Statistics Application

I. Overview and Purpose

The purpose of the statistics application is to allow users to extract numerical data and phrases from TRS files. The program provides several of types of "outputs" (instructions to process TRS segments in a specific way to produce data). These outputs can be saved in sets called configurations. Configurations can then be "run" on TRS files. The program itself does not provide any means of viewing the computed output – instead, the data is exported as spreadsheet file (comma-separated value (*.csv) format) that can be viewed and further manipulated using Excel.

II. Terminology

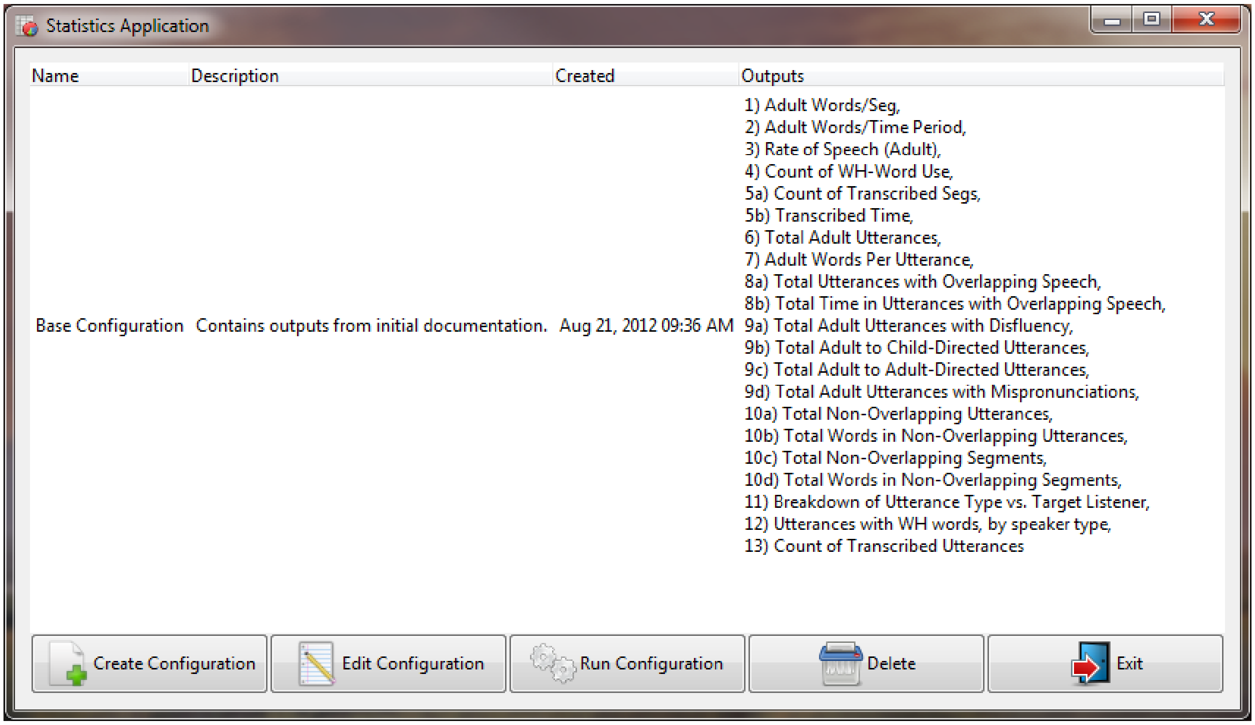
Below are some definitions that may be useful to look at before reading the rest of this manual:

Term:	**Definition:**
Segment	A LENA-delimited section of speech data from a TRS file. In some cases, LENA can make mistakes when it creates segment boundaries. Mistakes generally take one of two forms: 1) LENA may mistakenly think that a single utterance is actually two. In this case, LENA will create two segments when there should be only one. Transcribers can manually link these segments using I/C codes (see the transcriber manual for more information). The program keeps track of these links. 2) Alternatively, LENA may mistakenly consider two consecutive utterances to be a single utterance. In this case, transcribers can use the dot operator to separate the utterance codes. The program scans for this syntax and considers such a transcription as two separate, independent segments. There are two types of segments, described in the rows immediately below.
Unlinked Segment	This is a section of speech data *as it was originally delimited by LENA.*
Linked Segment	This is a group of Unlinked Segments *that have been chained together by transcriber I/C codes.*
Output	An output represents a particular type of statistic that you wish to include in the exported spreadsheet. There are different types of outputs, each with its own kinds of settings. Each output has zero or more "filters" (described below). **Example**: you could create an output to count the number of occurrences of "WH" words.
Filter	A filter can be thought of as a criterion that is used to "strain out" segments that you don't want to use when performing the calculations for an output. Alternatively, it can be thought of as a criterion that is used to include segments that you do want to use when performing the

