Building a Weighted Dictionary for Referential Activity

Wilma Bucci

Adelphi University
Derner Institute,
Garden City NY 11530
bucci@panther.adelphi.edu

Bernard Maskit

Stony Brook University
Mathematics Department,
Stony Brook NY 11794-3651
bernie@math.sunysb.edu

Abstract

The Weighted Referential Activity Dictionary (WRAD) is a dictionary (word list) containing 696 items, with weights ranging between -1 and +1, used for computer modeling of a psycholinguistic variable, Referential Activity (RA), in spoken and written language. The RA dimension concerns the degree to which language reflects connection to nonverbal experience, including imagery and bodily and emotional experience, and evokes corresponding experience in the listener or reader. RA is primarily indicated by attributes of style independent of content. High RA language is vivid and evocative; low RA language may be abstract, general, vague or diffuse. RA ratings have been widely used in psycholinguistic and clinical research. RA was initially measured using scales scored by judges; the CRA (Mergenthaler and Bucci, 1999), a binary dictionary, was the first computerized RA measure developed to model judges' RA ratings. The WRAD, a weighted dictionary, shows higher correlations with RA ratings in all text types tested. The development of the WRAD and its applications are made possible by the authors' Discourse Attributes Analysis Program (DAAP), which uses smooth local weighted averaging to capture the ebb and flow of RA and similar variables.

Introduction

Computerized text analysis procedures have largely emphasized aspects of content rather than style. Our new procedure has several innovative features: it assesses a psychological variable, Referential Activity (RA), that is primarily indicated by attributes of linguistic style; both RA as scored by judges, and its computerized measure are independent of specific content; and we use an empirical procedure for producing a dictionary with weighted items.

The RA dimension concerns the degree to which spoken or written language reflects connection to nonverbal experience, including imagery and emotional and bodily experience, and evokes corresponding nonverbal experience in the listener or reader. High RA language is vivid and evocative; see, for example, the following passage:

> In the late summer of that year we lived in a house in a village that looked across the river and the plain to the mountains. In the bed of the river there were pebbles and boulders, dry and white in the sun, and the water was clear and swiftly moving and blue in the channels. Troops went by the house and down the road and the dust they raised powdered the leaves of the trees. The trunks of the trees too were dusty and the leaves fell early that year and we saw the troops marching along the road and the dust rising and leaves, stirred by the breeze, falling and the soldiers marching and afterward the road bare and white except for the leaves.

As illustrated in this opening paragraph from Hemingway's "Farewell to Arms", the effective verbal communication of nonverbal experience often takes place in narrative mode (Bucci, 1995, 1997).

Speech and written language have somewhat different characteristics, which are often not well understood. Here is an example of high RA speech from JSI, one of our data sets; these are described below.

> Other than my finger? Uh, I don't remember how old I was but my grandmother came to live with us. Her husband had died and we had been in a two bedroom apartment and moved to a three bedroom but my sister and I still had to share a room. Grandmother got her own room and just at the time she came to live with us, she started to develop arthritis in her hands. And there was a decanter and glasses set I was very fond of. The decanter was all trimmed in gold and it was a beautiful shape and the glasses were very delicate all trimmed in the same gold. And she picked it up one night. She was having an argument with my parents. She used to fight with my father. This was my mother's mother and between her being upset and the fight, and what they told me was it was her arthritis, but now I wonder if she threw it. She broke this set, and it had always been my favorite. If I were home sick, my mother would fill up the glasses and I would have my juice out of the glasses and on special occasions the decanter would be on the table and I was very angry at her that it was broken and they kept

saying it was her arthritis, her hand had a spasm. And I wasn't allowed to be angry at her about this.

In the other direction, low RA language may be dominated by abstract ideas, as in the opening paragraph of Bertrand Russell's "A History of Western Civilization:

> The conceptions of life and the world which we call "philosophical" are a product of two factors: one, inherited religious and ethical conceptions; the other, the sort of investigation which may be called "scientific," using this word in its broadest sense. Individual philosophers have differed widely in regard to the proportions in which these two factors entered into their systems, but it is the presence of both, in some degree, that characterizes philosophy.

Low RA may also be expressed in spoken language that is vague and diffuse, as in this passage from an interview with a hospitalized depressed patient:

> I love people and I like to be with people. And right now I feel very bad because I can't be with them and do the things I would like to do. But I'm looking forward to a happier and healthier future and - I don't know what else to say. What else can I talk about? Well - I've had a very eventful life, I think. I've worked practically all my life and I love people.

