

## ONZEminer workshop

Jen Hay: [jen.hay@canterbury.ac.nz](mailto:jen.hay@canterbury.ac.nz)  
Robert Fromont: [robert@fromont.net.nz](mailto:robert@fromont.net.nz)  
University of Canterbury

This document walks you through the use of ONZEminer.

ONZEminer is database for transcribed sound files. Transcripts are first created using **Transcriber**, and then uploaded into the ONZEminer database.

ONZEminer is installed on your machines, with a small demo database of 26 speakers. Feel free to skip through this handout to whichever part most interests you, or – alternatively – not to follow the handout at all, and just follow your nose!

Much of the text here is also available from the help (?) screens on ONZEminer.

### **Demo database: The ONZE Corpus**

The ONZE Project has a large (more than 1000 hours) archive of recordings of people born in New Zealand from the 1850s to the 1980s, organised into three corpora.

- The oldest collection, the ‘Mobile Unit Corpus’, contains recordings of speakers born between 1850 and the early 1900s.
- The second collection, the ‘Intermediate Archive’, features speakers who were born between 1890 and 1930 (including some descendants of the ‘Mobile Unit’ speakers).
- The third collection, the ‘Canterbury Corpus’, contains recordings made by students of linguistics, and includes speakers born between 1935 and 1985. The Canterbury Corpus contains both interviews and wordlists.

Taken together, these three collections of recordings span the history of New Zealand English.

The demo version of ONZEminer installed on your computers includes just 16 speakers from the Canterbury Corpus (CC), and 10 speakers from the Intermediate Archive (IA).

The URL for ONZEminer on your machines is <http://localhost:8080/miner>. With luck, the browser should already be pointing to this site when you sit down at your machines.

## The filter page

The usual first port of call when you are doing an analysis is the 'filter' page. You can select 'filter' from the navigation bar, or follow the 'filter speakers' link from the home page.

The 'filter speakers' page is the page from which you start a search of the transcripts in the ONZEminer database. On this page, you specify which speakers to search – and then their utterances will be scanned when searching for the pattern you enter on the next page.

family	Selecting a transcript family from the dropdown list causes only speakers who appear in transcripts that belong to that family to display in the list. In the demo database, each interview is chunked into approximately 5 minute sound files, and all of the sound files from a particular interview constitute a family.
corpus	Selecting a corpus from the dropdown list displays only those speakers who belong to the selected corpus. This database has two corpora – the IA and the CC.
gender	The list can be filtered by gender.
job	A year-range can be entered here to filter by year-of-birth.
class	
region	
mother	
father	Each of these speaker attributes can be used to further restrict the list. The content of each of these dropdown boxes depends on what values are actually stored against speakers. If the dropdown list is empty, this means that no speaker has a value set for that attribute.

The speaker list can be filtered by various speaker attributes:

Note that the filters are cumulative, so if you select corpus=CC and gender=male, then the speakers who are both in the 'CC' corpus and male will appear in the list, and no others.

Changing any filter results in the page being reloaded. For the job date range, the list will refresh when the cursor leaves the text box.

Once you have selected your criteria, you can select the specific target speakers by clicking the tick box to the left of the speaker. To select or unselect all speakers, use the checkbox at the top of the list, in the same row as the filter conditions.

Clicking on the speaker's name in the list takes you to the speaker details page.

→*Play with the filter page to get a feel for how the filters work. When you are comfortable with this, create a subset of speakers that you are interested in working with, and hit 'search'.*

Make sure you have ticked at least one speaker before clicking the search button.

The search button allows you to do an orthographic search of the transcript files. The layered search button allows you to do a 'layered' search of information derived from the CELEX database, if this has been imported, or other tailor-made layers you may have created. We will come to this in due course.

## **The Search Page**

The list of speakers you selected appears at the top of the search page. The list can be hidden or shown by clicking the *speakers* heading.

For each speaker you can click the [notes] link to view (or hide) that speaker's notes, and you can click on the speaker's name in order to edit their details.

The **Text/Regular Expression** box is where you enter the regular expression you wish to search for. This regular expression is matched against complete lines in the orthographic transcript so you can enter patterns that would match across word-boundaries. For example you could enter 'not\seven' to find all instances where the word *not* is followed by the word *even*. (\s stands for any kind of white space). Regular Expressions allow you to search for text by pattern as well as finding text that matches exactly the text that you enter.