	calculations for an output. Filters can be set to operate on linked or unlinked segments. For more information, see the “Filters section” of this document. **Example**: you could create a filter that operates on the Transcriber code 1 (speaker type), straining out all segments that have an “Uncertain” speaker type.
Search Term/Pattern	Many types of outputs allow you to specify a “search term.” This text is searched for in the transcription phrases of the segments that the output goes through (exactly which segments it goes through depend upon the filters – see above). This works as one would expect basic search to. However one can also enter some special characters to search for patterns, (rather than exact words) in the phrase. Some common patterns are provided in a drop-down list in the output creation window. For more on this, see the output type sections below. If you need to make use of these special characters to search for a pattern (and your pattern isn’t listed in the “common patterns” dropdown list in the output creation window), feel free to consult your local computer scientist. **Search Example**: enter “every” in the search term box to search for occurrences of “every” in segment phrases. Note that in addition to “every”, this will also match “everywhere”, “everything.” (To search for only “every” as a word on it’s own, select the “Specific word” pattern from the dropdown and type “every” into the selected region as indicated.) **Pattern Example**: It is possible to search for *all words* beginning with the letter “e” using the pattern “\be[^\s]+\b”.
Configuration	Outputs can be grouped together into sets called configurations. These configurations can be saved. When you want to process a file, you select a configuration you’ve created, and select a TRS file. As the TRS file is processed, the program performs the calculations necessary for each of the outputs contained in the configuration. Finally, a spreadsheet file is generated, which contains a separate section with information for each output in the configuration. **Example**: You could create a configuration to compare two groups of speakers: children and adults. Such a configuration could contain two outputs: one that counts the number of words in all child-spoken segments (using filters and patterns, as described above), and another that counts the number of words in all adult-spoken segments. The configuration could be run multiple times on different TRS files. Each time it is run it will produce a spreadsheet file containing the results from both outputs.

III. Main Window

The main window displays a table in the middle, with a set of buttons across the bottom. Each row in the table corresponds to a saved configuration.



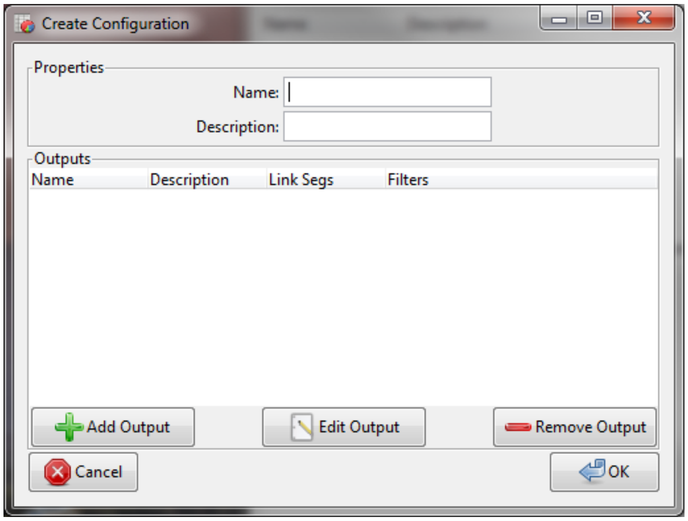
Name	Description	Created	Outputs
Base Configuration	Contains outputs from initial documentation.	Aug 21, 2012 09:36 AM	1) Adult Words/Seg, 2) Adult Words/Time Period, 3) Rate of Speech (Adult), 4) Count of WH-Word Use, 5a) Count of Transcribed Segs, 5b) Transcribed Time, 6) Total Adult Utterances, 7) Adult Words Per Utterance, 8a) Total Utterances with Overlapping Speech, 8b) Total Time in Utterances with Overlapping Speech, 9a) Total Adult Utterances with Disfluency, 9b) Total Adult to Child-Directed Utterances, 9c) Total Adult to Adult-Directed Utterances, 9d) Total Adult Utterances with Mispronunciations, 10a) Total Non-Overlapping Utterances, 10b) Total Words in Non-Overlapping Utterances, 10c) Total Non-Overlapping Segments, 10d) Total Words in Non-Overlapping Segments, 11) Breakdown of Utterance Type vs. Target Listener, 12) Utterances with WH words, by speaker type, 13) Count of Transcribed Utterances

The first three table column names should be fairly self-explanatory. The “Outputs” column lists the descriptions for each of the outputs in the configuration.

The buttons along the bottom of the window allow you to create, edit, run (perform the computation and export the results to a spreadsheet file), or delete the selected configuration in the list. These functions are described in more detail below.

