DICTIONARYMAKER USER MANUAL

VERSION 2.0(I)

M. Davel and M. Peche

14 September 2006

Table of Contents

1 Overview

The purpose of the *DictionaryMaker* system is to facilitate the creation of an electronic pronunciation dictionary in a target language, as originally described in [1]. Such a pronunciation dictionary consists of a list of words, each associated with one or more phonetic pronunciations. The developed pronunciation dictionary can be formatted for use by various speech processing applications, such as speech synthesis and speech recognition systems.

The system is designed to allow a speaker fluent in the target language to develop a pronunciation dictionary without requiring expert linguistic knowledge or programming expertise. Along with the pronunciation dictionary, a related set of grapheme-to-phoneme (g-to-p) rules is created automatically.

The system utilises a bootstrapping approach: improving models according to a controlled set of increments, at each increment utilising the previous model to generate the next. The system balances machine learning and human intervention with the aim to simplify and minimise the human intervention required during the bootstrapping process.

Only a word list, a grapheme set and phoneme set for the target language are required as inputs to the system, however a predefined dictionary can also be used to initiate a new project. Once initialised with these items, the system guides the target language speaker through the dictionary creation process.

2 Key concepts

2.1 Main functionality

The system's main mode of operation is an interactive one that requires user input during the execution of each task. The main interactive functions supported by the tool are:

- Creating a new electronic pronunciation dictionary through bootstrapping.
- Manipulating and updating an existing pronunciation dictionary.

In addition, some supporting (non-interactive) functionality is provided:

- Analysing a project, dictionary or rule set status.
- Verifying data during importing is in the correct formats.
- Creating a grapheme or phoneme set from an existing dictionary.
- Creating a g-to-p rule set from an existing dictionary.
- Identifying possible errors in an existing dictionary.

2.2 Main data components

The main (abstract) data components are:

1. A grapheme set *G*: a list of allowed graphemes (letters)
2. A phoneme set *P*: a list of allowed phonemes (sounds)
3. A rule set, describing how a grapheme in a specific graphemic context maps to a specific phoneme. The rule set format: $g_0..g_{i-1} – g_i – g_{i+1} ..g_n$ → p, g_j E *G*, p_i E *P.* The rule specifies that grapheme g_i, when found in the left context $g_0..g_{i-1}$ and right context g_{i+1} $..g_n$ is realised as phoneme p_i.
4. A pronunciation dictionary, consisting of word/pronunciation pairs currently being manipulated. For each pair a status indicator specifies whether a word has been verified, and if so, what the verdict is. (See for more info).
5. Activity log: a log of all significant activity associated with a specific dictionary, including timing information.

2.3 Dictionary project

The data components and history (log) associated with a specific dictionary creation project. A single dictionary is often created through a sequence of projects, each projects improving or manipulating the dictionary in a specific way.

2.4 Bootstrapping process

User Perspective

Before an experiment can be run, a list of target words, a grapheme set, a phoneme set and an associated set of sound samples are defined by the user.

The system runs through the word list one by one, predicts a pronunciation and presents the human with an audio version of the word: the human acts as a `verifier' and provides a verdict with regard to the accuracy of the word-pronunciation pair: invalid, correct, uncertain or ambiguous. or invalid. (An “invalid” word is not an actual word in the target language, e.g. a URL, half a word, or a word from a different language. The pronunciations associated with valid words are evaluated further: a “correct” pronunciation is 100% correct, an “ambiguous” pronunciation can have more than one pronunciation based on context, and an “uncertain” pronunciation indicates that the user itself is not sure how to pronounce the associated word.)

If the word is wrong, the verifier can correct the word on a phone-by-phone basis. This process is repeated (with increasingly accurate predictions) until a pronunciation dictionary of sufficient size is obtained.

System perspective

During each bootstrapping cycle, grapheme-to-phoneme rules are automatically extracted from the current version of the pronunciation dictionary, and used to generate additional word/pronunciation pairs based on the word list provided. The new word pairs are presented to the user for correction, and from the corrected words better rules are extracted.

The system initially predicts empty pronunciations, which, when corrected, form the basis for further bootstrapping.