Any type of contents, may be expressed in low or high RA ways; contrast the passage spoken by the psychiatric patient, which refers to emotion, with the manifestly neutral contents of the Hemingway passage, which is nevertheless far more evocative.

RA can be reliably scored by trained judges, using scales derived conceptually from the psychological features of the referential process as defined by Bucci (1997, 2002). The linguistic attributes that are associated with RA and that form the basis for the RA scales are *Specificity* (quantity of detail), *Imagery* (degree to which language evokes imagery), *Clarity* (organization and focus), and *Concreteness* (degree of reference to sensory and other bodily experience). These scales are described more fully in the RA scoring manual (Bucci et al. 1992). Measures of RA have been widely used in clinical and psycholinguistic research. The RA dimension has been shown to be operative in functions such as bodily and emotional self-regulation; interpersonal communication; and the capacity to synthesize cognitive, linguistic and emotional experience. Details concerning the applications of the RA measures and their reliability and validity can be found in Bucci (1997, 2002).

The development of a computerized RA measure has both practical and theoretical significance. On the applied level, while the scales are reliably and easily scored after brief training, computerized procedures are needed to permit assessment of RA in large sample and longitudinal studies, and to permit reliable micro-analytic tracking of fluctuation in RA within various forms of communicative discourse. From a theoretical, psycholinguistic perspective it is of considerable interest to distinguish the types of lexical items that figure in texts that are expressive and evocative, as opposed to texts that are vague, general or abstract.

The first method for computerized scoring of RA was the Computerized Referential Activity (CRA) measure of Mergenthaler and Bucci (1999). This is based on two dictionaries, comprising a total of 181 types. The CRA measure includes a set of items that are characteristic of high RA speech, the High CRA dictionary, and a set of items characteristic of low CRA speech, the Low CRA dictionary. (The definitions and selection procedures for “characteristic” vocabularies are not specifically described in that paper.)

CRA has generally been applied using one of two text analysis systems, the UNIX based TAS/C (Mergenthaler, 1985) and the Windows based CM (Mergenthaler 1998). Neither permit use of weighted dictionaries, and both require segmenting text into arbitrary units of fixed length (number of tokens). A new Windows based computer methodology, the Discourse Attributes Analysis Program (DAAP), which produces a continuous measure without use of segmentation into arbitrary units, has been developed (Maskit, Bucci and Roussos, in preparation). The DAAP produces a mathematically smooth local averaging that starts anew with each change of speaker, and was specifically designed to permit use of weighted as well as unweighted dictionaries. The availability of the DAAP system was a necessary condition for the application of our new Weighted Referential Activity Dictionary (WRAD) that was built to model the RA scores.

To the authors’ knowledge, all dictionaries thus far produced for computerized text analysis are unweighted; that is, an item either is in the dictionary or it is not. Weighted rather than binary dictionaries are particularly important for assessment of stylistic variables, which tend to vary in degree, in contrast to content features, which can usually be defined as present or absent. We anticipate that weighted dictionaries for other psychological and linguistic variables that are also more closely related to style rather than content can also be produced using this technique.

This brief presentation is concerned primarily with the method for producing a weighted dictionary that measures stylistic features of speech rather than content categories. A longer paper, including discussions of linguistic and psychological implications of the results presented here, as well as an explication of several other dictionaries and measures derived from these dictionaries, is in preparation (Bucci and Maskit, In preparation). Our procedure used a principle of modeling RA scale scores as rated by judges, similar to that introduced by Mergenthaler and Bucci (1999); but we used new techniques specifically designed to produce weighted dictionaries.

Methods

The Data Sets. We combined four distinct sets of spoken language texts for the initial construction of the WRAD. This Combined Data Set has a total of 763 texts, comprised

of 5,542 lexical items and 130,138 tokens. The average length was 170 items per text, with a range from 19 to 950 items.

The first data set of the Combined set, JSI, consists of 141 directed monologues, with a total of 2,609 items, and 32,316 tokens. For this data set, the speaker was asked to tell about an early memory, or a memory involving some particular emotion, such as shame or guilt. The data were collected over ten years ago, from 50 middle class women who were students at a graduate teachers training institution in an eastern U.S. city.

The second data set, JST, consists of 201 responses to Thematic Apperception Test (TAT) picture cards, with a total of 2,712 items and 46,711 tokens, from the same sample as in JSI.