IV. Creating a New Configuration

To create a new configuration, click the “Create Configuration” button in the main window. This causes the “Create Configuration window to appear:

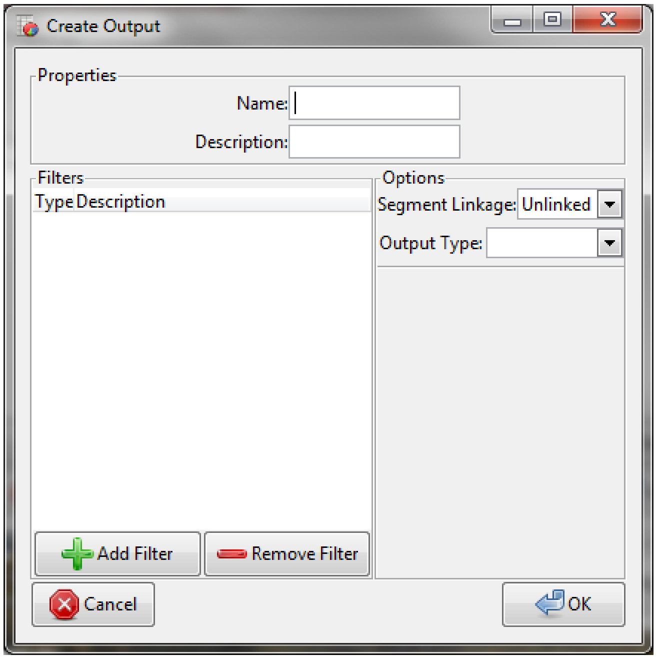


The two boxes at the top allow you to give the configuration a name and a description (This is the text that will display in the main window's table after the configuration is created).

The "Outputs" section consists of another table with buttons that allow you to add, edit, or remove outputs. After you have created outputs, they appear in the table.

<u>Adding Outputs:</u>

To add a new output, click the "Add Output" button. This brings up the "Create Output" window:



Segment Linkage: Unlinked

You can give the output a name and description using the two text boxes at the top of the window. The remaining functionality of this window is described in the sections below.

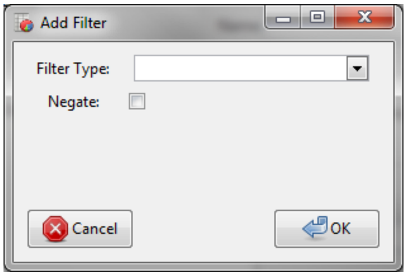
1) <u>Filters</u>

The "Filters" frame allows you to add or remove filters for this output. Filters control which segments the program goes through when calculating the output. You can add multiple filters to a single output. In such cases, the program will perform a logical "AND" operation on all of the filters.

For example, suppose you have a filter to include segments with a speaker code of "FAN." Now suppose you add another filter to include only those segments that start after the 10 minute mark. When the output is calculated, it will make use of only those segments whose speaker is "FAN" AND whose start time is greater than 10 minutes.

<u>Adding Filters:</u>

To add a filter, click the corresponding button at the bottom of the frame. This brings up the "Add Filter" window.



Filter Type:

Negate: □

There are different types of filters. You can choose which one you'd like using the "Filter Type" dropdown. Upon selecting an option, the window changes to display special inputs for that option.

Each filter type is discussed in detail further down.

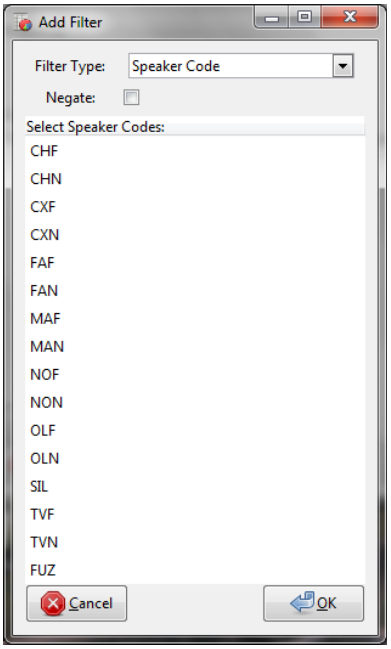
Negate Checkbox

By default, filters operate from an *"inclusive"* perspective. That is, segments that match the filter criteria will be *included* in output calculations (all others will be excluded).

The "Negate" checkbox can be used to invert the effects of the filter. When it is checked, the filter operates from an *"exclusive"* perspective. That is, segments that match the filter criteria will be *excluded* from output calculations (all other segments will be included).

Filter Types

- Speaker Code – this type of filter allows you to include/exclude segments that have one or more speakers. This is the "speaker code" that LENA has recorded for the segment (not the "speaker type" transcriber code).