The overall process consists of the following steps:

1. The system analyses its current understanding of the task (based on the word list characteristics and the type and status of pronunciations from the overall word list) and chooses the next word or set of words to be considered.

2. For each of the words on the above list, the system generates a new pronunciation using its current grapheme-to-phoneme rule set.

3. The system creates a ‘sounded' version of each word using the predicted pronunciation and the user-specified sound samples, and records the verifier's response.

4. Based on the status of each of the words in the newly verified word/pronunciation list, the system extracts a new grapheme-to-phoneme rule set. Only word-pronunciation pairs marked as “correct” are used when extracting a new rule set.

5. This process is repeated until a sufficient number of correct words are obtained.

Per experiment, the system logs the history of all activities and archives the intermediary data resources for further analysis.

2.5 Next word selection

In ‘system select’ mode, the system randomly selects the next word to be added to the dictionary. In ‘user select’ mode, the user selects the next word from a list offered by the system. (See for more info on system select mode).

2.6 Audio support

A user can listen to the phoneme set at any time. When a new word is predicted, a ‘sounded’ version of the word is generated. This sounded version consists of a concatenation of the phoneme samples. In future versions, TTS generation of the sounded version may be incorporated in the tool. Currently, sound samples must be added by the user. Sound files need to be recorded in a different application and saved separately. (We recommend *Praat* available at http://www.fon.hum.uva.nl/praat/ for this purpose.)

The user can choose to play the sounded version as many times as required. In auto-play mode, the word is played once when predicted, without requiring user interaction.

It is strongly recommended that the audio support is utilised when building a pronunciation dictionary, especially if the user building the dictionary has limited linguistic experience or is unfamiliar with the phone set used [2].

2.7 Updating modes

Rules are updated during the bootstrapping process according to the update mode chosen: ‘continuous update’ or ‘batch update’ mode:

- In continuous update mode, a new rule set is extracted every time a word verified as ‘correct’ is added to the system. (The updated rules are then used to predict the subsequent word.)

- In batch mode, the rule set is only updated after a user-specified number of words (batch size) have been verified as correct.

- In incremental batch mode, a modified version of rule updating is used after every word (incremental rule update – a faster, less accurate version of rule extraction). A full rule set update only occurs after a user-specified number of words (batch size) have been verified as correct.

An experiment is automatically set up such that a user starts in continuous mode, and automatically switches to incremental batch mode after the dictionary reaches a pre-specified size. However, the user can choose to switch modes at any time.

2.8 Error detection

The system has the ability to identify possible errors in a pronunciation dictionary by flagging words that create a large number of exceptional rules. However, the functionality is not yet available for the user as the possible error words are only printed out to a log file, which can be found in the specified project folder.

The number of errors identified is influenced by an error threshold (the lower the threshold, the more potential errors are identified). See [3] for more info.

2.9 Verification status

For each word/pronunciation pair in a pronunciation dictionary, a status indicator specifies whether a word has been verified, and if so, what the verdict was. An "invalid" word is not an actual word in the target language, e.g. a URL, half a word, or a word from a different language. The pronunciations associated with valid words are evaluated further: a "correct" pronunciation is 100% correct, an "ambiguous" pronunciation can have more than one pronunciation based on context, and an "uncertain" pronunciation indicates that the user itself is not sure how to pronounce the associated word. Only "correct" words are used during grapheme-to-phoneme rule extraction.

2.10 Grapheme-to-phoneme rule extraction

The Default&Refine algorithm is used for rule extraction and pronunciation prediction. Default&Refine [4] is a greedy search algorithm that extracts a list of increasingly specialised rules from a training dictionary. Prior to rule extraction, word-pronunciation pairs are aligned using an optimised version of Viterbi alignment [5].

The rule set format is similar to most rewrite rule schemes:

$g_0..g_{i-1} - g_i - g_{i+1} ..g_n \rightarrow p$, g_j E Graphemes, p_i E Phonemes

where this rule specifies that grapheme g_i, when found in the left context $g_0..g_{i-1}$ and right context $g_{i+1} ..g_n$ should be realised as phoneme p_i. Rules are ordered explicitly. When a new word is being predicted, the graphemes are processed one at a time. Each grapheme and its left and right context is compared to the rules in the rule set, and the first matching rule is applied.