For example, a search for 'our' will return all instances of text matching that precise sequence of characters - i.e. all instances of the word 'our' will be returned, as well as words like *your*, *hour*, *courage*, etc. By using the regular expression '[yh]our' you can restrict the result to only those instances of *our* that are preceded by the letter *y* or the letter *h* - the square-brackets mean 'any one of these characters', so will match both *y* and *h*. Help on regular expressions is available by clicking on 'Regular Expression' on the search page.

## Some Regular expression basics:

[abc]	a or b or c
^a	begins with a
a\$	ends in a
.	any character
a?	"a" zero or one times
a+	"a" one or more times
a*	"a" zero or more times
\s	any white space character.

So in a regular (non layered) search, the following searches translate to:

\sn	words beginning with n
^n	lines beginning with n
n\$	lines ending in n
[snmk]t	t, preceded by one of s, n, m or k.
the\s[aeiou]	the, followed by a word beginning with a, e, i, o or u.
st?r	s, followed by an optional t, followed by an r
\sah+	ah, or ahh, or ahhh etc.

Each transcript can optionally have a single speaker marked as the main speaker for that transcript - e.g. if the recoding is of an interview, normally the interviewee would be marked as the main speaker, and the interviewer and any other extraneous speakers would not. The **‘only search transcripts for which these are the main speakers’** checkbox restricts the search to only those transcripts in which the main speaker is one of the speakers you've selected.

The **‘export results to excel’** checkbox allows you to save the results of the search in a spreadsheet. You can specify which columns to include in the spreadsheet by clicking the **[options]** link. The format of the results is "CSV" - comma-separated values - which is a plain-text format recognized by Microsoft Excel and many other spreadsheet and database systems. If Excel will not open the file directly, save it somewhere locally first, then double-click the saved file. This feature is very useful for conducting an analysis – the spreadsheet contains hyperlinks to the matching sound excerpts. The analyst can add columns which record the results of the analysis.



Transcripts are assigned a 'type' when they are uploaded. If you click the **‘transcript types’** link you can restrict the transcripts searched by type, simply by ticking or unticking the checkboxes revealed. In this data set, all of the IA speakers are of the type ‘interview’, but the CC speakers have both ‘wordlist’ and ‘interview’ transcripts which can be searched.


→*Enter a search. This could be something nice and simple like a particular word (e.g. “house”). Or it could be something slightly more complicated (e.g. “[td]\$” will return all lines ending in a ‘t’ or ‘d’).*

## **The results page**

The results are ordered by speaker and then by transcript, and are sequentially numbered.

Each result line has:

- a checkbox, which allows the utterance to be selected for audio-export (see below)
- an html - no sound transcript icon  which, if clicked, takes you to the utterance in the HTML transcript (i.e. browser-friendly, but with no sound)
- an html - with sound transcript icon  which, if clicked, takes you to the utterance in the fully-interactive transcript (with sound)
- the text of the complete utterance which matched your regular expression, which, if clicked, takes you to the utterance in the fully-interactive transcript.

Below the results is an extract audio icon  which if clicked, allows you to extract the audio of the selected matching utterances - i.e. the results which you've ticked. You will be prompted to save a zip file, which contains extracted portions of the original recordings.

(Note: The extract audio process only works for utterances whose sound files are locatable by the ONZEminer server - i.e. where the wav files are stored on the server with the transcript files (or, more exactly, in a wav folder next to the trs folder in which the transcript files are stored by ONZEminer)).

→*Click on an utterance to go to the interactive transcript page. This will take you to the matching transcript, with the matching line aligned to the top of your screen. You can scroll up and down to study the full context of the utterance. If you click on the relevant line, you can hear it.*

→*Using the ‘back’ button on your browser, you can now return to your search results, and click on a different matching utterance in order to go through to that transcript, and listen to the utterance.*

→*Export the search results to excel, and use the spreadsheet hyperlinks to navigate to a couple of the matching utterances.*


Once you have experimented with navigating between search results, spend some time exploring the interactive transcript....