Filter Type: Speaker Code

Negate: □

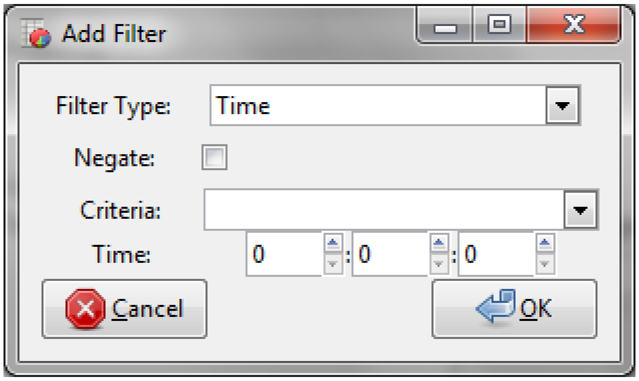
Select Speaker Codes:

CHF
CHN
CXF
CXN
FAF
FAN
MAF
MAN
NOF
NON
OLF
OLN
SIL
TVF
TVN
FUZ

To select a speaker, click the corresponding code in the list. To select additional speakers, hold down the control key and click another code. To unselect a previously selected speaker, hold down the control key and click the selected code again.

If you select multiple speaker codes, they will be combined using a logical “OR” operation. For example, suppose you select “FAN” and “MAN.” In this case, the filter will include segments whose speakers are “FAN” OR “MAN.”

- Time – this filter type can be used to include or exclude segments whose start/end times are before/after certain points.



Add Filter

Filter Type: Time

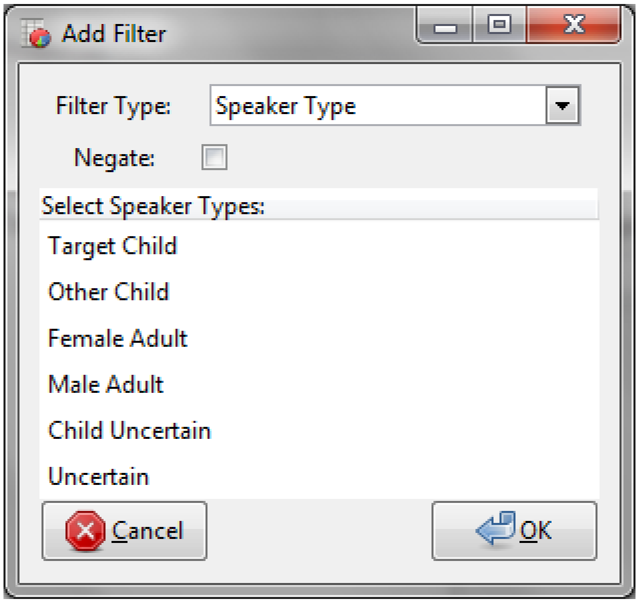
Negate: □

Criteria:

Time: 0 : 0 : 0

The “Criteria” dropdown allows you to pick whether the segment must start before, start after, end before, or end after a specific point in time. The “Time” inputs allow you to specify that point in time.

- Speaker Type – This filter type can be used to include/exclude segments whose speaker type (transcriber code 1) matches one or more selected values.



Add Filter

Filter Type: Speaker Type

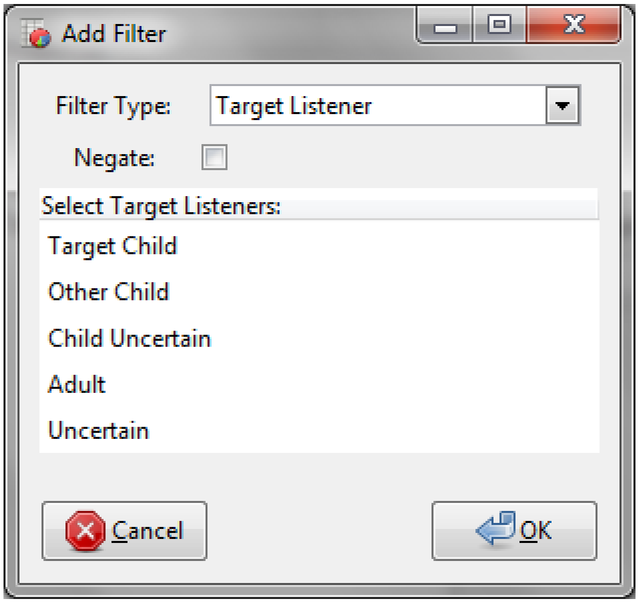
Negate: □

Select Speaker Types:

Target Child
Other Child
Female Adult
Male Adult
Child Uncertain
Uncertain

You can select/unselect multiple options by holding down the control key and clicking on the corresponding row. If you select multiple options, the filter will “OR” them together.

