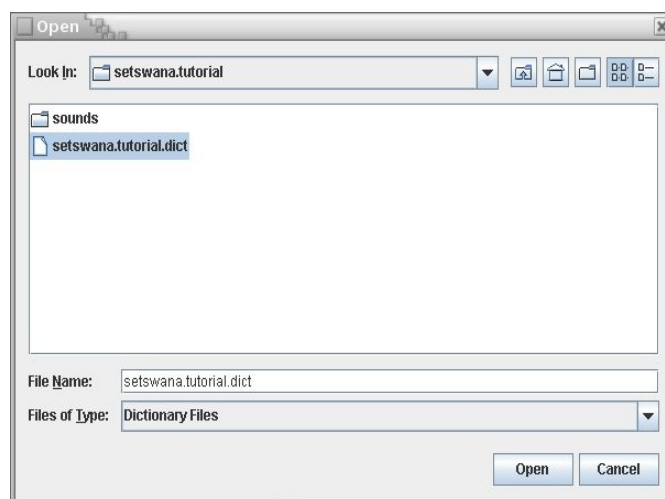


Figure 9: Creating a new project – importing a dictionary

`Setswana.initial.wdl` can be imported from the `Data/` folder as shown in figure 3. This dictionary contains correct pronunciations for a number of Setswana words.

If the dictionary import was successful, "Ok" is displayed.

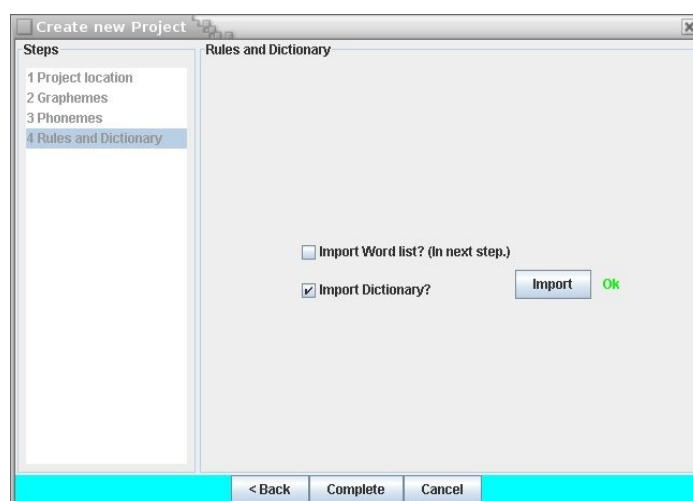


Figure 10: Dictionary import successful – now import word list

Now click on Complete to finish creating the project.

2.4.3 Starting with both a word list and a dictionary

To import both an initial dictionary and a word list, click both check-boxes, click on Import to import the dictionary, and then on Next> to import a word list.

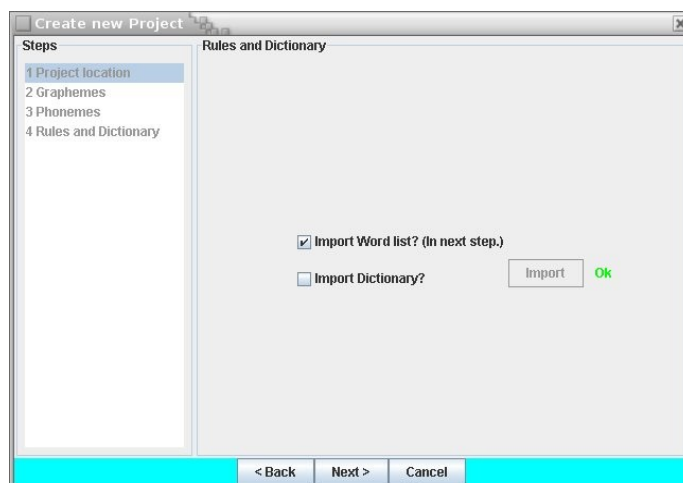


Figure 11: Creating a new project – importing a word list

The words in the initial dictionary have already been added to the word list. You can add and remove words individually, or import a word list.