3 Getting started: installation

3.1 Linux

The program had been tested on Fedora Core 4, Debian, and Mandria systems. *DictonaryMaker* will need Java-1.5 to compile the source code, and the build script uses the following:

- Apache Ant v1.6.2
- FindBugs 0.9.1
- Checkstyle v3.5-1
- Quilt v0.45
- java v1.5.0_06
- Junit v3.8.1-4
- JFCUnit

To build the "*.jar"-file:

- Copy "build.properties.example" to "build.properties", and edit it so that the necessary paths are valid for the local computer DictionaryMaker is to be run on.
- From the command-line, change directory to the DictionaryMaker folder.
- Run: `ant jar`

To run *DictionaryMaker*:

- Run: `java -jar dist/dictionarymaker.jar`
- For better performance, you can increase the memory used by the java VM by running:

```
          java -Xms128m -Xmx356m -jar dictionarymaker.jar
```

3.2 Windows

DictionaryMaker has also been tested on Windows XP, and have run smoothly as well. Execute the created *.jar file like any other *.exe application. Please make sure that under the *.jar file's 'properties' that it 'opens with...' the correct java platform binary.

A batch-file 'rundict.bat' is provided for better performance.

4 Getting started: using *DictionaryMaker*

This section provides a quick-start guide to using *DictionaryMaker*. For more detail, see the case study in section 7.

4.1 Starting a new dictionary

Starting a new dictionary is as easy as starting a new project; which can be found under the 'File' menu. You will need to provide the following to create a new project:

- *Name and Location*: Name of the new project and where it will be saved.
- *Graphemes*: The list of grapheme (alphabetical symbols) the target language uses. These can be added manually, or imported from a text-based "*.gra"-file.
- *Phonemes*: The list of phoneme (vocal sounds) used by the target language. Again, these can be added manually, or imported from a text-based "*.pho"-file. When adding sounds however, a phoneme name, category, and the location of the appropriate sound file must be provided:
    - *Categories* are used in DictionaryMaker to sort the available phonemes into groups for ease of use. All phonemes allocated to the same group are displayed together in one panel.
    - The *sound files* enable DictionaryMaker to play back a sound or pronunciation to the user, so that the word can be verified. Sound files need to be recorded in a different application and saved separately. (We recommend *Praat* available at http://www.fon.hum.uva.nl/praat/ for this purpose.)
- *Words list/Dictionary: Now, the words in the project is defined, using either a raw word list, or a starting dictionary.*
    - *Word list*: A list of words the DictionaryMaker will use to bootstrap the new dictionary. These can also be added manually, or imported from a text-based "*.wdl"- or "*.txt"-file .
    - Dictionary: The bootstrap process can be accelerated by beginning with another, smaller dictionary.

# **Note** that each grapheme, or word in the above text-based files should be on a line of its own! The phoneme name, file, and category (in that order) must be listed on the same line, but again each phoneme should be on a new line.

4.2 Building on an existing project

Find an existing dictionary under the 'File -> Open Recent' menu to continue building, or open the appropriate “*.proj”-file.

4.3 Using DictionaryMaker

The layout of DictionaryMaker is quite easy to understand:

- *Current Word Panel:* Displays the current word being added to the dictionary, and the array of phonemes used to pronounce the word. Buttons allows the user to listen to the phonemes in sequence, and then to either accept or reject the word. Pronunciations may also be marked 'Uncertain' (if the user is unsure of the pronunciation) or 'Ambiguous' (if one word can be realized as to two different pronunciations e.g. the present and past tense of the word read, pronounced as */ r iy d /* and */ r eh d/* respectively.
- *Phoneme Panel:* Displays the set of phonemes the *DictionaryMaker* associates with the target language. These phonemes are divided according to the categories provided when the phoneme-list was added, and double-clicking one of the phoneme labels will play the appropriate sound-file.
- *Word list Panel*: 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 a specific way. The option to view previous words in the order they were verified and correct any possible mistakes made during verification is also available under the 'History' tab.
- *Status Panel:* The status of the current project and the current batch (if *DictionaryMaker* is using “Batch” update-mode) are displayed. It shows the size, target-size, and percentage completed in each case.

