Guide to Keith Rayner Eye Movements in Reading Data Collection

Matthew J. Abbott

September 22, 2015

Contents

Data Processing Software

Here is a brief list of URLs for the different data processing programs that are mentioned throughout the collection (metadata and README files). They are listed in roughly the order one would use to process the data files.

Generally speaking, there are 3-4 steps to processing eye movement reading data collected using an SR EyeLink eyetracker and the UMass EyeTrack software:

1. Convert raw EDF file to ASC (plain text) format (**edf2asc.exe** or **EDF Converter** supplied by SR Research)
2. Clean artifacts from the data, including blinks, long fixations, indications of track loss, merge very short (< 80 ms) fixations within an adjacent fixation if within 1 character (**TimDrop.pl** for single-line sentences or **EyeDoctor** for passages)
3. (Display change experiments only) Extract information pertaining to the timing of critical display changes (i.e., using **jhook5b.pl**)
4. Extract relevant dependent measures (e.g., *first fixation duration*, *fixation probability*) from processed data (**EyeDry**, **DAMPDRY**)

TimDrop.pl

TimDrop.pl can be found here: https://sites.google.com/site/drtimothyjslattery/home/software

An excellent description of TimDrop.pl and subsequently mentioned data processing programs can be found on the same page ("PerlProgramDescriptions.docx").

Tim Slattery, a former Rayner Lab post-doc at UCSD, now Lecturer at Bournemouth University (UK), wrote a number of useful software tools. Among them is TimDrop.pl, Perl program that takes a list of **ASC** files produced by the UMass EyeTrack software and outputs one **da1** file ASC file. These files are compatible with Chuck Clifton's EyeDry software. The program also outputs a summary file that indicates the number of trials deleted for each participant (i.e., due to blinks on a target word), and other useful data processing information. It takes as input a **list file** containing the ASC filenames (one row per file), a **parameter file** describing properties of the experiment including screen resolution, and a **count file** that contains the character locations of target words in the sentences. Descriptions of these files are located in a later section of this document, and examples can be found throughout the data collection.

jhook5b.pl

jhook5b.pl can be found here: https://sites.google.com/site/drtimothyjslattery/home/software

jhook5b.pl takes a list of ASC files and a **boundary file** as input and outputs a file containing information regarding the timing of display changes and the fixations surrounding the boundary changes.

Questions.pl

Questions.pl can also be found here: https://sites.google.com/site/drtimothyjslattery/home/software

Questions.pl processes a list of ASC files and produces a file containing information about responses to comprehension questions.

EyeDry

EyeDry can be found at this link within the **dataanal** directories: https://blogs.umass.edu/eyelab/software/

EyeDry is a DOS utility written and maintained by Chuck Clifton (UMass Brain and Psychological Sciences) that computes a wealth of dependent measures (i.e., *first fixation duration*, *gaze duration*, *initial landing position*) from **da1** files.

DAMPDRY

DAMPDRY can be found within the **dataanal** directories at: https://blogs.umass.edu/eyelab/software/

Data File Types

This guide will walk through the different types of data structures and files that are contained in the data collection.

Count File

A count file (**.CNT** extension) delimits the different regions for subsequent analysis in **Eyedry** (and is also used for data processing using **TimDrop.pl**. Column definitions are as follows.

C1: Item number

C2: Condition number

C3: Number of regions

C4: Character onset of region 1 (always 0)

C5: Character onset of region 2 (always the space before a word begins)

C6 onward: Character onsets of subsequent regions

Boundary File

A **boundary file** indicates the locations of display change regions for each item and condition combination, and is used in conjunction with **jhook*.pl** to analyze what happened in and around display change regions. Column definitions are as follows.

C1: Item number

C2: Condition number

C3: Boundary location (pixels)

Example:

1 1 474

1 2 474

1 3 452

1 4 452

DAMPDRY Output

DAMPDRY is a DOS program that produces global eye movement measures (i.e., number of fixations, total sentence reading time) from raw ASC files. These measures are typically reported in moving window type studies (i.e., Bélanger et al., 2012; *Psych Sci*) Column definitions are as follows, drawn from the original documentation for the program.

The first section contains the trial-by-trial data. The second section contains the summary by condition for each subject. You can choose to not output the trial-by-trial data.

If you choose to include trial-by-trial data, then the following data are output for each trial (C# = column number).

C1 = Subject.

C2 = Condition.

C3 = Item.

C4 = Reading time (msec).

C5 = Reading rate (wpm).

C6 = Number of words.

C7 = Average fixation duration (forward).

C8 = Number of fixations (forward).

C9 = Average saccade length (forward, not including return sweeps).

C10 = Number of saccades (forward, not including return sweeps).

C11 = Average fixation duration (regressive).

C12 = Number of fixations (regressive).

C13 = Average saccade length (regressive, not including returns to prior lines).

C14 = Number of saccades (regressive, not including returns to prior lines).

C15 = Average saccade duration (forward and regressive combined).

C16 = Average return sweep duration. This should always be zero for single line stimuli.

C17 = Number of return sweeps. This should always be zero for single line stimuli.

C18 = Number of short return sweeps.

A short return sweep is a return sweep that doesn't reach the beginning of the next line. For example, if a reader is the end of the first line of text and makes a return sweep that lands on character 15 of the second line of text, and then makes another saccade to character 3 of the second line of text, then this is categorized as a short return sweep. I separate these because it is unclear whether the short fixation should count in the average, and it is also unclear whether the saccade following the short fixation should be called a regression because the reader is actually trying to move forward in the text. The program allows you to set the cutoff for short return sweeps. The default is 15 (i.e., if a return sweep doesn't pass character 15, it's labeled as a short return). The number of short return sweeps should always be zero for single line stimuli.

C19 = Line number of last fixation.

When a stimulus occupies more than one line of text, this is useful for determining if the entire text was read. For example, if a text has 4 lines, but the line number of the last fixation is 2, then the entire text might not have been read. This is labeled as an "incomplete trial" in the program. Of course, the reader might have read the entire text, regressed to line 2, then ended the trial. This would be a complete trial that would be erroneously labeled as an incomplete trial. The program allows you to exclude or include incomplete trials in the across-trial averages. The default is exclude incomplete trials. For single line sentences, this measure isn't very useful and you should probably include incomplete trials.

The trial-by-trial data are listed first in the output file. Note that condition 99 is mostly zeros. This is because the maximum number of conditions allowed is 20. Although you only have 11 different conditions, the program assumes sequential condition numbers. Thus, if you use 99 as a condition, the program assumes conditions 1 through 98 are also used. Anything listed for condition 99 is garbage.

The across-trial averages follow the trial-by-trial data. The across-trial averages are always output (i.e., you can't suppress the averages). Note that there are fewer data columns in the across-trial averages.

The following data are output as across-trial averages (C# = column number).

C1 = Subject.

C2 = Condition.

C3 = Number of trials included in across-trial averages.

C4 = Average reading time.

C5 = Average reading rate (wpm).

C6 = Average number of words.

C7 = Average fixation duration (forward).

C8 = Average number of fixations (forward).

C9 = Average saccade length (forward, not including return sweeps).

C10 = Average number of saccades (forward, not including return sweeps).

C11 = Average fixation duration (regressive).

C12 = Average number of fixations (regressive).

C13 = Average saccade length (regressive, not including returns to prior lines).

C14 = Average number of saccades (regressive, not including returns to prior lines).

C15 = Average saccade duration (forward and regressive combined).

C16 = Average return sweep duration.