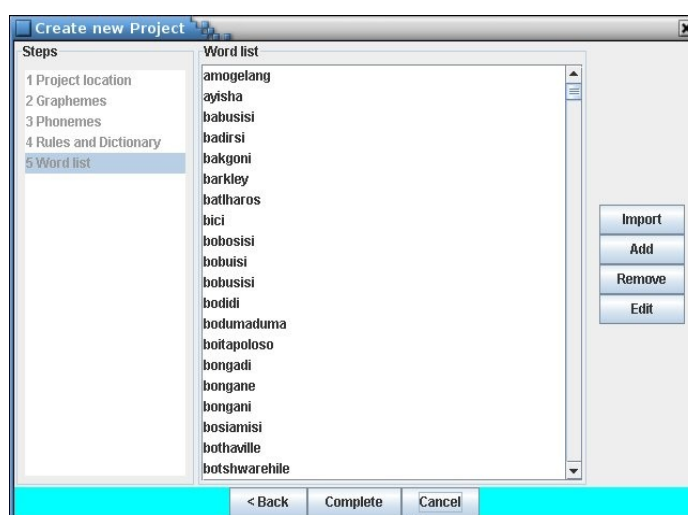


Figure 12: Importing an additional word list

When the word list is as you want it, click on Complete to finish creating the project.

You can now save your work before closing DictionaryMaker, or you can start verifying words.

3 Saving a project

DictionaryMaker currently does not automatically save projects. Projects are saved through the File-> Save menu option.

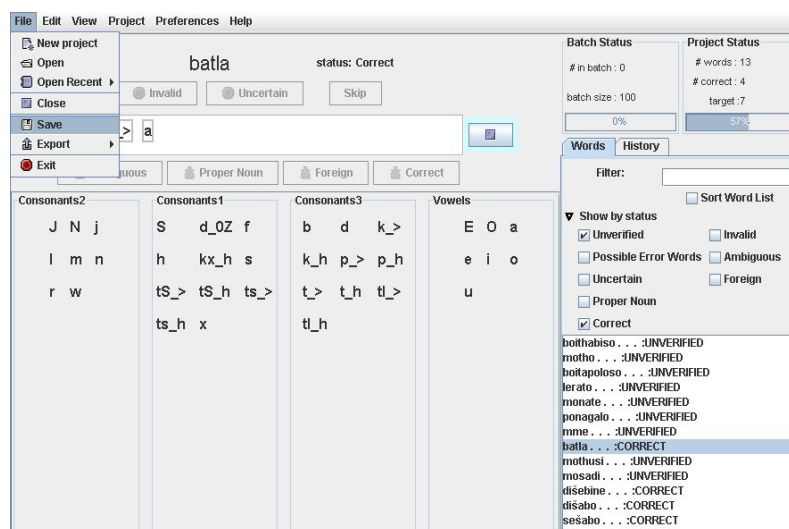


Figure 13: Saving a project

4 Opening an existing project

Existing projects are opened through the menu option File-> Open.

Recent projects can quickly be opened through the menu option File-> Open Recent.

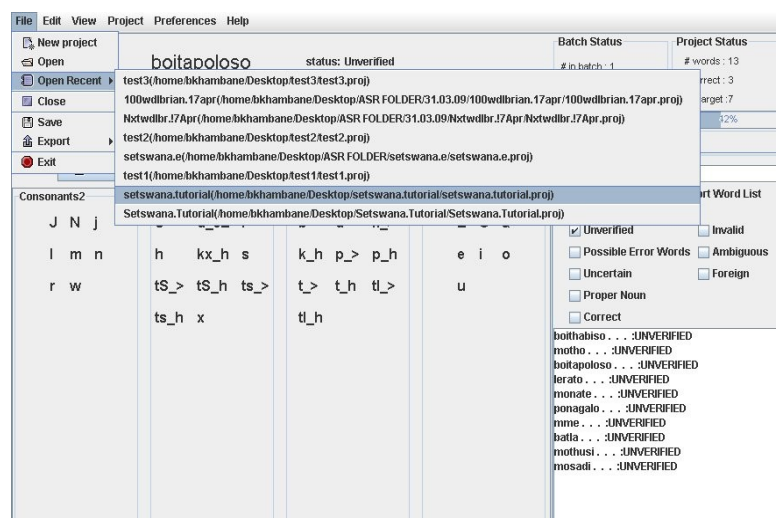


Figure 14: Opening a recent project

5 Using DictionaryMaker

The system runs through the word list word by word, predicts a pronunciation and sounds out the phonemes of the word. Based on this, the user provides a verdict with regard to the accuracy of the word-pronunciation pair:

- **Correct** – The word is a valid word in the language concerned and its pronunciation as displayed is correct.
- **Invalid** – The word is not a valid word (e.g. it is a URL), or it is spelled wrong, or it is only part of a word.
- **Uncertain** – The user is unable to decide whether the word and its pronunciation is valid.
- **Ambiguous** – There are multiple valid pronunciations, i.e. pronunciation variants.
- **Proper noun** – The word is a proper noun.
- **Foreign** – The word is a valid word from a foreign language, but not a word in the source language.