After having created a new project, the *DictionaryMaker* system runs through all the unverified words (if a batch-size is specified then it will run through the first batch of words) one at a time. The user then has to update the pronunciation by dragging the sounds as they are spoken from the Phoneme Panel into the text-box on the Current-Word Panel, or correcting the given pronunciation. Selecting the appropriate verdict button will process the current word and *DictionaryMaker* will continue with the next word.

Right clicking on a phoneme both in the text-box or the phoneme panel will open a menu which gives the user the option of removing, editing or playing the current phonemes. Right clicking on the text-box itself (if no phonemes are added yet, or in between present phonemes) it goes directly to the add-phoneme menu.

Batch update mode and batch sizes can be specified under the “Preferences -> System Defaults” menu item.

5 Advanced functionality

5.1 Changing grapheme set / phoneme set / word list

The grapheme sets, phoneme sets and the word list are defined during the creation of the new project. However, graphemes, phonemes and words may be added, edited or removed later during the development of the dictionary and from an existing project.

This is simply done by choosing the 'Project -> Edit Dictionary' menu-item and defining the new grapheme and phoneme sets and word list. Not all changes to the dictionary are allowed (for example a grapheme that is in use may not be removed). The system provides appropriate warnings if illegal changes are attempted.

5.2 Changing the update mode in which DictionaryMaker runs

Under the “Project -> Properties” menu item. By checking the “Use Custom Settings”, the user can set the update mode used to extract the rules from newly verified words:

- *Auto*: The system runs “Continuous” until a specified amount of words has been processed, then switches over to “Incremental”

- *Continuous*: The system continues through the whole list of words. A new rule set is extracted every time a word verified as ‘correct’ is added to the system.

- *Batch*: The system runs through a user specified amount of words, updates the rule-set before continuing with the next batch.

- *Incremental*: In incremental batch mode, a modified version of rule updating is used after every word: incremental rule update is a faster, less accurate version of rule extraction that is locally optimised. A globally optimal rule set update only occurs after a user-specified number of words (the batch size) have been verified as correct.

DictionaryMaker also gives the user the opportunity to choose if the system or the user should specify the next word to be verified.

These options is also available under the “Preferences -> System Defaults” menu item. However, it is important to stress that this second option will set the default options for all future projects, whereas the first will only apply to the currently loaded project.

5.3 Exporting data

Various types of data may be exported from DictionaryMaker to text-based files, using the 'File -> Export' Menu. Exporting the graphemes, phonemes, and word list will create files which may be used in creating another project which may utilise the same data.

5.4 Analysing a project, dictionary or rule set status

The current project statistics can be viewed under 'Project -> Statistics'. Here the amount of words, both according to their given verdict or not can be viewed. The options of viewing the rules, and amount of rules per grapheme (and per grapheme per context size) is able available here.

5.5 Synchronising the rule set.

The rule set of current project can be updated (by running a full rule extraction) without waiting for the specified batch to fill up. This event can be found under the 'Project -> Synchronise' menu.

6 Future functionality

Functionality we would like to add in the near future:

- Creating a dictionary from an existing rule set and a given word list.
- Error detection made available on requested for users, as well as setting the error threshold; The user can then correct any of the possible errors (in user select mode).
- Exporting the dictionary and/or rule set to pre-defined formats, including HTK dictionaries and Festival dictionaries and letter-to-sound rules.

- Recording the sound clips from within the *DictionaryMaker* application.

- Additional tools for manipulating grapheme and phoneme sets, e.g. splitting and merging of phonemes (with subsequent re-bootstrapping of relevant words).

- Determining whether a formant synthesizer works better than our simple concatenative phoneme synthesiser.

