

The Discourse Attributes Analysis Program (DAAP)

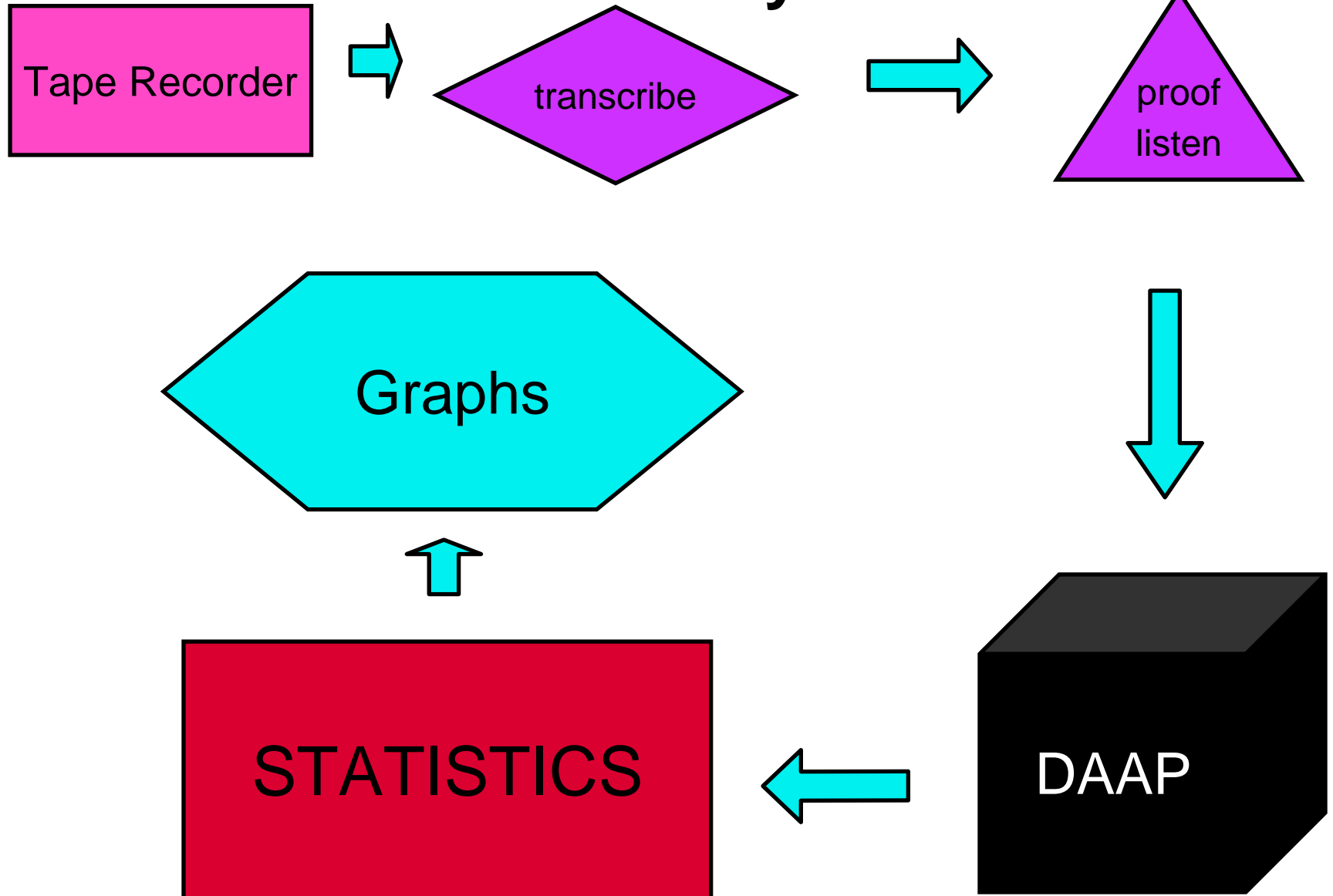
Version 1.1

Bernard Maskit

and

Wilma Bucci

Overview of the Basic DAAP System



Transcribing Rules

There are special rules for:

- Change of speaker (markers for turns of speech)
- Categories (markers for larger or smaller segments of text, perhaps including several turns of speech)
- Parenthetical remarks and non-speech (cough, sneeze, wheeze, etc.)
- hm's, umh's, etc., written as MM; but 'oh', 'ah', etc. are separate words
- Compound words (must use underscore rather than hyphen)
- Incomplete words (use hyphen)
- Numbers, ten and under, must be written out
- Disambiguations of: *kind, know, like, mean, well*
- Pauses (**BIG** topic - Time markers will be discussed below)

More special transcription rules

- ‘okay’, but not ‘ok’.
- ‘alright’, but not ‘allright’ or ‘all right’.

Transcribing Rules

- To obtain the latest version of the transcribing rules, go to:
- www.referentialprocess.org
- Click on the DAAP page, and download the file:
- DAAPTranscriptionRules2.pdf
- Please post any questions on the forum.

Prooflistening

There is not much to say about prooflistening; there are always errors that need to be corrected. Unfortunately, the basic rule of all writing, including transcribing, is:

There is always at least one more error.

DAAP Error Messages

- DAAP records errors in the LOG file. If your file is named Session.txt, then the LOG file is named SessionLOG.txt.
- After your file is run, a message appears on the screen telling you how many errors there are in the LOG file.

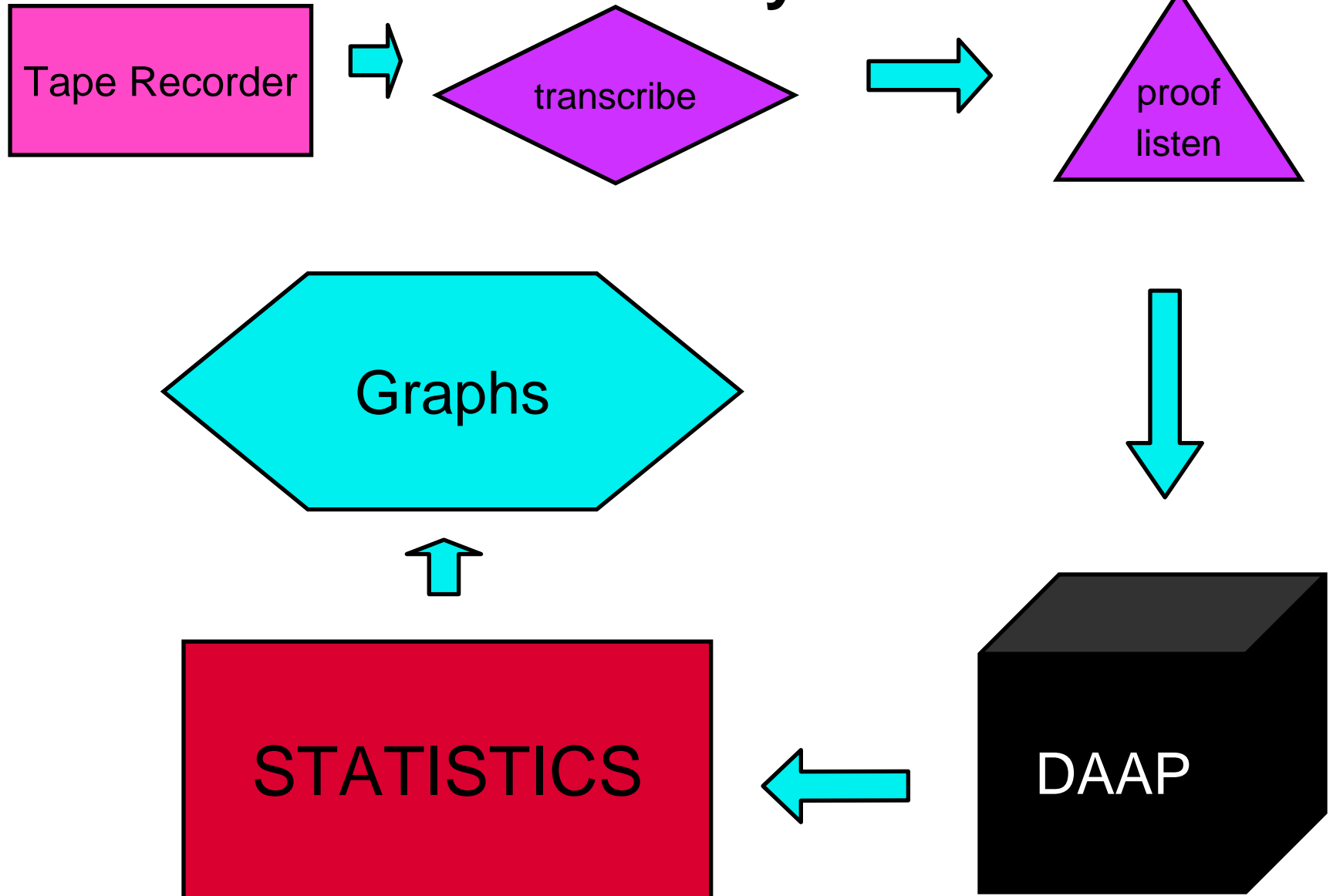
Usual Errors

- Strange words, containing numbers or hyphens in the middle, or strange symbols, such as em-dashes or en-dashes. The new LOG file gives you the word number (for use with the MTT file) for each of these.
- Misplaced backslash (\) (DAAP uses the backslash as its control symbol). The backslash must occur as the first symbol on a line. Otherwise DAAP crashes and sends you an error message on the screen.