- Target Listener – This filter type can be used to include/exclude segments whose target listener (transcriber code 2) matches one or more values.

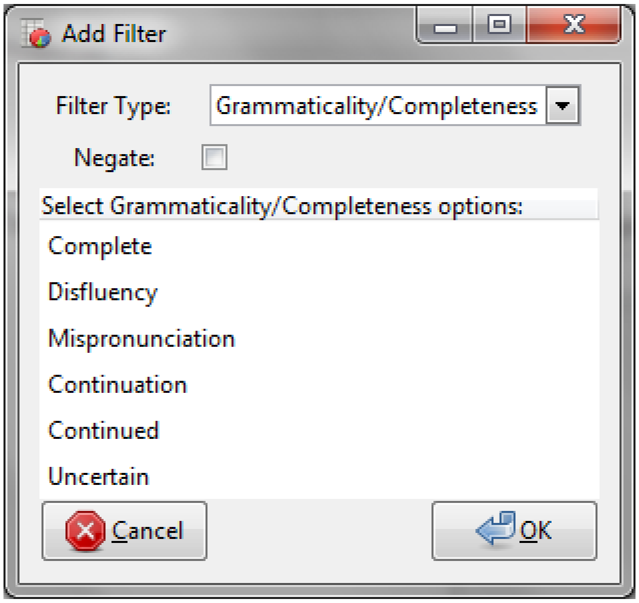


Filter Type: Target Listener

Negate: □

You can select multiple options using the control key.
These will be "ORed" together.

- Grammaticality/Completeness – This filter type can be used to include/exclude segments whose grammaticality/completeness (transcriber code 3) matches one or more values.

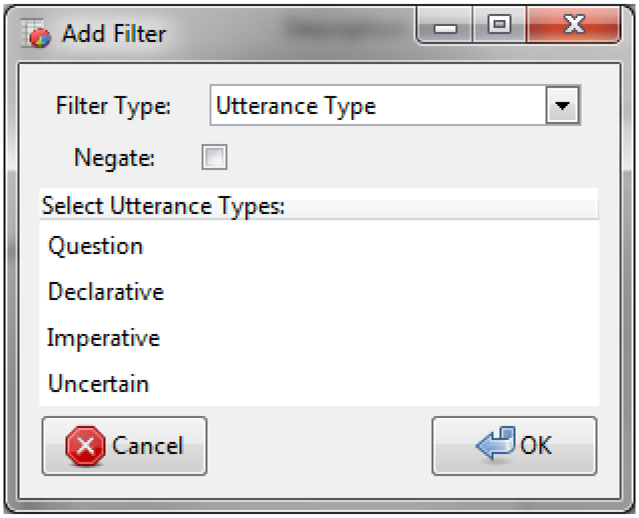


Filter Type: Grammaticality/Completeness

Negate: □

You can select multiple options using the control key.
These will be "ORed" together.

- Utterance Type - This filter type can be used to include/exclude segments whose utterance type (transcriber code 4) matches one or more values.



Filter Type: Utterance Type

Negate: □

You can select multiple options using the control key.
These will be "ORed" together.

- Overlapping Vocals – This filter type can be used to include/exclude segments that contain overlapping vocals. Segments contain overlapping vocals if LENA has given them a speaker code of OLN or OLF (overlapping near, overlapping far), or if they include the angle brackets (<>) in the transcription phase (see transcriber manual for additional details).
  This filter type has no special options.

After you have selected your options, click OK. The filter is appended to the list in the "Create Output" window.

Removing Filters

To remove a filter, select it from the list in the "Create Output" window and click the "Remove Filter" button.

2) Options

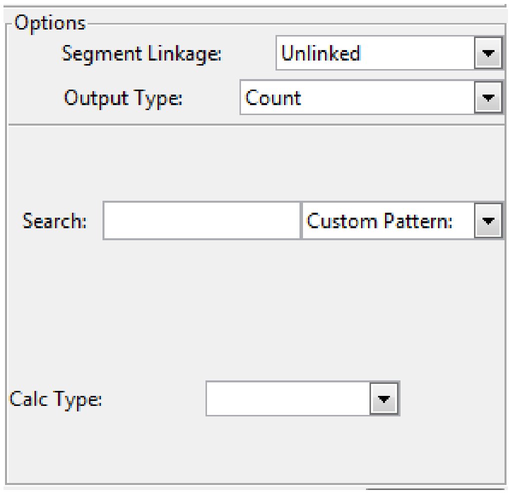
The "Options" frame allows you to set the type of output you wish to create as well as the type of segment (linked or unlinked) you would like the calculations to make use of.