## The interactive transcript


This page displays the text of the transcript. If this transcript is part of a 'family' that were all uploaded at the same time, then the <<previous and next>> links (in the top and bottom corners of the page) lead to the previous and the next transcripts in the family.

If the transcript has a 'main speaker' set, then that speaker's details appear in the speaker list, and all of that speaker's utterances are in bold.

If you have selected an 'interactive' format of the transcript, you can click on any line to hear the audio. In addition, a floating toolbar will be displayed in the top-right corner of the page, and various icons are available:

 start playback

Clicking this starts the sound playing, either from the beginning, or from the selected utterance.

 stop playback


This icon is displayed on the floating toolbar. Clicking it stops the sound playing.

 repeat

This icon is displayed on the floating toolbar. Clicking it causes the current playback to rewind approximately one second and play again.

 open praat

This icon is displayed both on the floating toolbar and on each utterance in the transcript. Clicking it on the toolbar opens Praat and imports into it the current sound file. If the icon is clicked on the utterance, then the utterance itself is extracted and displayed as a spectrogram in Praat.


 extract audio fragment

This icon is displayed on each utterance if the ONZEminer server can access the sound file (i.e. the server computer, not necessarily the computer on which the browser is running). Clicking it extracts just the selected utterance from the sound file, and allows you to save the extract to disk.


The transcript is available in various formats - links for which are at the top of the transcript page:

 plain text

Only the text of the transcript. This is suitable for use in a plain-text editor, a word-processor document, etc. There is no highlighting or indenting of text.

 html with no sound

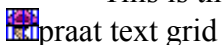
This is a 'browser friendly' representation of the transcript. Text is indented and highlighted, and there are links to other formats, and to the next and previous transcript (if the transcript was uploaded as part of a series or 'family' of transcripts).

 html with sound

This is the 'fully interactive' transcript. In addition to the 'html' features above, it is also possible to interact with the audio recording, including playback or export of selected portions of the sound file, and interaction with Praat.



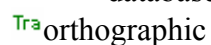
This is the original format of the transcript as produced by Transcriber.



This is a format that can be loaded into Praat to provide access to the transcript text in parallel to the recording. If you are viewing the 'layered' version of the transcript, then any layers you have selected will also be exported as tiers into the textgrid.




This is a word-by-word representation of the transcript which displays extra transcript layers if any exist. These layers, for example, might contain phonemic or syntactic information about the words taken from the CELEX database.




This is the original orthographic text of the transcript as entered into Transcriber.


If one of these icons is missing from the transcript, it's because you're already viewing the transcript in that format.

If ONZEminer is configured for correction suggestions, the submit correction  icon is displayed next to each utterance. Clicking this icon opens a window that allows you to enter a suggested correction to the transcript text.

The suggestion is sent to an email address nominated when ONZEminer was installed. The recipient of the email may then decide to implement your suggested changes to the transcript and re-upload the transcript file.

The  icon attempts a syntactic parse of the line using the Stanford Natural Language Processing Group parser. In order for the parse to be reasonable, breaks between lines should be between syntactically coherent units. Unfortunately these transcripts were transcribed before this utility was in place, and so the parser struggles with many of the lines. But if you find a shortest syntactically coherent line to try the parser on, you'll get some idea of the possibilities.

### **The layered transcript**

There are two broad representations of the transcript in ONZEminer - the orthographic representation (i.e. how the transcript was typed into Transcriber), and the layered representation. The layered representation can be accessed by clicking the  icon at the top of the interactive transcript page.

The layered transcript displays a number of selectable predefined layers, which may contain different representations of the words in the transcript. For example, in conjunction with the CELEX database, ONZEminer can be configured to look up information about the phonemic transcription, syntactic category, morphology, etc. of each word in the transcript. If this is done, then it's possible to view a phonemic representation of the transcript, simply by viewing the layered representation of the transcript, and selecting the phonology layer.

Clicking the name of the layer provides a little more information about it.