5.1 Providing a verdict

If no dictionary was imported, the project does not initially have a rule-set with which it can predict the current word, and the phoneme field is left blank. The user then has to provide the entire pronunciation.

DictionaryMaker will attempt to predict the phonemes of each new word, thereby enabling the user to correct the phoneme array (by adding, removing or changing certain phonemes), if necessary, before changing the status of the word.

Note that DictionaryMaker's accuracy will increase as the list of Correct words increases and more accurate rules are extracted by the algorithm.

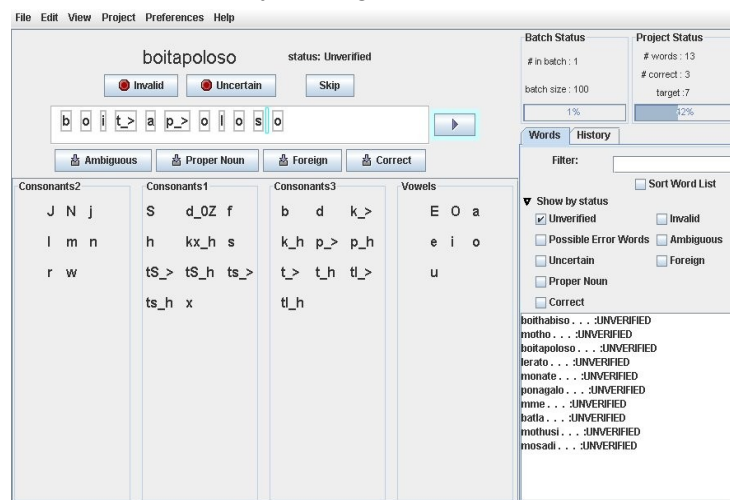


Figure 15: Main verification panel

The word currently being verified appears in the top left-hand corner.

The Current Word Panel displays the current word being verified, its current status and the list of phonemes used to pronounce the word. This panel also contains a number of buttons with specific functions.

- The Play button allows the user to listen to the phonemes in sequence.

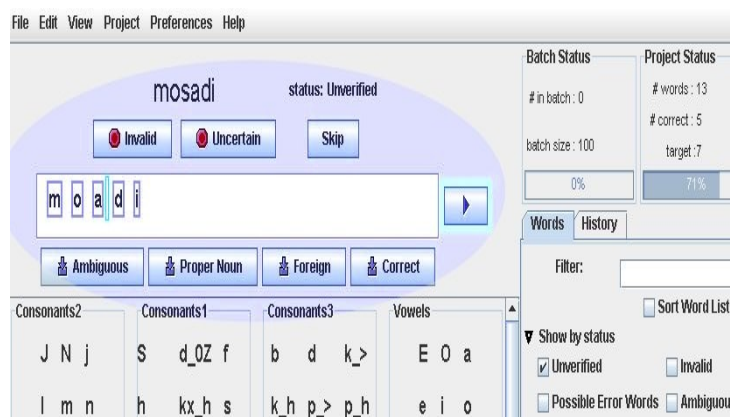


Figure 16: Current word panel

- The Skip button allows the user to move on to the next word to be verified without giving a verdict on the current word.
- The Correct button marks a word as correct.
- The Invalid button marks a word as invalid.
- The Uncertain button marks a word as uncertain.
- The Proper Noun button marks a word as a proper noun in the source language.
- The Foreign button marks a word as a foreign word.
- The Ambiguous button marks a word as having an ambiguous pronunciation, i.e. there is more than one correct pronunciation for the word.

The phonemes specified during the creation of the project can be seen, in their specified groups, in the lower left-hand corner.