More Errors

- Parentheses. DAAP does not process anything inside parentheses (round brackets only.) If you close a parentheses that has not been opened, DAAP shuts down and sends you an error message on screen.
- If you forget to close a parentheses, DAAP sends you an on screen error message; you can use the MTT file to find the source of the problem.

Overview of the Basic DAAP System



RA Dictionaries

We currently use seven RA dictionaries:

- ❖ The Weighted Referential Activity Dictionary (WRAD)
- ❖ Positive Affect (AffP)
- ❖ Negative Affect (AffN)
- ❖ Mixed Affect (AffZ)
- ❖ Disfluency (DF)
- ❖ Reflection (Ref)
- ❖ Sensory-Somatic (SenS)

RA Dictionaries (cont.)

- The WRAD, which is a style dictionary, is in its third incarnation. We have a new technique that seemed to work well for making the Italian WRAD; we will perhaps use this new technique to make a fourth generation WRAD.
- The other dictionaries contain words that are primarily content, rather than style, and are unweighted. They are currently in a state of flux; this will be discussed later.

Black Box Overview

- DAAP reads all the dictionaries.
- DAAP reads the transcript file word by word, tracks who is speaking, tracks categories, ignores parenthetical remarks, and compares each word with each dictionary, and tracks the WRAD weights.
- So, at first glance, DAAP produces, for each turn of speech, 7 lists of numbers, one list for each dictionary, where each list contains a number for each word in the turn of speech.

Black Box Overview

- For the unweighted dictionaries, this first number, which is called the *raw score*, is either 0 or 1. The raw score is 1 if the word is in the dictionary; it is 0 otherwise.

Black Box Overview

- For the WRAD, the raw score is already more complicated. We first have a *pre-raw* score, which is either the weight that the WRAD assigns to this word, or is 0, if the word is not in the WRAD. To get the *raw score*, these pre-raw scores are translated into positive scores between 0 (very low RA) and 1 (very high RA), with .5 as the neutral value; words not in the WRAD have raw score of .5.

Basic Example

- Text: “Well, I think I’m happily excited, worried, anxious and my back hurts, / (chair creaks) // emotionally.”

	AffP	AffN	AffZ	DF	Ref	WRAD	SenS
Well	0	0	0	1	0	0.375	0
I	0	0	0	0	0	0.125	0
think	0	0	0	0	1	0	0
I	0	0	0	0	0	0.125	0
m	0	0	0	0	0	0	0
happily	1	0	0	0	0	0.5	0
excited	1	0	0	0	0	1	0
worried	0	1	0	0	0	0.5	0
anxious	0	1	0	0	0	0	0
and	0	0	0	0	0	1	0
my	0	0	0	0	0	0.8125	0
back	0	0	0	0	0	1	1
hurts	0	1	0	0	0	0.5	1
emotionally	0	0	1	0	0	0.5	0

The Smoothing Operation

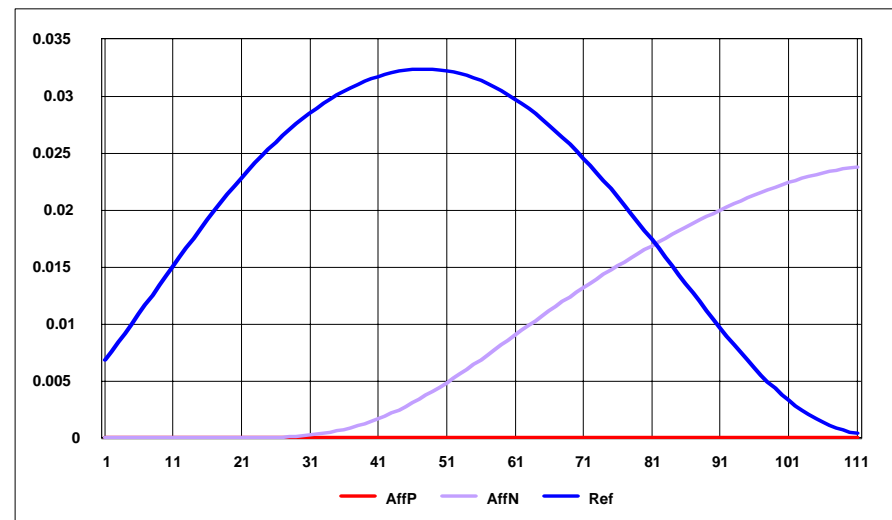
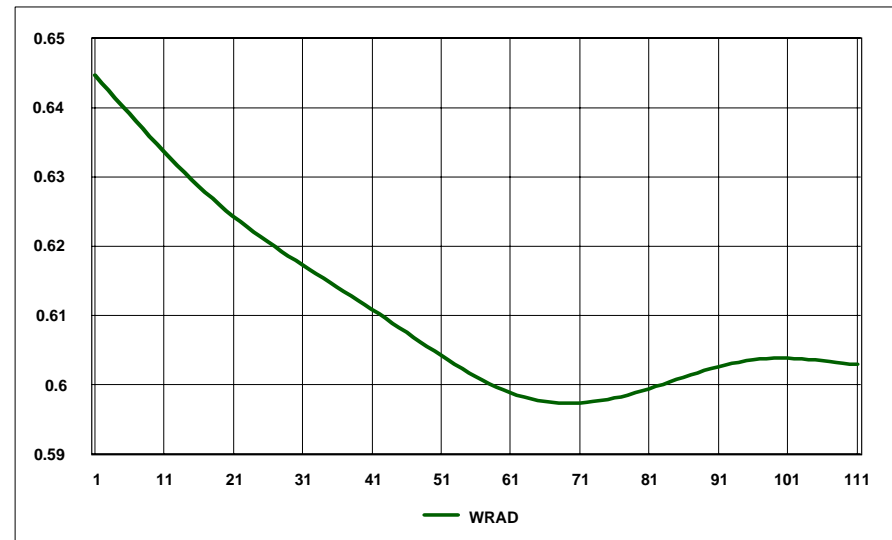
- The next step is to transform each of these lists (one list for each dictionary and each turn of speech), into numbers that give you a continuous curve.