Selecting an output type from the dropdown list causes additional options to appear in the area below the dividing line. Each of the output types is described in detail below:

Count Outputs

These outputs search for something in the segment phrase and count them.

The "Search" text box allows you to specify *what* you want to count. You can enter some regular text to search for, or choose a predefined pattern from the dropdown list. These predefined patterns let you search for things like: any word, a specific word, a WH word, or any transcription (the latter will match any segment with a transcription phrase, so it used for things like counting the number of transcribed segments).



Options

Segment Linkage: Unlinked

Output Type: Count

Search:

Custom Pattern:

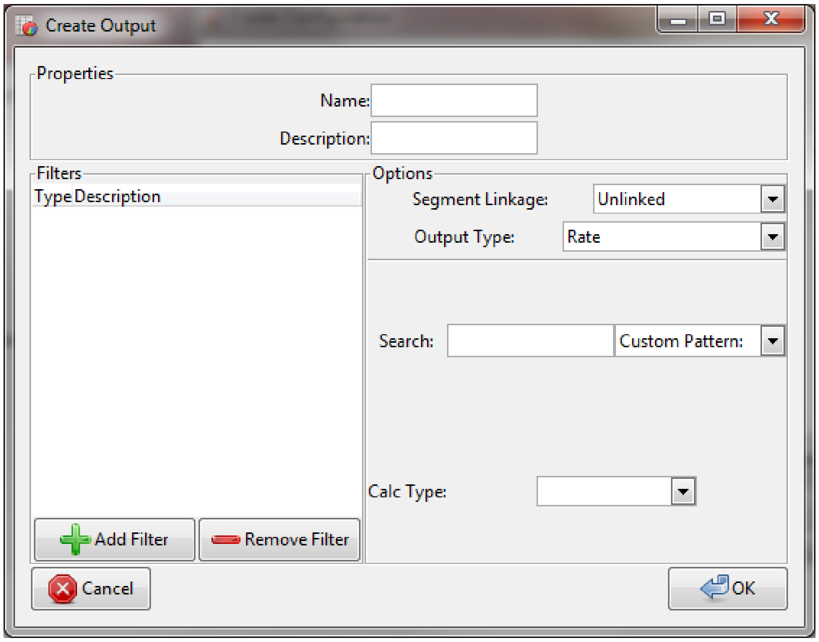
Calc Type:

The “Calc Type” dropdown allows you to specify *how* you want to count things. There are three options:

1) “Per Seg” – This option causes the program to search each segment individually. The resulting spreadsheet will contain one row for each segment in the TRS file, each with an individual count number beside it. In addition, it will add up all of the individual counts and display the total in an extra row at the bottom.
2) “Avg Across Segs” – This option cause the program to keep a running sum of all occurrences of the search text/pattern as it goes through the TRS file (across all segments). After it reaches the end of the file, it divides the final sum by the number of segments it went through. The resulting average is printed on a single row in the spreadsheet.
3) “Sum Across Segs” - This option cause the program to keep a running sum of all occurrences of the search text/pattern as it goes through the TRS file. When it reaches the end of the file, it prints the final sum in a single row in the spreadsheet.

Rate Outputs

These outputs search for something in the segment phrase, count it, and then divide the resulting number by the length (in seconds) of the segment they were found in. In other words, you get a measure of count/time.