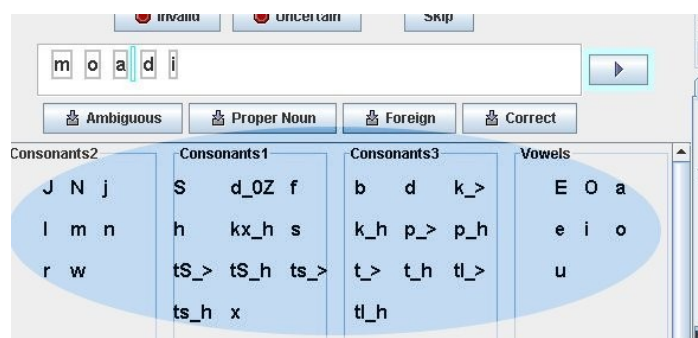


Figure 17: Phoneme panel

5.2 Updating the pronunciation

The phonemes are placed in the pronunciation list on the Current Word Panel in order for the user to correct the predicted pronunciation. The Phoneme panel is used in the following ways:

- Right-clicking one of the phoneme labels will play the appropriate sound file.
- Left-clicking a phoneme adds it to the pronunciation list at the current cursor location. The cursor can be moved with the right and left arrow keys on the keyboard, or by moving the mouse over the phoneme entry panel. The backspace key deletes the phoneme to the left of the cursor.
- Phonemes can be dragged to the pronunciation list and dropped in location, and can also be dragged from the list to remove them.
- Left-clicking on a phoneme in the pronunciation list pops up a context menu that lets you play, remove or change the phoneme.
- Right-clicking on a phoneme in the pronunciation list lets you quickly replace it and gives a drop-down list of phonemes with which it can be replaced.

5.3 Viewing the word list

The word list can be found to the right of the main window.

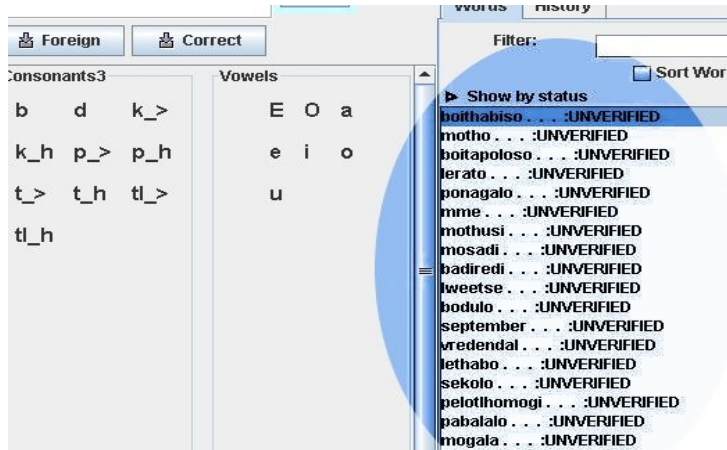


Figure 18: Word list panel

The Words tab displays the list of words which DictionaryMaker uses to create a dictionary. This list can display all words of a preferred status (Unverified, Correct, Uncertain, etc.), and can be filtered to display only words that start with specific letters.

It is possible to view previous words in the order in which they were verified, and to correct possible mistakes made during verification. This can be done by using the History tab.

5.4 Displaying the status

The current number of correct words can be viewed in the status panel, just above the word list. Here, the system shows you how many words of the supplied word list, as well as how many words of the current batch, have been completed.

6 Advanced use

6.1 Selecting different views of the word list

Words with a status other than Unverified, can be viewed and verified/corrected by selecting the appropriate option under Show by Status. You can also reduce the words in the list without having to remove words from the dictionary, by using the Filter field. In such a case, only the words that begin with the characters typed into the field, will be displayed.

