# DAAP Math I: Word Count Base

Bernard Maskit

February 8, 2014

## 1 INTRODUCTION

This note is intended as a description of the DAAP measures where the basic unit is word count. There is a separate technical manual for the DAAP measures where the basic unit is time; in that case, the words are chunked into blocks of variable length by utterance.

Our goal is to have a description that is on a sufficiently technical level so as to include exact statements of the definitions of the measures, and mathematically correct statements as to their properties. We include definitions of the smoothing operator, as well as the measures that are defined in terms of the smoothing operator: the Mean High WRAD, the High WRAD Proportion, and the covariations.

We assume throughout that we are given a text, perhaps a written text, perhaps a transcription of spoken material, perhaps a mixture of both. We assume that this text has been segmented by speaker, if there are distinct speakers, and perhaps by content. There are two prototypical examples of such segmentation:

1. A psychotherapy session that has been segmented by speaker.
2. An interview that has been segmented by question or topic. This segmentation might occur on several levels; i.e., there might be several general topics; there might be several questions within each topic; there might be some overarching concern that includes some questions from different topics.

In any case, as a result of this segmentation, the entire text is divided into a collection of subtexts, which we call *segments*; each segment consists exactly of a set of contiguous words of the text, and the segments are non-overlapping; that is, there is no word (token) that is included in more than one segment.

We also assume that we are interested in one particular dictionary; later, in the section on covariations, we will examine the relationship between two dictionaries.

DAAP reads a text, checks each word against the dictionary and assigns a number, called the *dictionary value*, to each word of the text. If the dictionary is unweighted, then the dictionary value at a word is $+1$ if that word is in the dictionary, and is 0 otherwise.

For computational purposes, weighted dictionaries are defined with weights lying between $-1$ and $+1$, with 0 as the neutral value. Then, as with unweighted dictionaries, words not in the dictionary are assigned the value 0.

Since it is difficult to conceptualize measures such as RA as having a negative value, these dictionary values, lying between $-1$ and $+1$, are linearly transformed so as to lie between 0 and $+1$, with .5 as the neutral value. This linear transformation has no effect on any statistical comparisons, such as the correlation coefficient. From here on, we assume that all dictionary values lie between 0 and $+1$.

# 2 The Smoothing Operator

## 2.1 Extension of the Dictionary Values to a Periodic Function

The smoothing operator, which is defined for each dictionary for an entire text, is defined separately on each segment of text.

We start with a single dictionary and a text containing $N$ contiguous words, labeled as $w_1, \ldots, w_N$. For each $j = 1, \ldots, N$, let $R(j)$ denote the dictionary value at $w_j$.

Conceptually, the smoothing operator consists of two steps: a weighted moving average and a wrap-around. At each word, the weighted moving average uses the dictionary value of the word itself, along with the dictionary values of both the preceding 99 words and the following 99 words. This procedure causes some difficulties at the first and last 99 words of each segment. The wrap-around procedure described below takes care of this difficulty in a particular way.

We first set up the wrap-around by extending the definition of the dictionary values, $R$, to all the integers. We start the definition of the new function $\tilde{R}$ as follows. For $j = 1, \ldots, N$, $\tilde{R}(j) = R(j)$, and $\tilde{R}(N + j) = R(N + 1 - j)$ (that is, in the range, $1 \ldots, 2N$, $\tilde{R}$ is invariant under the reflection $x \mapsto -x + 2N + 1$. We complete the definition of $\tilde{R}$ by requiring that it be periodic with period $2N$.

This function $\tilde{R}$, which is defined for every integer, is called the *periodic extension* of the dictionary value function.

## 2.2 The Moving Weighted Average

We next define the moving weighted average for all integers, using all of $\tilde{R}$. We first define the weighting function, $W$, which depends on the parameter, $m$. For $x \leq -m$, we set $W(x) = 0$; we likewise set $W(x) = 0$ for $x \geq +m$. For $x$ lying between $-m+1$ and $+m-1$, we set

$$W(x) = \frac{e^{-2m^2 \frac{m^2+x^2}{(m^2-x^2)^2}}}{\sum_{j=-m+1}^{m-1} e^{-2m^2 \frac{m^2+j^2}{(m^2-j^2)^2}}}. \tag{1}$$

For the applications discussed here, which involve word count as the independent variable, the parameter $m$ has the value 100.

The smoothed dictionary function $S(\tilde{R})$, defined for every integer $n$, is given by the convolution product,

$$S(\tilde{R})(n) = \sum_{j=-m+1}^{m-1} W(n-j)\tilde{R}(j). \tag{2}$$

The *smoothed dictionary function*, $S(R)$ that we actually use is the restriction of $S(\tilde{R})$ to the values $1, \ldots, N$.