Ozymandias

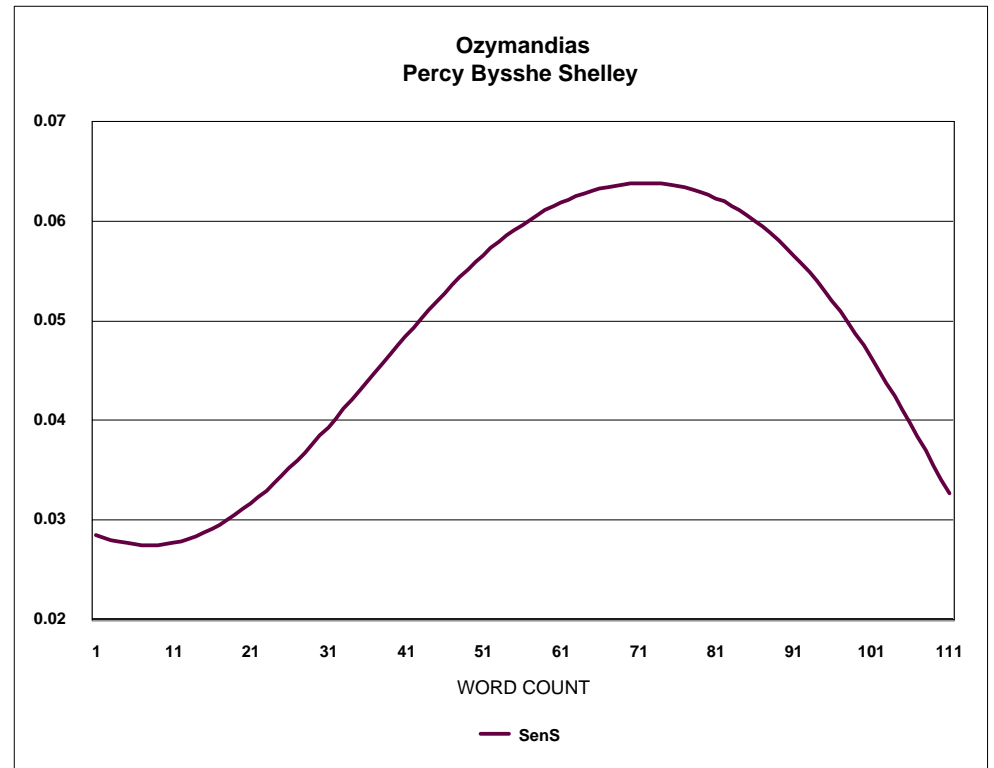
Ozymandias

by Percy Bysshe Shelley

I met a traveler from an antique land
Who said: [10] Two vast and trunkless legs of stone
Stand in the [20] desert. Near them, on the sand,
Half sunk, a shattered [30] visage lies, whose frown,
And wrinkled lip, and sneer of [40] cold command,
Tell that its sculptor well those passions read [50]
Which yet survive, stamped on these lifeless things,
The hand [60] that mocked them and the heart that fed:
And on [70] the pedestal these words appear:
"My name is Ozymandias, king [80] of kings:
Look on my works, ye Mighty, and despair!" [90]
Nothing beside remains. Round the decay
Of that colossal wreck, [100] boundless and bare
The lone and level sands stretch far [110] away.



- Words 15 (*legs*), 60 (*hand*), 66 (*heart*), 68 (*fed*), and 83 (*look*) all lie in the Sensory-Somatic (SenS) dictionary; the others do not.



Means and Weighted Averages

- If you have three numbers x_1 , x_2 and x_3 , you can write the mean:

$$\begin{aligned}\underline{x} &= (x_1 + x_2 + x_3)/3 \\ &= (1/3)x_1 + (1/3)x_2 + (1/3)x_3.\end{aligned}$$

The last expression is written as a weighted average, with weights, $1/3$, $1/3$ and $1/3$.

- The sum of the weights must always be equal to 1, in order not to distort the scale.

Weights

- As long as we keep the sum of the weights equal to 1, we can change them at will. For example, to more strongly weight the middle term, we could change this to:

$$(1/4)x_1 + (1/2)x_2 + (1/4)x_3$$

or

$$(1/8)x_1 + (3/4)x_2 + (1/8)x_3$$

- If we had 5 terms, and we wanted to weight them in a linear fashion similar to the first, we would write:

$$(1/9)x_1 + (2/9)x_2 + (3/9)x_3 + (2/9)x_4 + (1/9)x_5.$$

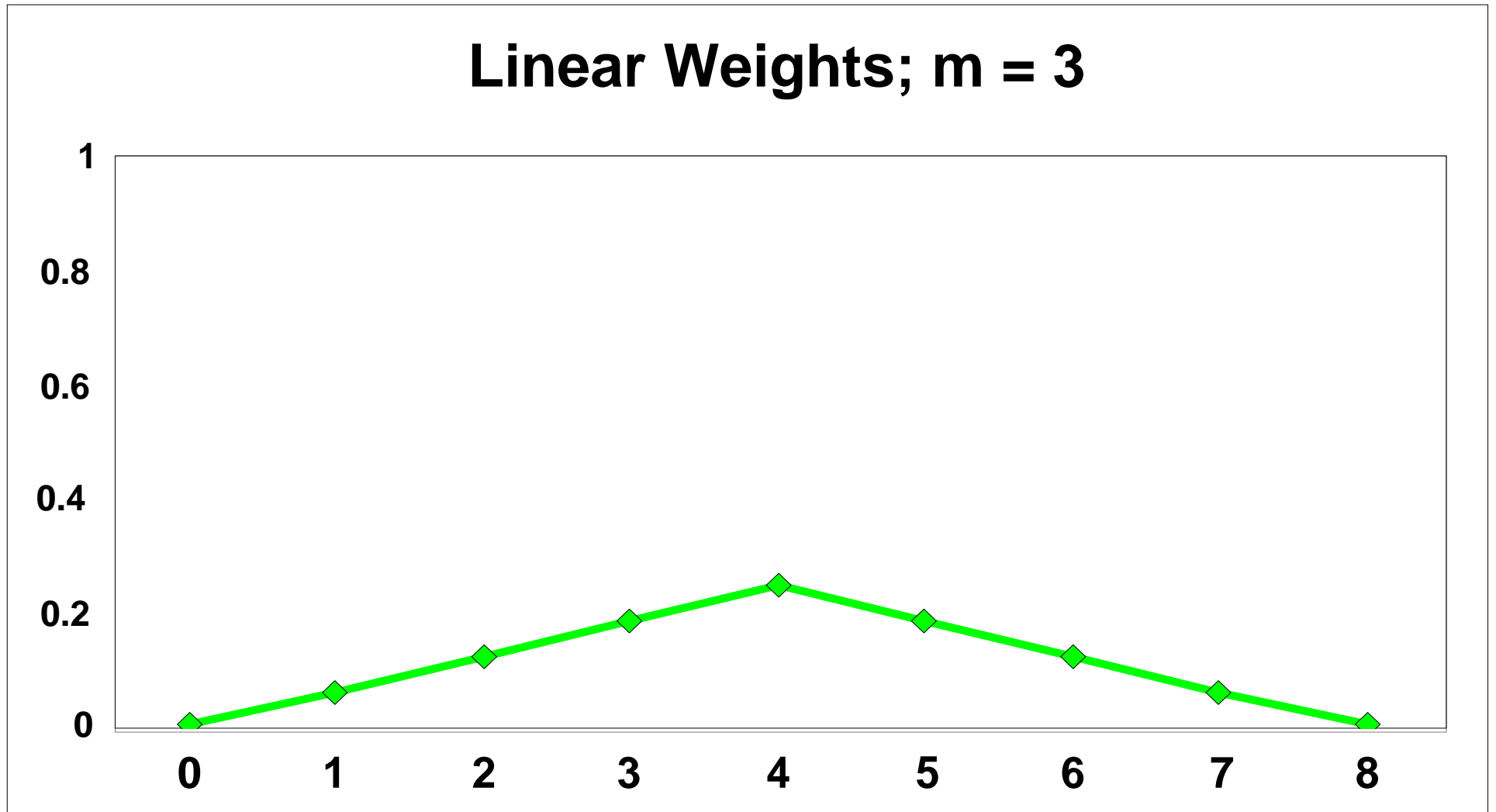
- Similarly, if we had $2m + 1$ numbers (every odd number can be written this way), we could write the linear symmetric weighted average as:

$$(1/M)x_1 + (2/M)x_2 + \dots + (m/M)x_m + \\ ((m+1)/M)x_{m+1} + (m/M)x_{m+2} + \dots + \\ (1/M)x_{2m+1},$$

where M is the sum of the weights:

$$1 + 2 + \dots + m + (m+1) + m + \dots + 2 + 1 \\ = (m+1)^2.$$

Visualization of linear weights



Smoothing

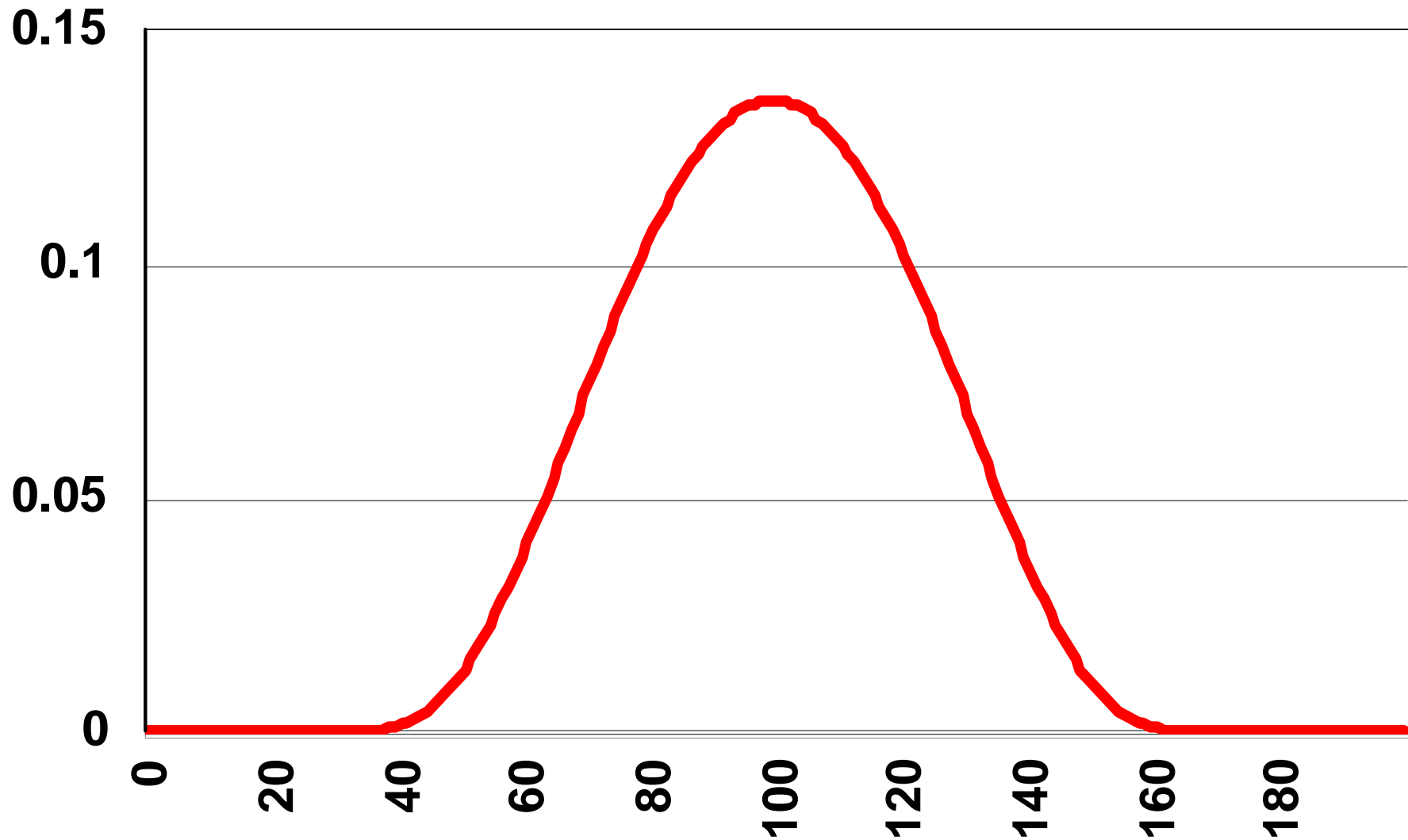
- Notice the point at $x = 4$ in the above chart, where there is an abrupt change in direction.
- This discontinuity of the first derivative can cause problems for us in that, if we use this linear function as our weighting function, then the resulting density curve will not have a continuous second derivative.
- We solve this by changing the weighting function to something without any noticeable breaks or abrupt changes in direction

Smoothing

- The new weighting function looks something like a normal curve; in fact, its formula is closely related to the formula for the normal curve. The major difference is that the normal curve never quite gets to zero, while this one does.
- Our weighting function is 0 at $x = 0$, then minutely positive for x positive, increases to its maximum at $x = 99$, and then decreases back to 0 at $x = 199$.

DAAP Smoothing Function

$m = 100$



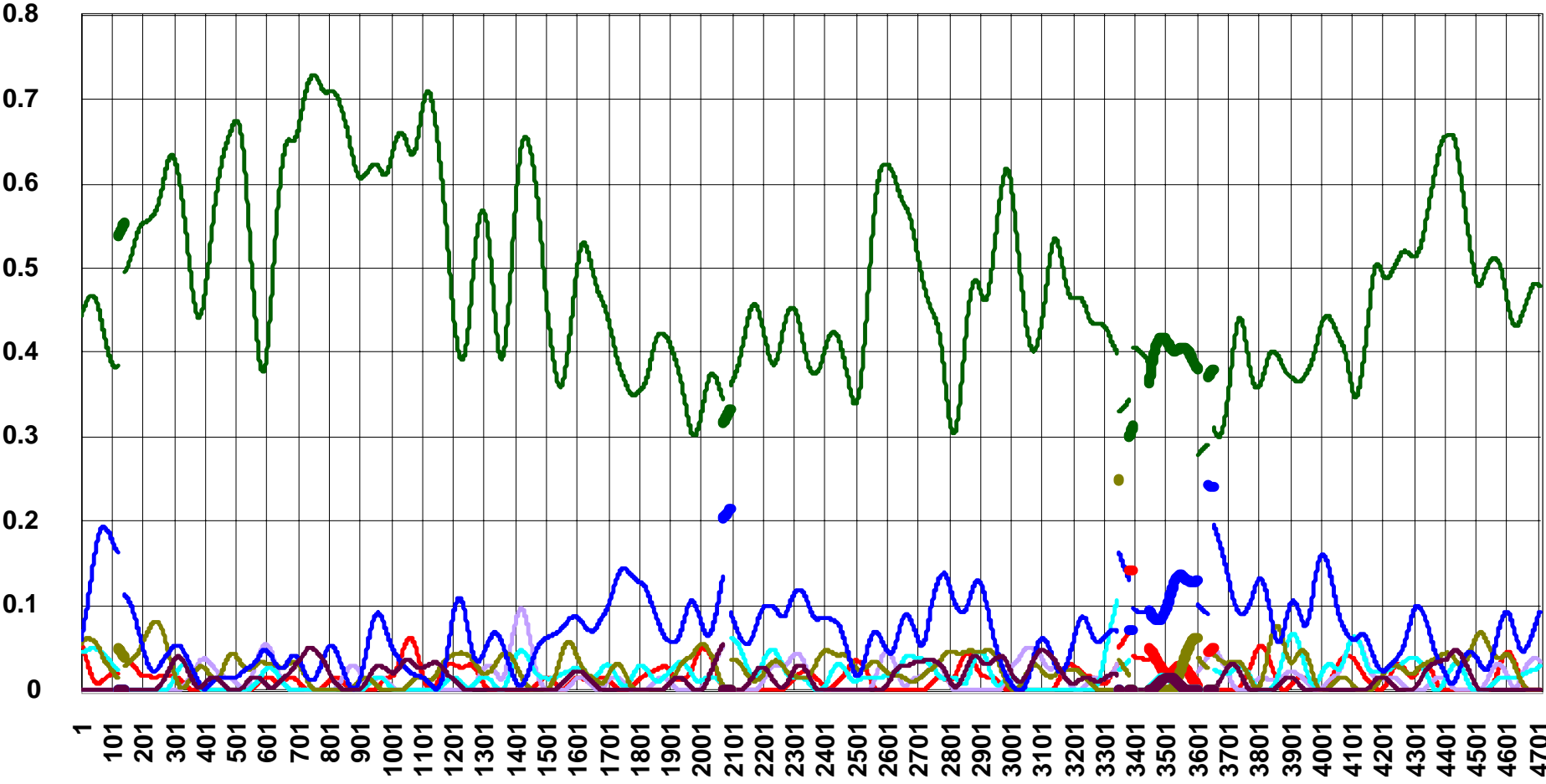
DAAP Smoothing

- DAAP uses the above curve as a weighting function to make a weighted average for each dictionary at each word.
- There are special procedures for obtaining this weighted average near the beginning and end of each turn of speech. The weighting function is appropriately truncated.

Smoothing (Cont.)