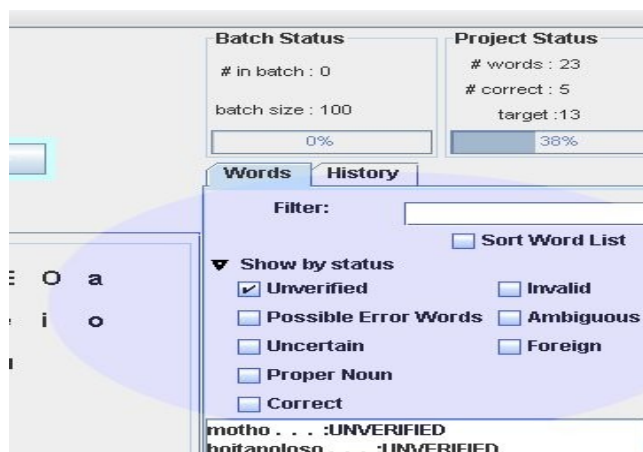


Figure 19: Selecting different views of the word list

6.2 Changing the system defaults

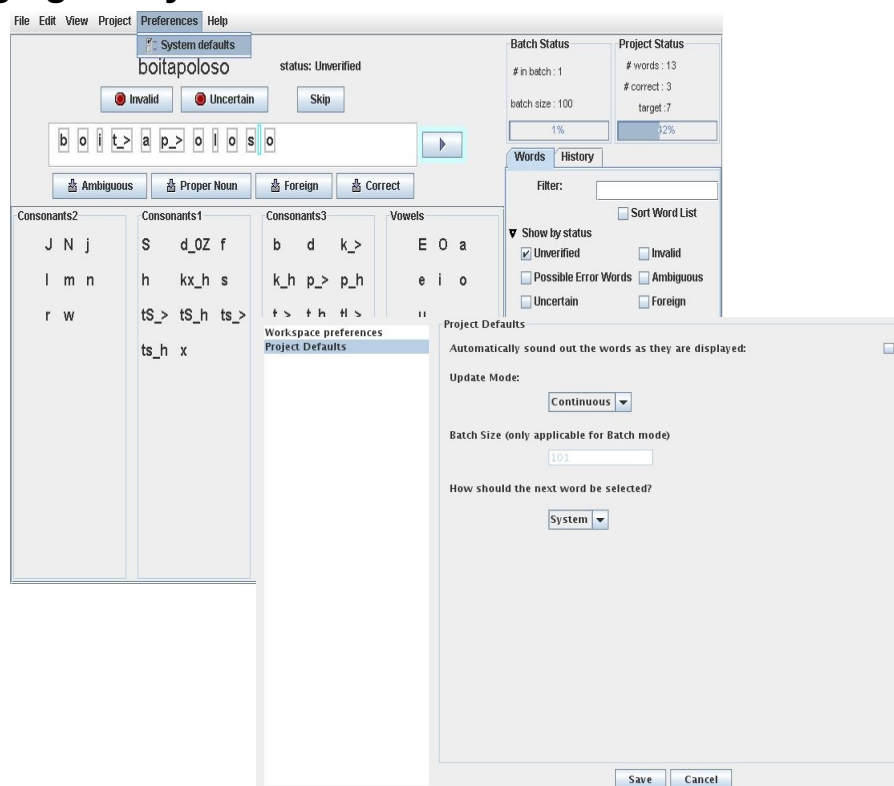


Figure 20: System defaults

System defaults are settings applied to any project opened by the current DictionaryMaker application.

Setting up the batch size can be done under the Preferences-> System Defaults menu item. The drop-down menu Update Mode contains the following options:

- Continuous: The system continues through the whole list of words. A new rule-set is extracted each time a word verified as Correct is added to the system.
- Batch: The system runs through a number of words, which you can specify in the lower field, without updating the rule-set. When a batch is completed, the rule-set is updated before the system continues with the next batch.

The next word selection defaults to System Select. It can be set through the drop-down box with the following options:

- In System Select mode, the next word to be verified is randomly selected from the list of unverified words.
- In User Select mode, the next word to be verified is the next unverified word, as displayed in the word list panel.