If it's a phonological layer, then you can select how to represent the phonology on the page:

- Use IPA - this option displays IPA-like characters on the page, as best as possible with the fonts already installed on your computer. This is not always very pretty, but will work ok on most systems.
- Use SIL IPA Font - this option displays the phonetic transcription using characters as defined with the SIL IPA fonts. These are freely-downloadable fonts with the full IPA alphabet (You can download them by following the IPA Font link). For this option to work, you must have at least one of the SIL IPA fonts already installed on your computer.
- Use ASCII representation - this option displays the 'raw' phonetic transcriptions as they are stored in the ONZEminer database. These follow the conventions they are followed by the CELEX database (i.e. either the 'CELEX' character set or the 'DISC' character set, depending on what options are selected for the layer)

Each line of the transcript is displayed with the selected layers stacked on top of each other, with the 'transcript' layer (if selected) at the bottom. The following example has the morphology and phonology layers selected, in addition to the transcript layer:

**we+had real+ly like to get a bit further with**  
wi:d    ɹi:lɪ    laɪk tu:    get eɪ    bɪt    fɜ:ðə    wɪð  
**we'd    really    like to get a bit further with**

Clicking on a given word allows you to view or edit the layer values of that particular word in the transcript.

Clicking on  plays the line below.

In our database, the most common pronunciation / part of speech etc is displayed on the screen. No attempt at intelligent part of speech parsing has been made. When layers are searched, any sequence which could match the search string is returned. E.g. 'drink' matches queries for both nouns and verbs.



## Layered Searches

The layered transcripts can be searched by conducting a 'layered' search.

→ *Return to the filter page, select some speakers, and select 'layered search'.*

From this page you can search the layered representation of the transcripts, using regular expressions. Only utterances spoken by the speakers you selected are returned. This search will only work if you have layered transcripts - i.e. you have uploaded the CELEX database into ONZEminer.

Unlike the orthographic search page, the layered search is by word, so regular expressions you enter will be matched against individual words, not entire utterances. For example if you enter "not\seven" to find all instances where the word *not* is followed by the word *even*, you will get no results because the regular expression "not\seven" will never match a word in any transcript (because the \s between *not* and *even* means 'any whitespace character', and words don't contain whitespace). Layered searches are achieved using a matrix of regular expressions - a search can span several successive words, and can match regular expressions at different layers.

A simple search might be one word wide and one layer deep - e.g. to return all words that could be nouns, the search would be across one word and involve only the syntax layer. To return all words containing the TRAP vowel (—), the search would be across one word and involve only the phonology layer.

A slightly more complicated search might be, for example, to find all possible prepositions that begin with a vowel which follow the word "not". This would be a search across two words, and involve three layers (the orthography layer, to match the word "not", the syntax layer to find prepositions, and the phonology layer to match words beginning with a vowel).

*layers*

- syntax
- morphology
- phonology
- orthography
- transcript

search across  words.

search regular expressions

<b>syntax</b>	<input type="text"/>	followed	<input type="text" value="^PREP"/>
<b>phonology</b>	<input type="text"/>	immediate	<input type="text" value="^[aeiouIE\{\} \&lt;&lt;"/>
<b>orthography</b>	<input type="text" value="^not\$"/>	by	<input type="text"/>

Only search transcripts for which these are the r

Tick which layers you want to search, and how many words 'wide' you want to search. For any layer you select, you can enter a regular expression for matching a single word. If a word's representation at that level matches the expression, the match is successful for that word.

If you enter more than one expression in a column, then all of the expressions for the word must match, for the match to be successful on that word.

You can match patterns against adjacent words - each column represents one word - or against words with up to one or two intervening words.

For example, to find all instances where a word's spelling ends in 'r' and the next word begins with the /ai/ diphthong:

1. tick the transcript layer and the phonology layer
2. search across 2 words
3. click set search matrix
4. enter r\$ in the transcript row of the first column (the \$ matches the end of the word)
5. enter ^ in the phonology row of the second column (^ means the beginning of the word)
6. click the « link - this opens the 'phoneme picker'
7. click the ai link - this will insert a 1 into the phonology row of the second column (1 is the CELEX DISC representation for the ai phoneme)
8. click search

Boxes at the beginning and end of the search also allow you to optionally restrict the search to be turn- or line-initial, or turn- or line-final.