Segment Linkage: Unlinked

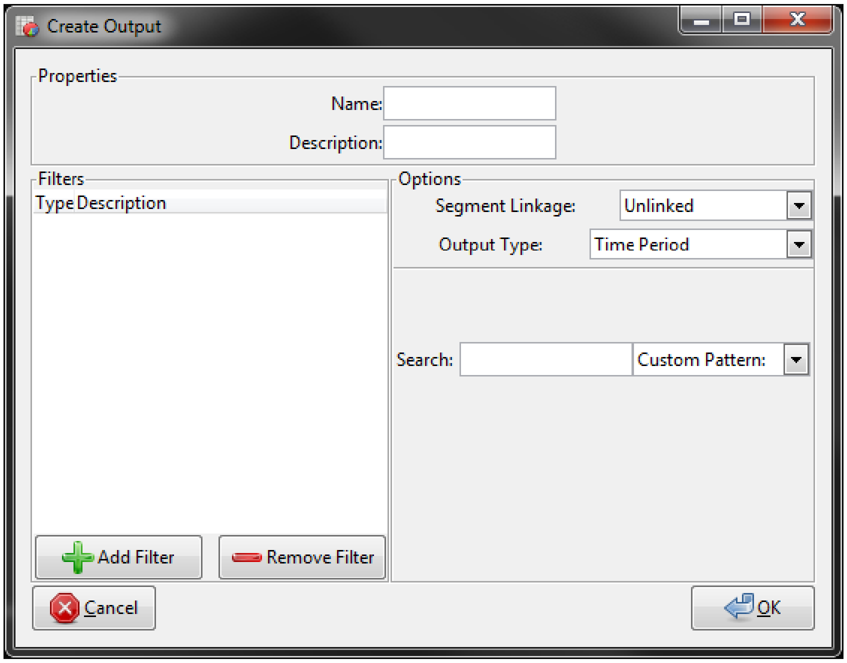
Output Type: Rate

As with the count output, you can enter some text to search for, or choose an existing pattern. The "Calc Type" dropdown allows you to select one of the following options:

1) Per Seg – This causes the rate calculation to be done individually for each segment whose phrase contains a search match. The final spreadsheet will contain a row for each segment with a match.
2) Avg Across Segs – This causes the rate calculation to span all of the segments containing a match. In other words, the program scans the TRS file, keeping a running sum of the number of search matches in each segment. It also keeps a running sum of the lengths of segments with matches. After the scan is completed, the rate calculation is performed as follows:
(sum of counts of matches in all segments) / (sum of lengths of all segments with matches)

Time Period Outputs

These outputs search for segments containing a particular search phrase/pattern, then sum the lengths (elapsed time) of the segments that contain a match.



Create Output

Properties

Name:

Description:

Filters

Type Description

Options

Segment Linkage: Unlinked

Output Type: Time Period

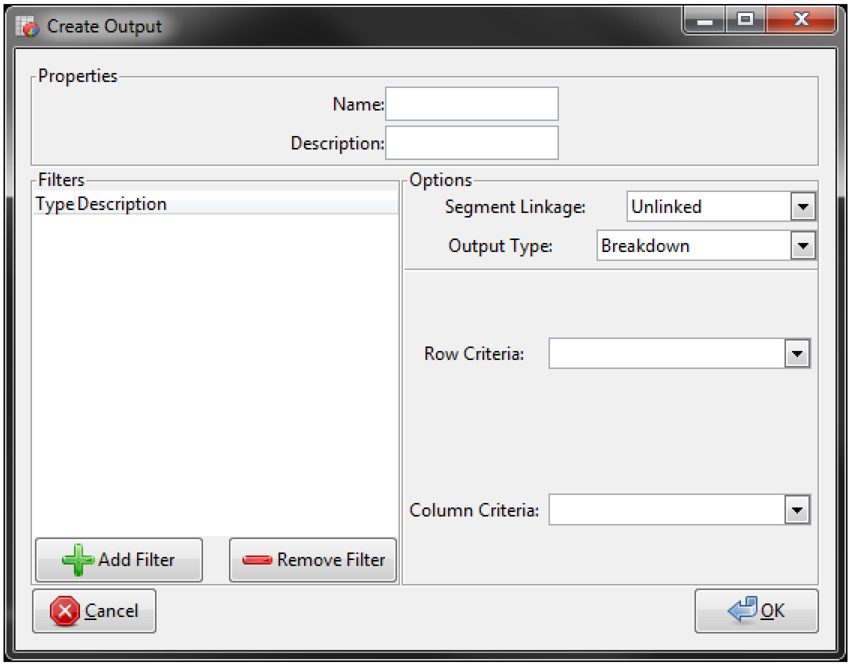
Search: Custom Pattern:

The search box is the only input for this type of filter. The spreadsheet will contain a single row for this output, displaying the total time in the format hh:mm:ss.

The program takes any segment overlap into account (this is particularly an issue for linked segments), ensuring that overlapping time periods are never counted twice.

Breakdown Outputs

This type of output generates a table of numbers. Each of the two axes in the table represents a particular property (or “criteria”) that a segment can take on. Each cell in the table contains a number indicating the number of segments that have both the (corresponding) x-axis and y-axis criteria.



Create Output

Properties

Name:

Description:

Filters

Type Description

Options

Segment Linkage: Unlinked

Output Type: Breakdown

Row Criteria:

Column Criteria:

For example, one could set the x-axis to represent "Speaker Type" (transcriber code 1) property, and the y-axis to represent the "Target Listener" (transcriber code 2) property.

This would produce a table in the spreadsheet file. There would be a single row for each Target Listener code, and a single column for each Speaker Type code. Each cell would contain the number of segments that have both the row's target listener and the column's speaker type.