- This moving weighted average, computed at each word appears, when put on a graph, as a smooth curve.
- There is in fact a mathematical theory behind this smoothing operation which ensures that the graph is a smooth curve.
- The next two slides illustrate these smooth curves, called *density curves*, for a psychoanalytic session.

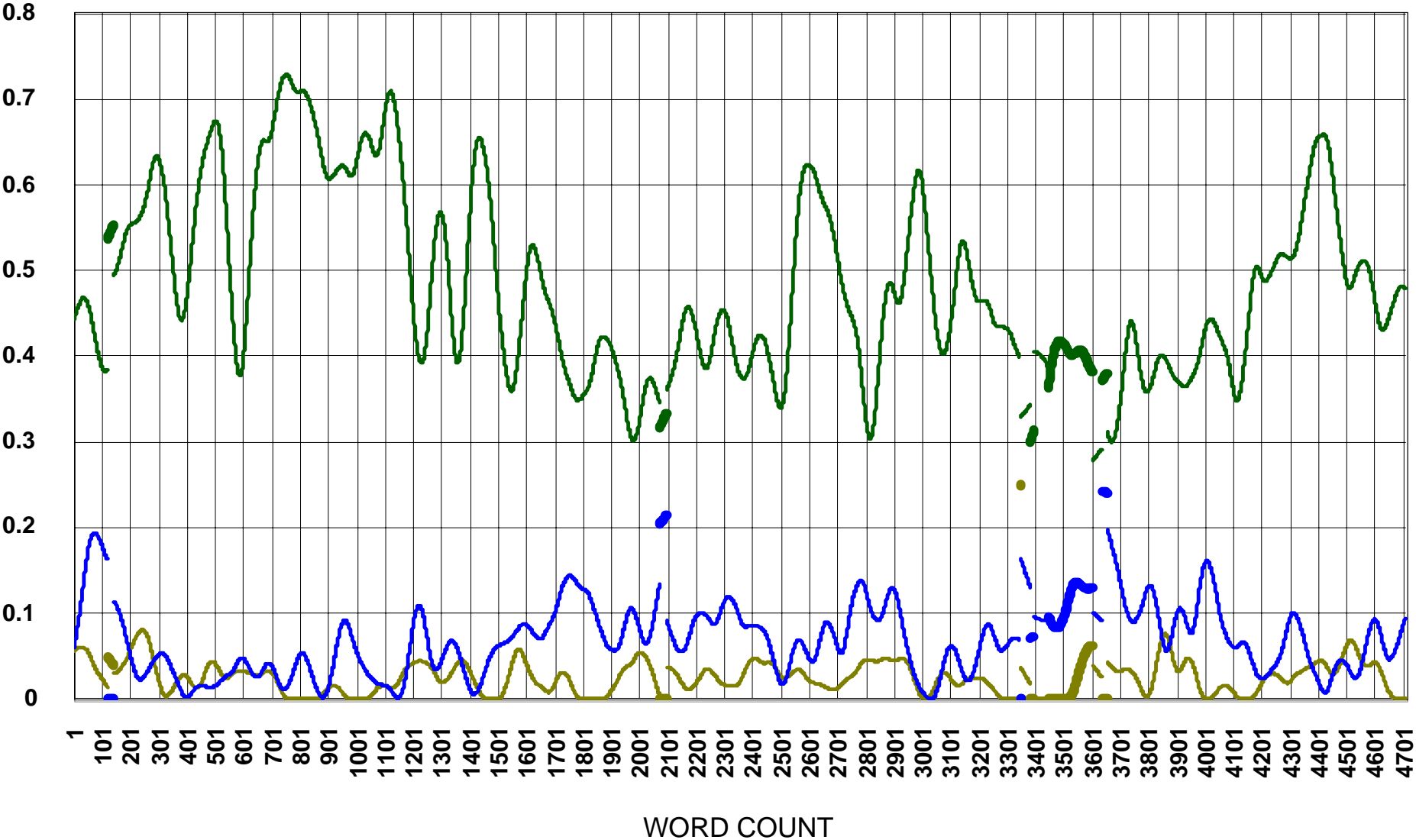
Ms Y Session 257



WORD COUNT

- | | | | |
|---------------------------------------------|---------------------------------------------|---------------------------------------------|---------------------------------------------|
| — PAffP | — PRef | — AAffN | — ARef |
| — PAffN | — PWRAD | — AAffZ | — AWRAD |
| — PAffZ | — PSenS | — ADF | — ASenS |
| — PDF | — AAffP | | |

Ms Y Session 257

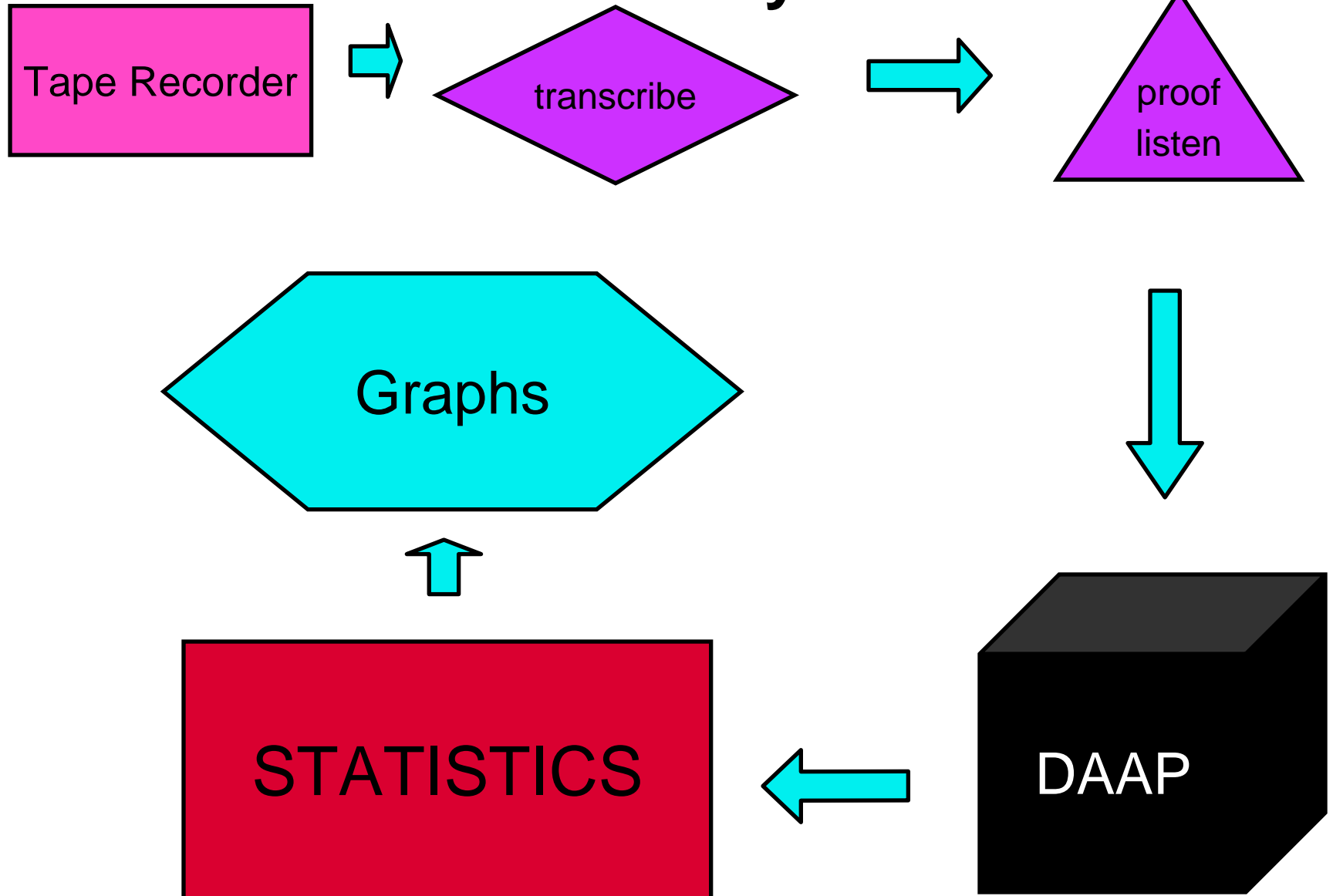


PDF PWRAD ARef AWRAD
PRef ADF

Word Count vs. Time

- Notice that the x-axis is marked “WORD COUNT”. This is to note that the smoothing or averaging is done with regard to words, rather than with regard to time. There is a separate DAAP program that smoothes with respect to time.