### Example Layered Regular Expression searches:

In the phonology layer, the following regular expression searches translate to:

```
^n      words beginning with n
n$      words ending in n
^n..    words containing at least three phonemes, beginning with n
^n..$   words containing exactly three phonemes, beginning with n
[nm]    words containing n or m
^[nm]   words beginning with n or m
.+n.+   words which contain an n, which is neither word-initial nor
        word-final.
```

The **‘only show results from the first n transcripts’** checkbox allows you to restrict the search results to a given number of transcripts. Because layered searches can sometimes take time to complete, this option is useful when refining searches - a search can be tried, a small number of results checked, and the search adjusted accordingly.

When you initiate a search, the **progress** pane is displayed to give an indication of how long the search might take, and to provide the option of cancelling it.

→ *Try to conduct these searches*

*Sequences of prepositions and nouns*

*Words which have the dress vowel as the second phoneme*

*The word ‘the’ followed by a word beginning with the phoneme /k/*

*Words ending in schwa, followed by words beginning with /p/ or /b/.*

*Words which begin with ‘k’ in the spelling, but begin with the phoneme /n/*

Once a layered search is complete, you have the option of **exporting the results to an excel sheet**. By default, the layers you have searched are included in the excel spreadsheet. By clicking ‘options’ you can configure which other fields are also included.

→ *Repeat one of the above searches, and create a spreadsheet from the results. Configure the options so that the spreadsheet also includes word frequency information.*

You've now seen most of the features which are available to 'regular' users of ONZEminer. We're now going to move on to administrative features, which you might be using if you were setting up your own database. Any user who is not configured as an 'administrator' would not have access to the below functions.



### **Uploading transcripts and ONZEminer administration.**

Often, an ONZEminer database will run as a central server, and changing its configuration will affect all users. For the purposes of the workshop, you have your own personal version installed on your machine, so you are welcome to reconfigure it as much as you want!

#### **Adding a new corpus**

A given speaker can belong to as many or as few corpora as you like. Which corpora a speaker is in affects whether or not they appear on the filter page when you filter by corpus. Each transcript is also assigned to a corpus, but unlike speakers, a given transcript can only belong to one corpus, and it cannot be changed after the transcript is uploaded. A transcript's corpus is determined when the transcript is uploaded, and determines which folder the transcript's files are saved in.

Corpora are managed from the 'corpora' page.

To add a corpus type, enter the name in the bottom text box, and click the New  icon. To delete an existing corpus, click the Delete  icon for the chosen corpus. You cannot delete a corpus if there are any transcripts or speakers using it.

→ *Add a new corpus called 'mycorpus'.*

#### **Uploading transcripts into your corpus.**

On the Desktop you should find a folder containing the sound files and transcripts for a new speaker 'fyp98-10a'.

You might like to open the transcript files in the 'Transcriber' software, so you can get a sense of how they were created (just double-clicking the transcripts should achieve this). There are three interview transcripts, and one wordlist transcript produced by the same speaker.

→ *Select 'upload' on ONZEminer.*

Transcripts are produced using Transcriber, and then they are submitted to the ONZEminer database using this 'upload' page. When transcripts are uploaded, ONZEminer analyses them, creating new speakers or linking to existing speakers as required, and if CELEX is being used, the entire contents of the transcript text is split into words and the various CELEX layers are built.

Uploading new transcripts is a generally a 3-step process:

1. Enter the number of transcripts being uploaded. The transcripts that are uploaded together are grouped together into a transcript family.
2. Select the transcript files (called something.trs), in the order that you want them to appear in the transcript family (each transcript will have 'next' and 'previous' links to neighbouring transcripts in the family, so the order is important). You can also specify:
  - CD Number: e.g. XYZ-123 - this is the optional number or name of the CD that contains the sound recordings (something.wav). When you later view the interactive transcript, a prompt will be displayed, asking for the CD number or name that you enter here. Because our sound file is available on the machine, we can leave this blank.
  - CD Path: e.g. D:\soundfiles\something - this is the optional path to the folder that contains the sound recordings, so that the interactive transcript can find the .wav file to play. If this is incorrect, then each time you view the interactive transcript, you will have to browse to the sound recording manually. If you leave this blank, ONZEminer will look in the default position (so just leave it blank, here).
  - Corpus: this determines the folder name that ONZEminer uses to store its own copy of the .trs file, and other related files. It also determines the corpus for the speakers that appear in the transcript. Select 'mycorpus'
  - Transcript Family: e.g. XYZ - this is a descriptive name for the transcript family, which is used to determine the subfolder where the transcript files and other related files will be stored. As you enter filenames, ONZEminer will try to guess a sensible family name, but you can change this default if you like. We tend to use the speaker ID, in this case fyp98-10a.
  - Type: this is the type of the transcripts, which can be used to include/exclude the transcripts from searches.