### 2.3 Mathematical Asides

1. The construction of $\tilde{R}$ is easily described in mathematical language. We construct the group generated by the reflections $x \mapsto 1-x$ and $x \mapsto 2N+1-x$. and then extend $R$ to a function that is invariant under this group, call it $\tilde{R}$. We remark that $\tilde{R}(n)$ is well defined for every integer $n$; it is periodic with period $2N$, and it is even; that is, $\tilde{R}(-n) = \tilde{R}(n)$, for every integer $n$.

2. We could perform the entire construction, involving both the extension to $\tilde{R}$ and the construction of the smoothed function, $S(R)$, in terms of functions based on real numbers, not just the integers. We first note that the weighting function as given by Equation **??** is already defined for all real $x$, and that it has continuous derivatives of all orders for all $x$. We extend the function $R$ so that, except for a discrete set of points, it is defined on the entire interval $(1/2, N+1/2)$, as follows. For each $x \in (1/2, N+1/2)$, if there is a unique integer $j$ so that $|x-j| < 1/2$, set $R(x) = R(j)$; if $x$ is a half-integer, leave it undefined. Let $G$ be the group generated by the reflections: $x \mapsto -x+1$ and $x \mapsto 2N+1-x$. Then let $\tilde{R}$ be the function defined for all numbers other than the half-integers, by the property that it is invariant under $G$.

   Then $S(\tilde{R})$ is the convolution product,

   $$\tilde{R} * W(x) = \int_{-\infty}^{\infty} W(t-x)\tilde{R}(t)\,dt.$$

   Finally, $S(R)$ is the restriction of $S(\tilde{R})$ to the domain $[1/2, N+1/2]$.

3. One of the points of the above is that, for a given segment and dictionary, the graph of $S(R)$ is the graph of a mathematically smooth function; that is, the function is continuous, and has continuous derivatives of all orders. The weighting function $W$ has been defined so that it has derivatives of all orders; it follows that the function $S(\tilde{R})$ also has derivatives of all orders.

4. For the first and last 99 words of each segment, one could define the weighted moving average by changing the denominator in Equation **??** so that it is variable, and equal to the sum of the weights actually used. While this approach seems natural, property 2 of Section **??** would not hold, and it is not clear whether or not Property 3 would hold. The wrap-around process was designed to ensure that these properties hold.

5. The definition in terms of continuous variables is necessary for the version of DAAP based on time rather than word count. This is explained in the separate "DAAP Math II: Variable Time Marked Base".

### 2.4 Properties of the Smoothed Function

1. The maximum (minimum) value of $S(R)$ is not greater than the maximum (minimum) value of $R$.

2. The mean of the smoothed dictionary values is equal to the mean of the original dictionary values; that is

$$\frac{1}{N}\sum_{j=1}^{N} S(R)(j) = \frac{1}{N}\sum_{j=1}^{N} R(j). \tag{3}$$

3. The standard deviation of $S(R)$ is not greater than the standard deviation of $R$.

4. For any integer, $i$, in the interval, $[1, N]$, the smoothed dictionary value $S(R(i))$ is independent of the dictionary value $R(j)$, for all integers $j$ for which $|i - j| > 99$.

5. For very short segments, the wrap-around feature dominates the moving average and has the effect that $S(R)$ shows very little variation; it is almost constant. This effect is noticeable for segments having fewer than 25 words, and is generally negligible for segments having at least 100 words.

## 3 Single Variable Derived Measures

The smoothed dictionary function described above is computed separately for each segment or turn of speech. However, we regard it as a function defined on the entire text, so that we can construct the derived measures for each speaker, for the whole text, or for that part of a text satisfying some property, such as all interview responses to a certain question.

### 3.1 Basic Definitions of the Derived Measures

At present, DAAP only uses one weighted dictionary, the WRAD. The definitions below are all given in terms of the WRAD, but could be applied to other weighted dictionaries having a neutral value as well.

From here on we assume that we are concerned with either a single segment of text, or perhaps a disjoint union of related segments of text, such as all the words spoken by one speaker, or perhaps all answers to a related set of questions. These might be the full text, or some subset of it. We label the words in this segment or union of segments as $w_1, \ldots, w_N$. We also assume that we have computed the smooth dictionary values $S(WRAD)$ as above for each element of this disjoint union of text segments.

The WRAD has a neutral value of .5. We need to look at those words for which S(WRAD), the smoothed WRAD function, lies above the neutral value, and those words for which it lies below. We divide the set of integers, $j = 1, \ldots, N$ into the two disjoint subsets, $H$ (for High) and $L$ (for Low), by the property that the number $j$ belongs to $H$ if $S(WRAD)(j) > .5$, and belongs to $L$ otherwise. That is, $H$ is the set of words for which the smoothed WRAD function lies above its neutral value of .5, and $L$ is the complementary set of integers where $S(WRAD) \leq .5$. We let $|H|$ be the cardinality of $H$ (this is the number of words in $H$), and let $|L|$ be the cardinality of $L$, so that $|H| + |L| = N$.