Overview of the Basic DAAP System



**ON TO
STATISTICS**

But First, the generations of DAAP

1. BADAAP: Windows platform; written in Visual Basic; one dictionary at a time; up to 10 speakers; no statistics. No longer supported.
2. HDAAP: Console platform; written in C++; call to dictionaries (AffP, AffN, AffZ, DF, Ref, WRAD) built in; arbitrary number of speakers; means, and primitive covariations. No longer supported.

But First, the generations of DAAP

3. HDAAPP: Console platform; written in perl; arbitrary number of speakers; supports categories of text (superdivisions and subdivisions of turns of speech); call to dictionaries (AffP, AffN, AffZ, DF, Ref, WRAD, SenS) built in; means and covariations. No longer distributed, but still supported.

But First, the generations of DAAP

4. DAAP01 and DAAP01.1 (Final version?): Console platform; written in perl; arbitrary number of speakers; arbitrary number of dictionaries; categories and statistics as in HDAAPP. DAAP01.1 also outputs a list of types and tokens by speaker. DAAP01.1 is currently being distributed.

Back to Statistics

- Each of the different versions of DAAP has its own set of formats for the output files. These are described in the “DataFiles.doc” file that comes with the program.

Means

- First basic statistic is the mean of the dictionary values for a given dictionary. For example, the Mean AffP (MAffP), for a given turn of speech, is the number of words in the turn of speech that match the AffP dictionary, divided by the total number of words in the turn of speech.

Means

- The Mean WRAD (MWRAD) is somewhat more complicated. First of all, the WRAD had weights that vary between -1 (low RA speech) and +1 (high RA speech), with 0 as the neutral value. But we want to think of RA as a positive quantity, so we linearly transform the weights to lie between 0 and 1, with .5 as the neutral value.

Means

- The mean WRAD (MWRAD) for the turn of speech is obtained by averaging all the WRAD weights of the words in the turn of speech. Going back to our very short example:

Basic Example

- Text: “Well, I think I’m happily excited, worried, anxious and my back hurts, / (chair creaks) // emotionally.”

	AffP	AffN	AffZ	DF	Ref	WRAD	SenS
Well	0	0	0	1	0	0.375	0
I	0	0	0	0	0	0.125	0
think	0	0	0	0	1	0	0
I	0	0	0	0	0	0.125	0
m	0	0	0	0	0	0	0
happily	1	0	0	0	0	0.5	0
excited	1	0	0	0	0	1	0
worried	0	1	0	0	0	0.5	0
anxious	0	1	0	0	0	0	0
and	0	0	0	0	0	1	0
my	0	0	0	0	0	0.8125	0
back	0	0	0	0	0	1	1
hurts	0	1	0	0	0	0.5	1
emotionally	0	0	1	0	0	0.5	0

We see that, for example,

$$\text{MAffP} = 2/14 = .1429$$

and

$$\text{MWRAD} = 5.3125/14 = .3795$$

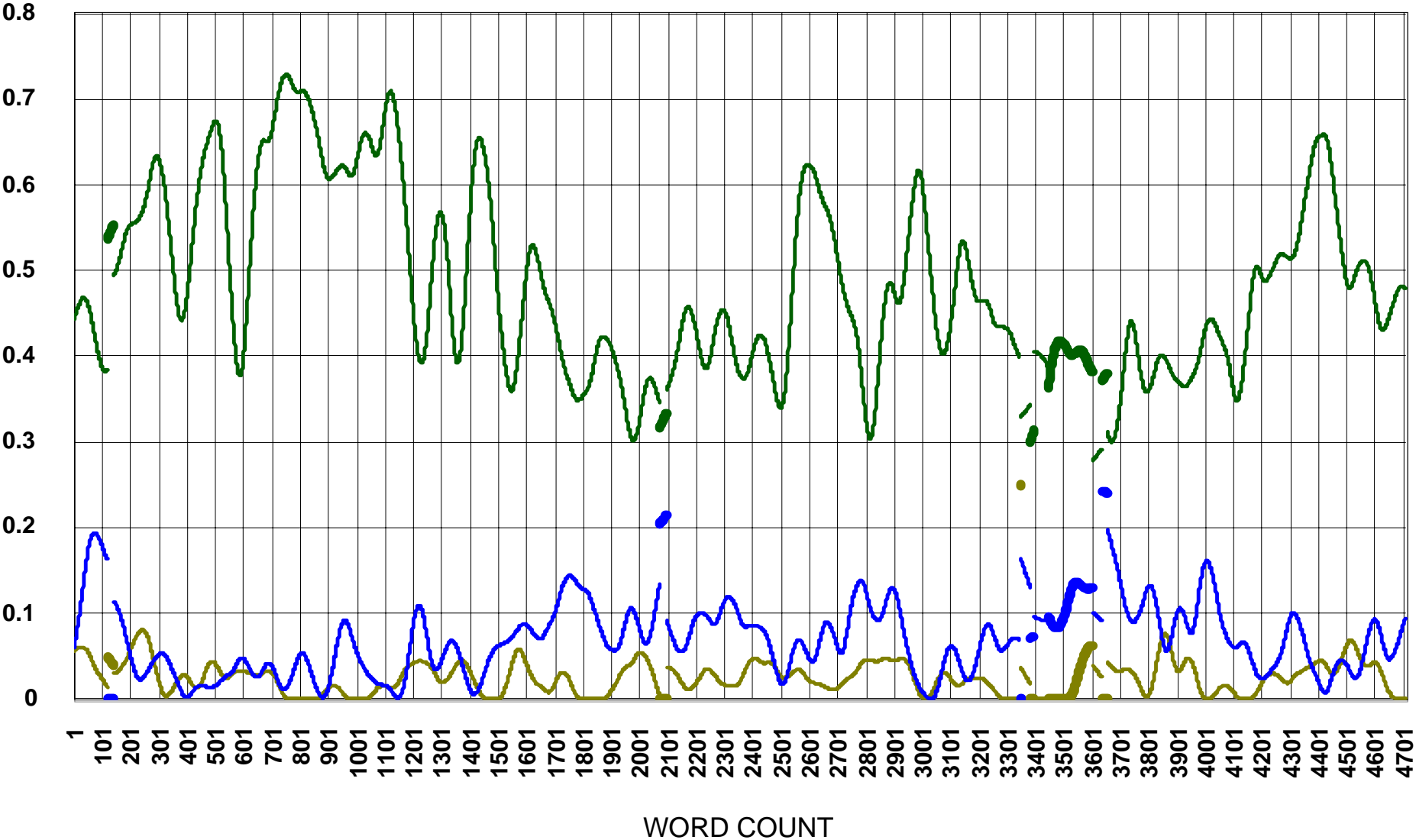
Other Means (to what end?)

- We also compute the mean of the smoothed density for each variable; this mean is usually only minutely different from the mean of the unsmoothed data.
- We also compute the *variablility* of the smoothed dictionary data; this is computationally not different from the standard deviation; however it is statistically different.

Mean High WRAD (MHW)

- MHW is a measure of how high the WRAD is, when it is high. It is obtained by looking only at the words in the turn of speech for which the WRAD density is above its neutral value. We compute the difference between the WRAD density value and .5 (the neutral value); add these all up, and divide by the number of words for which the WRAD density is above .5

Ms Y Session 257



PDF PWRAD ARef AWRAD
PRef ADF

MHW

- The turns of speech with positive MHW are as follows:
- Turn 1 (Analyst): $MHW = .05$
- Turn 4: $MHW = .12$
- Turn 6: $MHW = .07$
- Turn 7: $MHW = .06$
- (Sorry for the confusion: for the computer, the first turn of speech is labeled turn 0.)

MWP

- This measure is the proportion of high WRAD words; that is, it is the total number of words for which the WRAD curve is above its neutral value, divided by the total number of words in the turn of speech. It, along with MHW, are measures of the intensity of the speaker's connection to emotion.

Covariation

- The covariation between two densities gives a measure of the extent to which they are simultaneously above or below their neutral value. For the WRAD, the neutral value is at .5. For the unweighted dictionaries, the neutral value is the mean of the smoothed density values for the entire session (HDAAPP06 and later.)