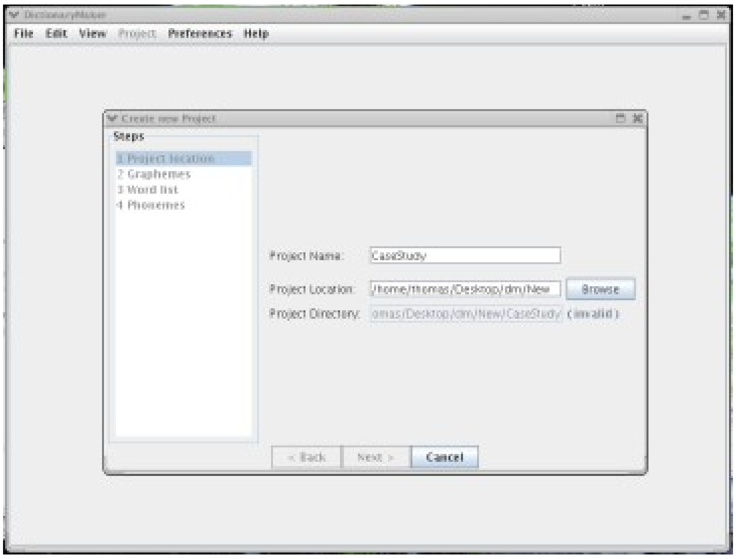
Or, if you would like to contribute to this project – please help us by adding some of this functionality!

7 Case study

In this section we build an Afrikaans pronunciation dictionary. The incomplete project is also available in the './data' folder, project name 'CaseStudy.proj'

Step 1: Creating a new project

In order to bootstrap a new dictionary, we need to create a new project that will contain the raw words, graphemes and phonemes used in our new dictionary, as well as the half finished dictionary itself. To achieve this is easy: just select the 'New Project' option found under the 'File' Menu. This will open the following pane:



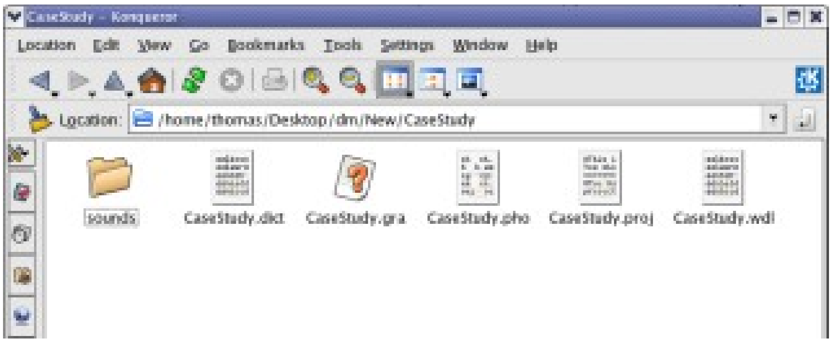
Project Name: CaseStudy

Project Location: /home/thomas/Desktop/dm/New

Project Directory: omas/Desktop/dm/New/CaseStudy (invalid)

Figure 1: Creating a new project

Here you can fill in the project name, and select the desired location for the project to be stored at. I have set the project name to 'CaseStudy'. Please note that the project is already encapsulated inside a created folder, therefore (referring to this example) I don't need to create a folder 'CaseStudy' to hold all the project's files – the folder is created automatically:

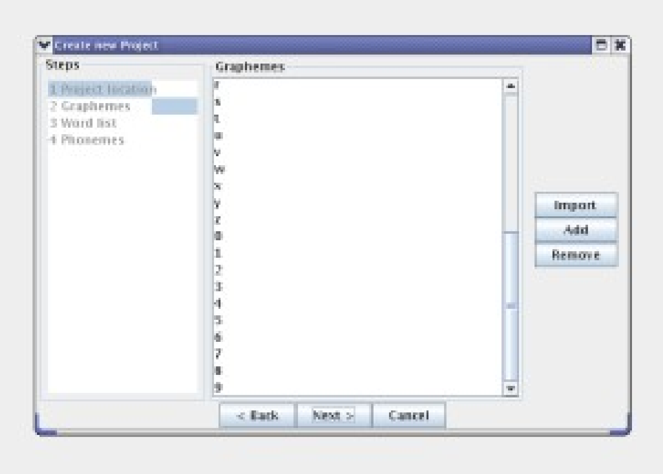


/home/thomas/Desktop/dm/New/CaseStudy

sounds
CaseStudy.dict
CaseStudy.gra
CaseStudy.pho
CaseStudy.proj
CaseStudy.wdl

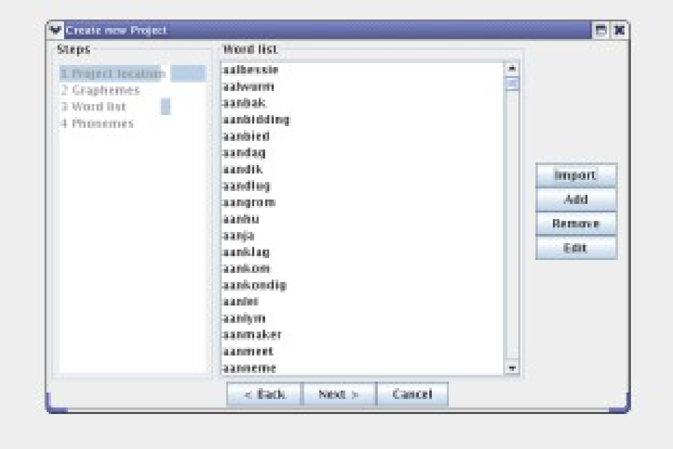
Figure 2: Files created during project initialisation

Next, we define the graphemes, word list and phonemes that the new dictionary will require. You can add, remove or import from a predefined text-file: *.gra for graphimes; *.wdl for the list of words; and *.pho for the phonemes. Note that such a text-file must contain only one entry on each line (or one phoneme per line in the case of the *.pho file, which must contain the phoneme name, sound-file and class on the same line!)



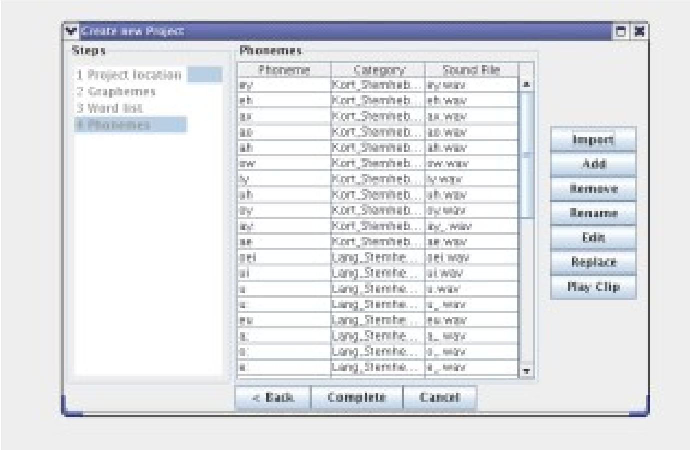
r
s
t
u
v
w
x
y
z
0
1
2
3
4
5
6
7
8
9

Figure 3: Editing the graphemes



aalbessie
aalwurm
aanbak
aanbidding
aanbied
aandag
aandik
aandlug
aangrom
aanhu
aanja
aanklag
aankom
aankondig
aanlei
aanlym
aanmaker
aanmeet
aanmeme

Figure 4: Editing the word list



Phoneme	Category	Sound File
ay	Kort_Stemheb...	ay.wav
eh	Kort_Stemheb...	eh.wav
ax	Kort_Stemheb...	ax.wav
ao	Kort_Stemheb...	ao.wav
ah	Kort_Stemheb...	ah.wav
aw	Kort_Stemheb...	aw.wav
ly	Kort_Stemheb...	ly.wav
uh	Kort_Stemheb...	uh.wav
oy	Kort_Stemheb...	oy.wav
ay:	Kort_Stemheb...	ay_.wav
ae	Kort_Stemheb...	ae.wav
aei	Lang_Stemhe...	aei.wav
ai	Lang_Stemhe...	ai.wav
u	Lang_Stemhe...	u.wav
u:	Lang_Stemhe...	u_.wav
eu	Lang_Stemhe...	eu.wav
a:	Lang_Stemhe...	a_.wav
o:	Lang_Stemhe...	o_.wav
e:	Lang_Stemhe...	e_.wav

Figure 5: Editing the phonemes

And there you have a new project. Like normal projects, you can then save your work before closing it, or open an already existing project to continue with it.

Step 2: Using *DictionaryMaker*

After having created a new project, the *DictionaryMaker* system runs through all the unverified words one at a time.