The High WRAD Proportion (HWP) is the proportion of words lying in $H$; i.e.,

$$HWP = \frac{|H|}{N}. \tag{4}$$

The Mean High WRAD (MHW) is 0 if there are no words for which $S(WRAD)$ is greater than .5; that is, if $|H| = 0$. If $|H| > 0$, MHW is the average value of the difference, $S(WRAD)(j) - .5$, for those integers in H. That is,

$$MHW = \frac{1}{|H|} \sum_{j \in H} S(WRAD)(j) - .5. \tag{5}$$

# 4 The Covariation between Two Measures

## 4.1 The Basic Formula

As above, we consider a text consisting of $N$ words labeled $w_j$, $j = 1 \ldots, N$. This text is either a single contiguous segment of text, such as a turn of speech, or is the disjoint union of related such segments of text. Each of the words in this text is evaluated by two distinct dictionaries, which we call $A$ and $B$, with smoothed dictionary functions $S(A)$ and $S(B)$, respectively. The computation of the covariation requires a separate *neutral value* for each of $A$ and $B$; these are described below. Let $M(A)$ be the neutral value for $A$, and let $M(B)$ be the neutral value for $B$.

We will need both smoothed functions to have some variation, so we compute the *skewed variances*, $V(A)$ and $V(B)$. We call $V$ the skewed variance because it would be the variance if the neutral value were the mean.

$$V(A) = \sqrt{\sum_{j=1}^{N} (S(A(j)) - M(A))^2},$$

and

$$V(B) = \sqrt{\sum_{j=1}^{N} (S(B(j)) - M(B))^2}.$$

If either of these is equal to 0, that is, the smoothed dictionary function is constantly equal to its neutral value, then the covariation is equal to 0. If the two skewed variances are both positive, then the covariation, $[A, B]$, between $A$ and $B$ is given by

$$[A, B] = \frac{1}{(V(A))(V(B))} \sum_{j=1}^{N} (S(A(j)) - M(A))(S(B(j)) - (M(B)). \tag{6}$$

We note that $[A, B]$ can be regarded as the cosine of the angle between two vectors; that is, it is the inner product divided by the product of the norms. Hence $|[A, B]| \leq 1$.

### 4.2 The Neutral Value of a Weighted Dictionary

Each weighted dictionary, such as the WRAD, comes with its own neutral value. For the WRAD, this neutral value is .5. We expect that all future weighted dictionaries will have this same neutral value.

### 4.3 Neutral Values for Unweighted Dictionaries for Entire Texts

If $A$ is an unweighted dictionary, and the text we are evaluating is the entire text under consideration, then we take the neutral value to be the mean. For example, if we are looking at all speech by the patient in a psychotherapy session, then the neutral value is the mean of the dictionary function, taken over the entire session. That is, if $A$ is an unweighted dictionary, and our entire text consists of these $N$ words, then

$$M(f) = \frac{1}{N} \sum_{j=1}^{N} S(A(j)). \tag{7}$$

Again, this is also the mean of the original dictionary scores; that is;

$$M(f) = \frac{1}{N} \sum_{j=1}^{N} A(j). \tag{8}$$

### 4.4 Neutral Values for Unweighted Dictionaries in General

Finally, if we are considering a single segment within a larger text, as for example, a single turn of speech in a psychotherapy session, then we take the neutral value to be the mean of the smoothed dictionary function for the same speaker for the entire text. We use the same procedure for a set of segments within a larger text, such as all answers concerning a given topic within a larger interview. That is, the neutral value is the mean of the smoothed dictionary function for the largest appropriate subtext; usually all text by the same speaker.

### 4.5 Connection with the Correlation Coefficient

In the case that, for both $A$ and $B$, the neutral value is the mean of the dictionary values, then, from the point of view of computation, the covariation is indistinguishable from the (Pearson) correlation coefficient. However, from the point of view of statistics, they are quite different. The correlation coefficient requires that the individual items in $A$, and in $B$, be statistically independent, such as test scores on the same test for two distinct groups of people. The values for both $A$ and $B$ at nearby words are not statistically independent, since the smoothing operator takes all nearby words into account.

Even though we cannot treat the smoothed dictionary function at individual words as independent, we can treat the covariations of two dictionaries across a set of distinct texts as independent variables.

## 5 Notes

1. The High WRAD Proportion (HWP), the Mean High WRAD, and the covariations can all be defined in terms of integrals rather than sums, and then one can view the formulae given above as being numerical approximations to the actual value. Even for relatively small texts, these approximations are very close.

2. Time based DAAP has the same derived measures; these are necessarily defined as numerical approximations of integrals; see DAAP Math II: Variable Time Marked Base.

3. Since the smoothed dictionary function tends to be close to constant for short segments (i.e., 25 words or less), the covariations for these short segments tend to be either very close to $-1$ or very close to $+1$, so that very small changes in the dictionary values can cause large changes in the covariations. For this reason, covariations should not be computed for short segments of text or for texts consisting primarily of such short segments.