Covariations

- Going back to Ms Y, session 257, the covariations between Ref and WRAD for the three long turns of speech are:
 - Turn 2: $D_R = .04$; $D_W = -.27$; $R_W = -.66$
 - Turn 4: $D_R = .35$; $D_W = -.30$; $R_W = -.29$
 - Turn 12: $D_R = -.23$; $D_W = .09$; $R_W = -.56$

The Marked Text

- The program reproduces the text with markers placed every 10 words. These markers tell you the word count; the word count for the turn of speech; the number of the turn of speech; and who is speaking.

DAAP Outputs

- Suppose your text file is named OyVey.txt
- Then DAAP will output several data files; these are:
- OyVeyLOG.txt, OyVeyLOL.txt, OyVeySMT.csv, OyVeyTRN.csv, OyVeyAG0.csv, OyVeyAG1.csv, OyVeyAG2.csv, OyVeyMTT.txt, OyVeyTTR.txt, OyVeyGLB.csv.

The Log and LOL Files

- It is important to look in the LOG file. It should be empty; if not, then it is pointing to a problem that probably needs to be fixed.
- The LOL (Left Over List) file is also worth looking at. It is a list of all words in this text that are not in any dictionary, and not in the current Left Over List. It will include numbers and some misspelled words.

The Smooth Data File

- The Smooth Data File (...SMT.csv) contains the numerical data needed to make the charts, such as you just saw.
- The file contains N rows of numbers, where N is the number of (counted) words in the file, and $8M$ columns of numbers, where M is the number of speakers. These columns are, in order: AffP, AffN, AffZ, AffS, DF, Ref, WRAD, SenS.

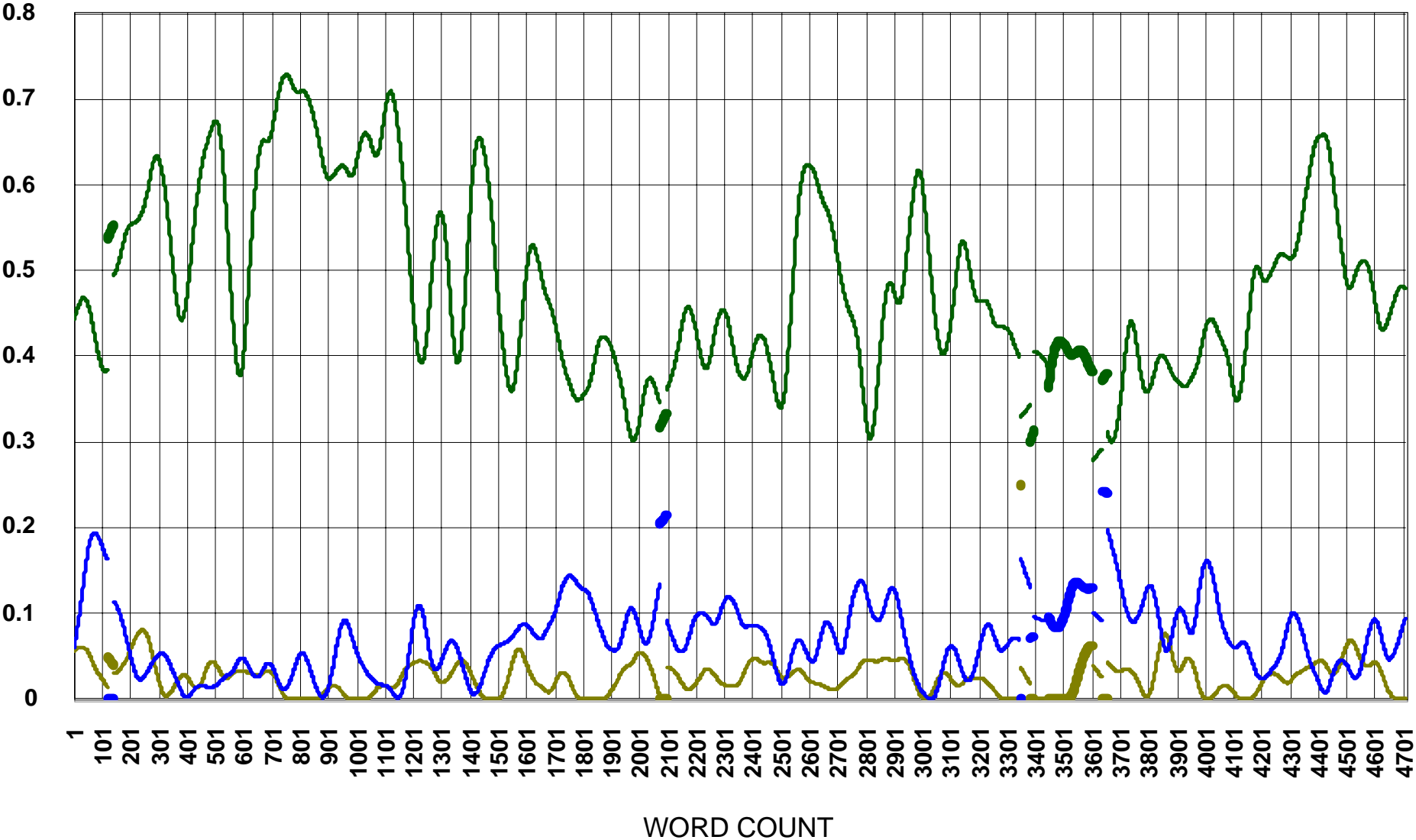
Technicalities

- In older versions of DAAP, labeled as HDAAP..., the SMT File is labeled as the raw file, and has only 7M columns of numbers for each speaker; these are AffP, AffN, AffZ, AffS, DF, Ref and WRAD.

Making Charts

- Both Excel and Lotus 123 recognize .csv files and can make charts using the data in these files.
- **SUGGESTION:** Insert a row at the top of the ...SMT.csv file and write in labels at the top of each column. The spreadsheet will use these to label your chart.

Ms Y Session 257



PDF PWRAD ARef AWRAD
PRef ADF

The Turn Data

- The ...TRN File (...TRN.csv) contains data for each turn of speech. There is one row for each turn of speech. The columns are labeled.
- The data include the number of counted words in the turn of speech, the starting point, the total number of matches for AffP, the mean number of AffP matches, the mean of the smoothed AffP density, the AffP variability, and so on for each dictionary.

The Turn Data

- There are also columns for the number of High WRAD words, the Mean High WRAD, MWP, and the total High WRAD.
- Finally, there are columns for all the covariations: AffP_AffN, AffP_AffZ, etc.

Text Sections

- For most purposes, the natural section of a text is given by turns of speech. However, we sometimes want to form larger or smaller sections; this is done via special markings, called *category markers*, in the text.
- The category markers can delimit a portion of a turn of speech, or encompass several turns of speech.

Example of Category Markings

- Suppose we had a set of interviews of both male and female college students, where each person was interviewed three times, and each interview consisted of a response to a TAT card, an early memory, and a description of a recent event.

- Each response would be marked as to whether this was the first, second or third interview, whether the interviewee was male or female, and which question was being asked. Possible *units of text* for statistical purposes include, for example:
 - A single response to a prompt;
 - All male responses to the TAT on the first interview;
 - All responses by one interviewee to the early memory prompt;
 - All female responses on the first interview.

The AG1 File

- The AG1 File (...AG1.csv) contains one row of basic statistical data (i.e., the same data as is given in the TRN File) for each speaker, and for each instance of each category. The data for distinct turns of speech are aggregated. Since the instances of the categories must be labeled, the number of columns depends on the number of categories. However, the rows and columns are all labeled.

The AG2 File

- ❖ We can also put in special markers to further aggregate data. For example, as above, we could aggregate all the data coming from responses to TAT prompts, or we could aggregate all the data coming from male responses to TAT prompts.
- ❖ However, each such aggregation requires a separate run of the program (a matter of at most a minute or two).
- ❖ The AG2 File (...AG2.csv) has the same columns as the AG1 File, but fewer rows, as some data is aggregated.

The Global Data File

- The Global Data File (...GLB.csv) contains data for the entire file: number of words, number of turns of speech, number of matches for each dictionary and coverage for each dictionary.