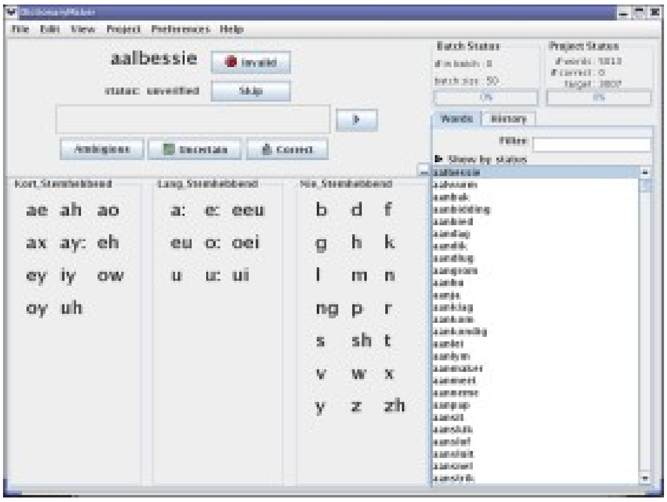


Figure 6: Main bootstrapping panel

The word currently being verified is at the top left corner:

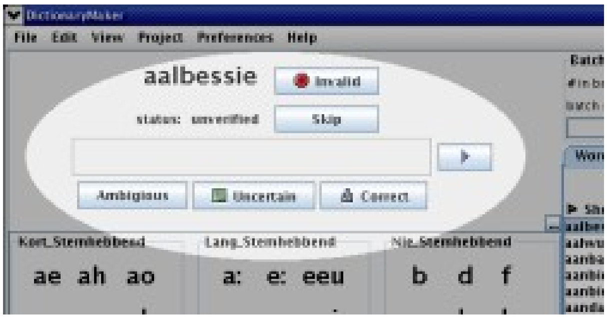


Figure 7: Current word being verified

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

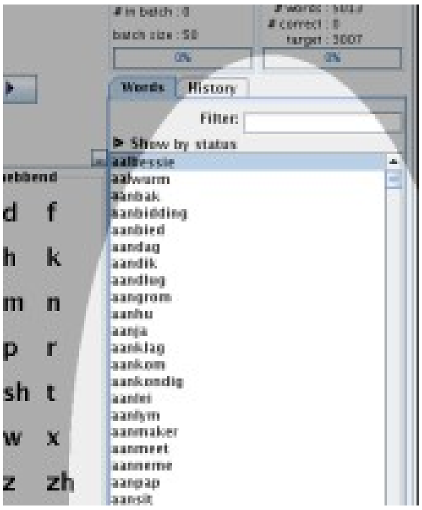


Figure 8: Word list panel

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

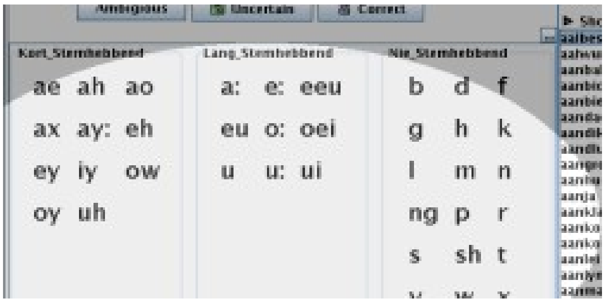


Figure 9: Phoneme panel

At first, the project doesn't have a rule set with which it can predict the current word, and the phoneme field is left blank. To start building your words, drag the correct phonemes into this field from below. A pop-up menu enables you to remove or change a phoneme, when an inserted phoneme is selected, otherwise it allows to insert a phoneme by listing all the phonemes used by the project.

Before evaluating the pronunciation, first decide whether this actually is a valid word for the target language. If not, press “Invalid” and the next word will be displayed.

The menu gives the option to play a single phoneme, which is also achievable by double-clicking the required phoneme on the lower panel. The entire array of selected phonemes can also be played by clicking the “Play” button to the right of the field.