Example:

Target Listener/Speaker Type	M	F	T	O	C	U
T	0	0	0	0	0	1
O	1	0	0	0	3	0
C	0	0	5	0	0	0
A	0	0	0	1	0	0
U	0	2	0	0	0	0

In the table able, 5 segments were found that have Target Listener code "C" and Speaker Type code "T". There were no segments with Target Listener code "T" and Speaker Type code "M". Etc.

Use the criteria dropdowns to select the codes that the rows and columns will represent.

<u>Behaviour on Multi-Character codes</u>

The "Grammaticality/Completeness" (transcriber code 3) is the only transcriber code that allows for multiple character (eg. "MF", "IC", etc.).

If you select “Grammaticality/completeness” as one of your axis criteria, the axis headings will contain only the single code characters (F, D, M, C, I, U).

Segments that contain codes with multiple characters will increment two cells in the table. For example, a segment with code MF will bump up the M count AND bump up the F count.

<u>Behaviour on linked segments</u>

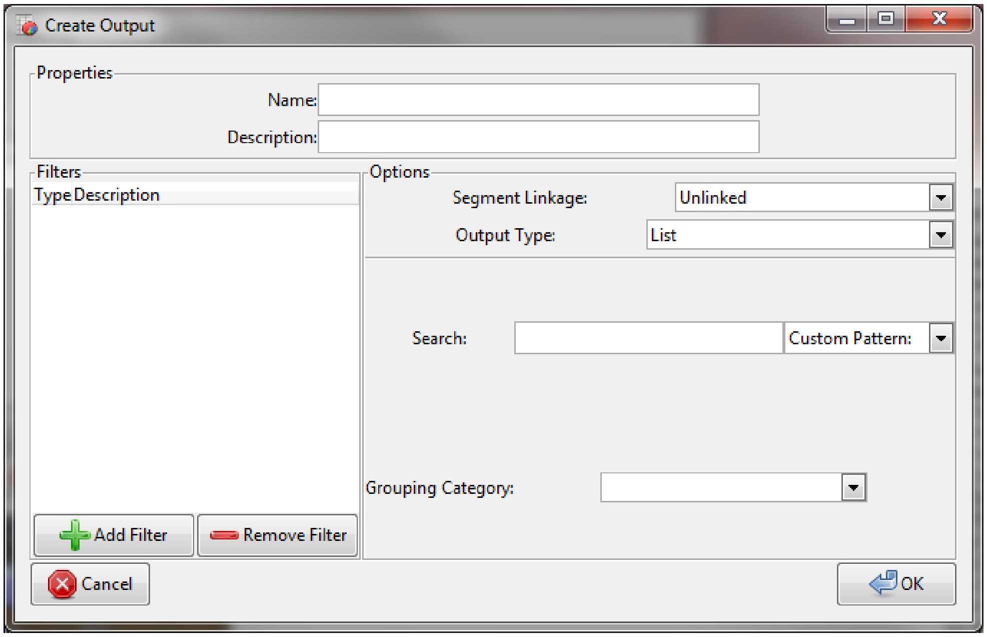
When the segment linkage is set to “linked,” only the codes of the first segment in the chain will be used when generating the breakdown table. (Note: In most cases, this does not mess with the accuracy of the data, since linked segments only exist where LENA has mistakenly split a single segment in two – therefore (in theory) linked segments’ transcriber codes should almost always be the same).

One noteworthy effect that this does have, however, is that it messes with the I and C code 3 counts. Since only the first segment in the chain is considered (and it will have an I code), the corresponding C code (at the end of the chain) will never increment any of the cells in the table.

Note: It was difficult to decide upon the appropriate behaviour in these situations. If you would like something changed, please don’t hesitate to let your programmer know.

<u>List Outputs</u>

This type of output can be used to search for segments with a particular text pattern and list them, grouped by a particular transcriber code.



Segment Linkage: Unlinked

Output Type: List

Search: Custom Pattern:

For example, you could use the pattern dropdown box to search for “Any Transcription”, and set the “Grouping Category” dropdown to “Speaker Type.” This would generate a list of all segments that have a transcription phrase. The list would be grouped by Speaker Type code, as follows:

M

<segments with transcription phrase and speaker type M>

F

<segments with transcription phrase and speaker type F>

T

<segments with transcription phrase and speaker type T>

etc.

If the grouping category is “Grammaticality/Completeness” (transcriber code 3), the code can have multiple characters (eg. MF, IC, etc.). In these cases, the list headings will each be single code characters (eg. M and F will be separate groups). Segments with multi-character code (eg. MF) will appear in the list multiple times (once under the M heading, and once under the F heading).

When linked segments are being used, only the first segment in the chain will be considered. As with the breakdown output, the desired behavior here was a little difficult to decide upon. Feel free to discuss things with your programmer if you’d like different behavior here.