6.3 Changing the phoneme panel settings

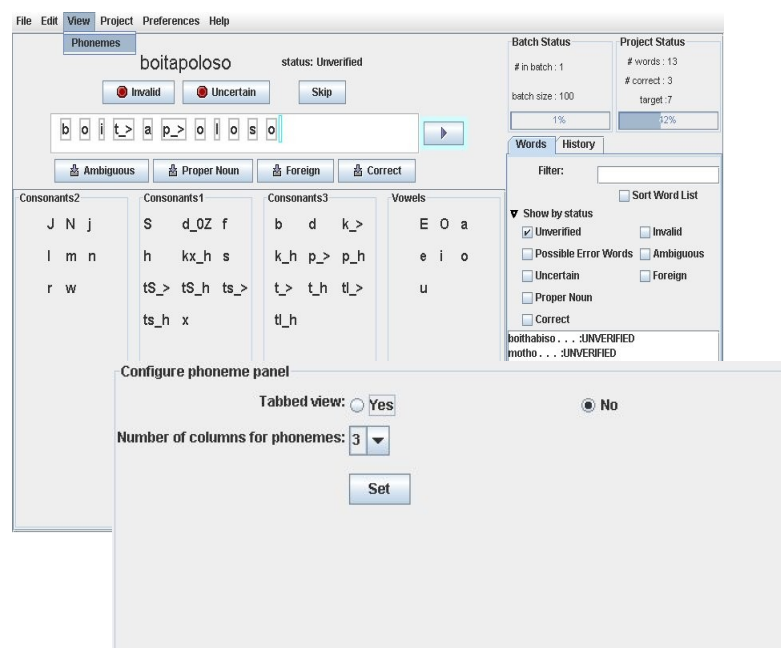


Figure 21: Phoneme view settings

The number of columns used to display the phonemes in each of the category panels of the Phoneme Panel can be changed through the menu item View-> Phonemes. It is also possible to change to a tabbed view for each category.

6.4 Changing the project defaults

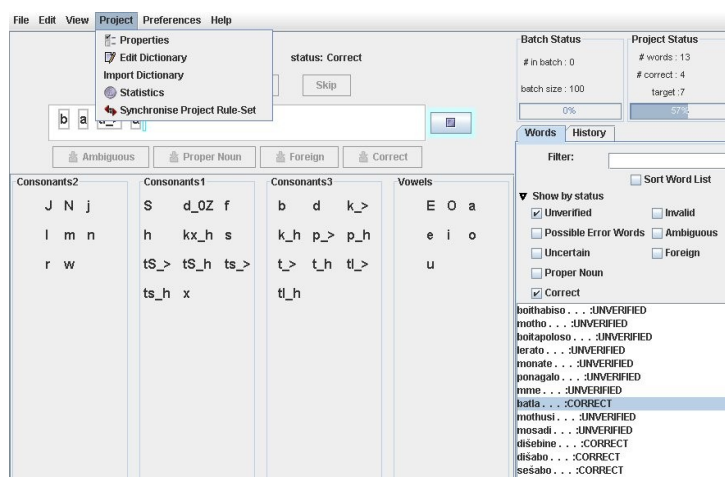


Figure 22: Project defaults

The Project Defaults are settings applied only to the current project. These settings can be enabled through the menu option Project-> Properties. If the Use Custom Settings check-box is checked, the custom settings override the system defaults.

6.5 Displaying project statistics

Number words in dictionary:

6055

Number of rules in dictionary:

1037

Number of words per status/verdict:

Unverified

0

Correct

5655 Proper Noun

Invalid

274 Foreign

Ambiguous

28

Uncertain

18

Project Rule Statistics

Display All Rules in Project

☒
☐
☐

Rules per Grapheme

☐
☐

Rules per Grapheme per context size

☐

Context	Predicted Phoneme
-a-	a
a-a-	0
ha-a-	a
ja-a-k	a
-b-	b
-d-	d
-e-	E
s-a-	e
l-a-	a

Close

Figure 23: Project statistics

The following Project Statistics are available under the Project-> Statistics menu item:

- Number of words in the dictionary
- Number of unverified words in the dictionary
- Number of correct words in the dictionary
- Number of words of the following verdicts: invalid, ambiguous, uncertain, proper noun, foreign
- Number of rules in the dictionary.

It is also possible to view the rules extracted from the dictionary in a number of ways:

- All rules in the project
- Rules per grapheme
- Rules per grapheme per context size.

6.6 Changing dictionary data

The data specified when the project was created is not final. You can add, remove or even import more words, graphemes or phonemes. You can even alter the words and phonemes currently in the project.

These options can be found under the Project-> Edit Dictionary menu item. The dialogues are similar to those used during project creation.

The Project-> Import Dictionary menu item, can be used to import an entire dictionary.

6.7 Exporting data files

You can also export the entire project to text-based files, using the File-> Export menu item. Exporting the graphemes, phonemes, and word list as a project will create files which may be used to initialise a new project which utilizes the same initial data.

The exported dictionary file will contain all the words specified in the word list, but only those specified as Correct will have accompanying phoneme definitions. The rule-set used by DictionaryMaker to bootstrap the pronunciation dictionary, can also be exported to a text-based file.