When you are satisfied that the array of phonemes in the field accurately pronounce the given word, press the “Correct” button just below the field. If you are unsure of the pronunciation, there exists both an “Uncertain” button (if you really don't know what the word sounds like) and an “Ambiguous” button (if the specific spelling happens to have several pronunciations). There also exists to the top of the field, a “skip” button if you'd like to return to the word later. Note that when you press “Skip”, the status of the word is not altered.

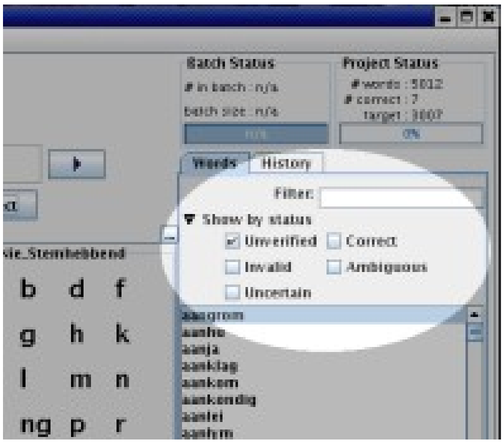
When one of the buttons (other than “play”) is pressed, the current word is moved to the specified status, and the next unverified word is loaded and the process is repeated. In the background, an algorithm extracts rules specifying how each grapheme (in a specific context) maps to a specific phoneme. This is done after each word, but *DictionaryMaker* can also be set to do it after a certain amount of valid words have been completed. This setting is determined by the user.

DictionaryMaker will attempt to predict the phonemes of each new word, therefore enabling you to just 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 increase and more accurate rules are extracted by the algorithm.

Step 3: Advanced usage

Selecting different views of the word list

Words with a status other than 'Unverified', can be viewed and verified/corrected by selecting an 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 starts with the characters typed into the field will be displayed. This is useful when the user can define the next word to be verified:



Batch Status
# in batch : n/a
batch size : n/a

Project Status
# words : 5012
# correct : 7
target : 3007

Filter:
▼ Show by status
☑ Unverified ☐ Correct
☐ Invalid ☐ Ambiguous
☐ Uncertain

aaagrom
aanhe
aanja
aanklag
aankom
aankondig
aanlei
aanhym

Figure 10: Selecting different views of the word list

Changing the update mode

Setting up the batch size can be done under the “Preferences” menu. The drop down menu “Update Mode:” contains the following options:

- Auto: The system runs in “Continuous” mode until a specified amount of words have been processed. This amount of words you can specify in the upper field. It then switches over to “Batch” mode.

- Continuous: The system continues through the whole list of words. A new rule set is extracted every time a word verified as ‘correct’ is added to the system.

- Batch: The system runs through a specified amount 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.

- Incremental: In incremental batch mode, a modified version of rule updating is used after every word. (Incremental rule update is a faster, but less accurate version of rule extraction.) A full rule set update only occurs after a specified number (batch size) of words have been verified as correct.

Displaying the status

The current amount of correct words can be view 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 are completed.

Step 4: Changing dictionary data.

The data specified when the project was created, isn't 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” menu, at the “Alter Dictionary” item. The dialogs resemble those first seen during project creation.

Step 5: Exporting Data files.

To export the dictionary you have just created, you can use the “Export” menu, which you can find under the “File” Menu. The exported 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.

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

References

1 M. Davel and E. Barnard, "*Bootstrapping for language resource generation*", in Proceedings of the Symposium of the Pattern Recognition Association of South Africa, South Africa, 2003, pp. 97–100.

2 M. Davel and E. Barnard, "*The efficient creation of pronuniciation dictionaries: human factors in bootstrapping*," in Proceedings of Interspeech, Jeju, Korea, October 2004, pp. 2797–2800.

3 M. Davel and E. Barnard, "*Bootstrapping pronunciation dictionaries: practical issues*," in Proceedings of Interspeech, Lisbon, Portugal, September 2005.

4 M. Davel and E.Barnard, "*A default-and-refinement approach to pronunciation prediction*," in Proceedings of the Symposium of the Pattern Recognition Association of South Africa, South Africa, November 2004, pp. 119–123.

5 M. Davel and E. Barnard, "*The efficient creation of pronunciation dictionaries: machine learning factors in bootstrapping*", in Proceedings of Interspeech, Jeju, Korea, October 2004, pp. 2781–2784.