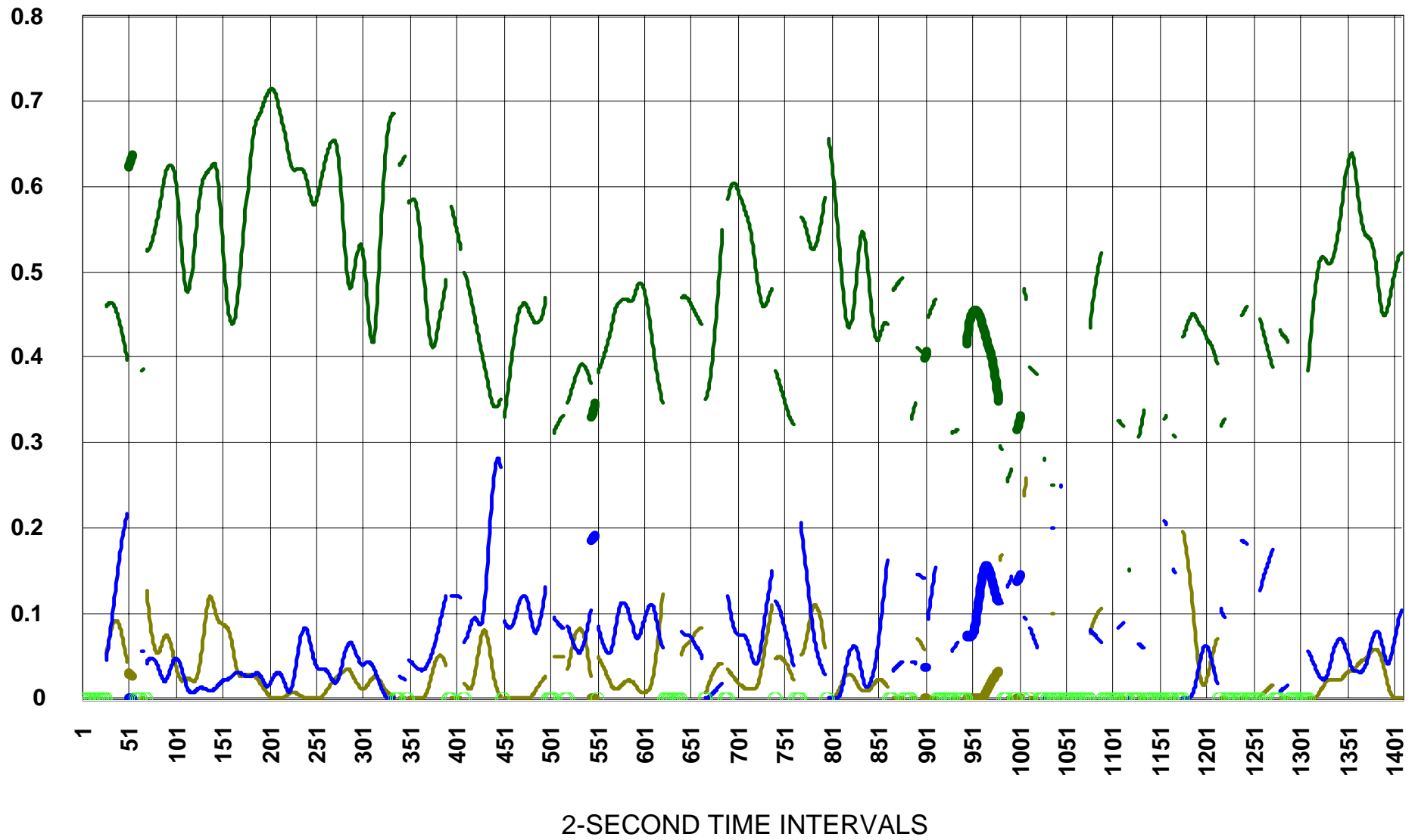
The Type-Token Register

- The TTR file (...TTR.txt) has an alphabetical list of all the types used by each speaker, along with the number of tokens for each type. It also lists the number of NTV's (Non-Turn Vocalizations) for each speaker.

TIME

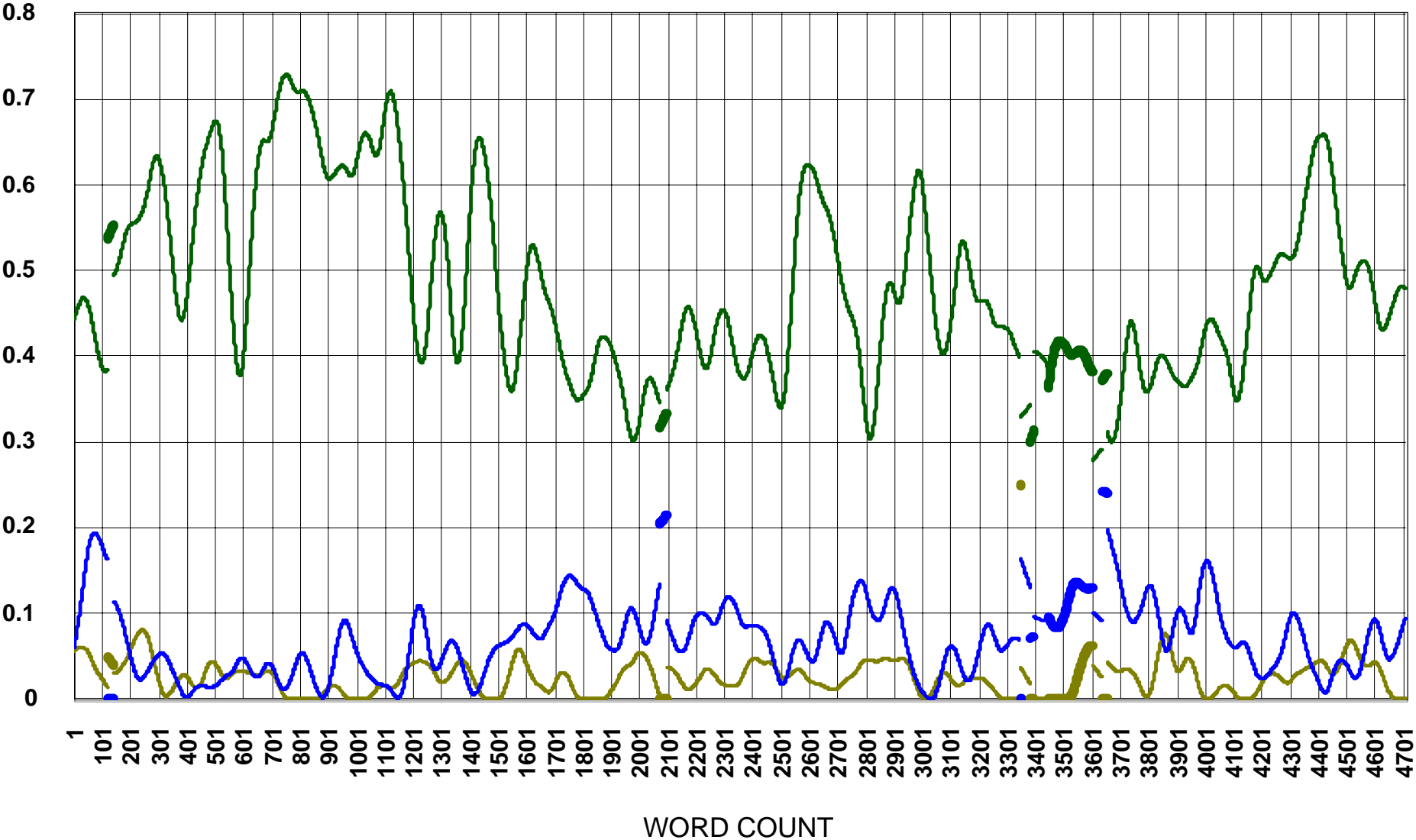
- A new version of DAAP (HDAAPPT) uses time as the independent variable.
- The technology for inserting time markers in the text every 2 seconds was worked out by Ethan Graham, and we now have several such texts.
- This program, which is not yet ready for distribution, produces the same statistics as the preceding ones, but using time as the independent variable.

Ms. Y Session 257



— PDF — PWRAD — ARef ○ Spkr0
— PRef — ADF — AWRAD

Ms Y Session 257



PDF PWRAD ARef AWRAD
PRef ADF