3. Set the 'main speaker' for each transcript. This allows the utterances searched later to be narrowed to only those where the speaker is marked as the 'main speaker'.

→ *Upload the three interview transcripts as a single 'interview' family into your new corpus, and then upload the wordlist transcript as a 'wordlist'.*

You can leave the CD number and CD path blank. In this case, ONZEminer will look for the sound files in the default path. So the last step will be to put copies of the sound files in the default position:

```
C:\Program Files\Apache Software Foundation\Tomcat  
5.5\webapps\transcript\files\mycorpus\fyp98-10a\wav\
```

If you browse around the files in this area, you will get a sense of the organisation of the .wav files and .trs files. They are in folders organised into  
...\corpusname\familyname\trs\ and ...\corpusname\familyname\wav\


You can also use the upload page to re-upload new versions of previously-uploaded transcripts, by ticking the only updating existing transcripts checkbox. The effect that this has is:


- You don't have to re-specify the CD details, etc. all over again.
- ONZEminer ensures that no file locations are changed - the new transcript replaces the old one.
- Also the transcripts will keep the same family and order that they originally had - so it's possible to update a single transcript within a large family and not have it orphaned from that family or appear out of order.


Once the transcripts are uploaded, ONZEminer has its own copy of the transcripts files, so the original files can be edited, deleted, moved, etc. without affecting ONZEminer.

### **Editing the speaker details**

You can specify information about each speaker. The '**speaker**' page is a list of all speakers in the database, for which you can change basic information. Speaker information specified here can be viewed in various places in ONZEminer, and can be exported with search results.


To change an attribute of a given speaker, edit the details and then press the Save  icon on that row. You can only save the details of one speaker at a time.

The Edit  icon allows you to edit more details about the speaker, including general notes.

The Delete  icon can only be used when all transcripts referencing that speaker have been deleted.

The button labelled 'tra' conducts training on transcribed audio files in order to automatically find word boundaries. This feature of ONZEminer is still a work in progress.

→ *fyp98-10a is a female professional, born 1968.*  
*Update her details on the speaker page.*


You can also change the name of the speaker on this page, and on the Edit  page. Be aware that this is the name that ONZEminer uses when uploading new transcripts to determine whether a new speaker record must be created, or an existing one used. So changing the name may affect what happens when subsequent transcripts are uploaded. For example, imagine that a speaker is called "Noam Chomsky" in all transcript files. If you upload half of the transcripts, then change the speaker name in ONZEminer to "N. Chomsky", then when you upload the rest of the transcripts, the "Noam Chomsky" of the transcripts files will not match the "N. Chomsky" that's stored in the database, and so a second speaker record will be created.


It is advisable to:


- ensure that the same speaker is named consistently throughout all transcript files
- leave the speaker name in the database unchanged until all transcripts for that speaker have been uploaded.

### **The 'transcripts' page.**


From the **transcripts** page you can specify various attributes of transcripts you have uploaded.

To change an attribute of a given transcript, edit the details and then press the Save  icon on that row. You can only save the details of one speaker at a time.


The Delete  icon allows you to delete the transcript from the ONZEminer database, and optionally delete the transcript file and any other associated files from disk.

The Set Main Speaker  icon allows you to specify which of the speakers who appear in the transcript should be considered the main speaker. This is so 'non-main-speaker' utterances can be filtered out of transcript searches.


The last three icons allow you to view the transcript, in a format of your choice:

 plain text

Only the text of the transcript. This is suitable for use in a plain-text editor, a word-processor document, etc. There is no highlighting or indenting of text.



 html with no sound


This is a 'browser friendly' representation of the transcript. Text is indented and highlighted, and there are links to other formats, and to the next and previous transcript (if the transcript was uploaded as part of a series or 'family' of transcripts).

 html with sound

This is the 'fully interactive' transcript. In addition to the 'html' features above, it is also possible to interact with the audio recording, including playback or export of selected portions of the sound file, and interaction with Praat

The **type** attribute allows the transcript to be categorised (in our case, as an interview or wordlist) so it can be easily included or excluded in searches.

**Prompt** is the text of a prompt that is displayed when the html with sound  view of the transcript is selected. When first uploading the transcript, if the CD Number box is filled in, the prompt is automatically set to "Please insert CD "... followed by whatever text is entered. Otherwise, the prompt is left blank, and no prompt appears when the html with sound  transcript is displayed.

The **location url** is a URL for the browser to use to locate the sound file when the html with sound  transcript is displayed.

When first uploading the transcript, if the CD path box is filled in, this converted into a file:... URL, so that the browser knows where to look for the sound file when the CD is inserted. You can change this URL to another one that leads to the location of the sound file for this transcript (either a file:... URL or an http://... URL). It must be a URL which, if the name of the sound file itself is appended, would resolve to the sound recording. If location url is left blank, then ONZEminer assumes that the sound file is stored with the transcript file on the server.

The **filename** is the name of the sound file, excluding its suffix. This is appended with ".wav" to the location url in order to determine where to find the recording.

Note: if the ".wav" suffix is included (e.g. if the filename is set to interview1.wav instead of interview1, then ONZEminer will look for a sound file name to end with ".wav.wav" (i.e. interview1.wav.wav).





## Creating a Custom Layer

It is also possible create a custom layer of information about particular words, or parts of words. This could include beginning and end points for vowels, recording times at which particular analyses are taken, or codes for realisations of particular phonemic variants. If you use Praat you can download a Praat text grid, add a custom layer, and then upload it for display in ONZEminer.

*This example assumes you know how to use Praat.*

Imagine you want to identify the start and end point of key vowels produced by a speaker, and provide a phonetic transcription of the realisation of that vowel.

First, you need to define your new layer. Go to **'layers'**, select the layer type 'Time Intervals', and enter an appropriate description for your new layer (e.g. 'vowels'). Press the  button.

Next, go to a layered view of a transcript for Basil Grither – turning off all layers but the 'transcript' layer, and download a Praat textgrid by pressing the  button.

Open this textgrid, together with the corresponding wav file in Praat.  
The sound file should be available here:

```
C:\Program Files\Apache Software Foundation\Tomcat  
5.5\webapps\transcript\files\IA\BasilGrither\wav\
```

Add an interval tier to the bottom of the textgrid.

One challenge with respect to representing your analysis correctly in ONZEminer is that it needs to know which word your analysis relates to. We are in the midst of implementing automatic word segmentation for our speakers, so that the textgrid automatically has the word boundaries in the correct place. This has been done for Basil Grither, so the 'transcript' tier of the textgrid will have the word boundaries (more or less) accurately identified. However if word-alignment is not done, the transcript tier will simply have the word boundaries placed at equal intervals for each line of the transcript (this is the case for all other speakers in the demo database).

You need to:

- (a) enter your chosen intervals on your new tier, with whatever phonetic labelling you so desire; but also
- (b) ensure that the chosen interval falls within the appropriate word on the transcript tier.

For Basil Grither, (b) should hopefully just involve double checking that the word you are analysing does, indeed, appear in the expected place on the transcript tier.

For speakers for whom automatic word-alignment is not done, (b) will involve dragging the boundaries for adjacent words on the transcript tier backwards or forwards so that the boundaries for the relevant word are correctly placed.

Once you have entered a few new intervals, you can save your textgrid, and then try uploading it to your new layer. Go to the **upload** page, and select **upload praat textgrid**. Browse to your text grid and select upload.

Make sure that 'transcript' is mapped to the 'transcript' layer, and the tier you created is mapped to the new layer you have created. Then select **update layers**.

Now navigate back to the layered view of the transcript you have been working on. Select your new layer, and click **display layers**. Your new layer should be available for viewing.

This layer is now also available on the 'layered search' page, and its contents can optionally be included in excel spreadsheet representation of the search results.

---

We hope you've had fun playing with ONZEminer. We have primarily developed this software for our own use, but are happy to make it available to others for whom it might be useful.

However please note that future maintenance and support relies on the availability of future funding and so can not be guaranteed.

<http://www.ling.canterbury.ac.nz/jen/onzeminer>