The third data set, EKM, has 133 texts, with a total of 2,258 items and 22,053 tokens. This data set consists of segmented interviews, carried out over ten years ago with 41 working class men and women in a middle western U.S. city; only interviewee speech was used for this analysis.

The fourth data set, MSC, has 288 texts, with a total of 1,777 items and 29,058 tokens. This data set was derived from a set of psychoanalytic sessions, all with the same patient and analyst, that had been segmented into idea units, and then scored for RA. Sessions were collected more than twenty years ago, but were scored within the past several years.

The Combined Data set consists of the four previously described data sets, JSI, JST, EKM and MSC.

The fifth data set, TSI, was used for the penultimate step in the construction of the WRAD. This data set consists of 64 texts, with a total of 16,301 tokens. The data set was derived from a set of directed monologues, including early memories, as in JSI, responses to TAT cards, as in JST, and undirected monologues, where the speaker was asked to speak on any topic for five minutes. This sample is less than five years old, and comes from a population of college students.

The sixth data set, MPJ, which was used only to test the final dictionary, and to compare it with CRA, consists of 72 responses, from 36 interviewees, to questions such as in JSI, asking for early memories and recent memories. The interviews were conducted in 2003 with a varied general population. These 72 responses were further divided by judges into 113 idea units, each of which was then scored for RA using the scales. There were a total of 14,495 tokens in this data set.

The basic information for the texts in these data sets is summarized in Table 1.

RA scoring. All the texts were scored for the four RA scales by at least two trained raters, who followed the RA scoring manual (Bucci et al., 1992), and who had achieved reliability of at least .80 measured as Cronbach's alpha, or .75 measured as the single measure intraclass correlation coefficient. Each rater scored each text on a scale of 0 to 10, on each of the four scales. The final RA score for each rater is the average of the four scales, and the final RA score used in our computations is the average of the RA scores of the raters. (The MPJ data set, however, which is used only for testing the final dictionary, has thus far been scored for RA by only one rater, who had achieved excellent reliability in previous work.)

Table 2 shows RA data for the sets of texts used in the construction and testing of the WRAD.

Table 1: Basic Data Set Information

NAME	No. of Texts	No. of Types	No. of Tokens	Mean Text Length	Max. Text Length	Min. Text Length
JSI	141	2,609	32,316	229	950	62
JST	201	2,712	46,711	232	619	63
EKM	133	2,258	22,053	166	620	19
MSC	288	1,777	29,058	101	682	20
Combined	763	5,542	130,138	171	950	19
TS	64	1,582	16,301	254	1,138	43
MPJ	113	1,881	14,495	128	273	36

Table 2: RA Scoring Data

Data Set	RA Mean	RA Maximum	RA Minimum	Standard Dev.	RA/Text Length Correlation
JSI	5.62	10	1	2.217	0.397
JST	4.53	10	.25	2.109	0.298
EKM	4.32	8.75	.625	1.789	0.522
MSC	3.72	7	1.0625	1.167	0.212
Combined	4.39	10	.25	1.898	0.428
TS	4.09	8.25	1.25	1.698	0.329
MPJ	5.37	8.125	2	1.147	0.363

Table 3: The RA Range Categories

RA Range Category	Range of RA scores	Number of Tokens
Very Low	0 - 2.75	22,804
Moderately Low	875 - 3.75	22,354
Mildly Low	3.875 - 4.75	22,242
Mildly High	4.875 - 5.75	21,236
Moderately High	5.875 - 7.125	20,218
Very High	7.25 - 10.00	21,284
Total	0 - 10.00	130,138

Transcription modifications. A set of text preparation transcription rules (Maskit et al., in preparation) has been developed for applications of dictionaries, such as WRAD, within the DAAP system. Special rules for WRAD and related dictionaries include disambiguations of some frequently used words with multiple meanings such as "like", “kind”, “know” and “mean”. Transcribers modify the data sets by introducing new lexical items with distinguishing suffixes. For example, the item **likeC** represents the word "like", when used as a comparative, as in, "This looks likeC a good paper."; **likeV** is the word "like", when used as a verb, as in, "I likeV to go to the movies", and the word **like** itself is reserved for the filler use of the word, as in “So then, like, we like went to the mall”. Specific rules are provided for events and sounds other than words. Transcribers use the item "MM" to represent all meaningless, neutral sounds generally characterized as filled pauses, such as “hm”, “mm”, “um”. Items such as "oh", or "ah", judged by the transcriber to have intended communicative meanings, are transcribed as separate items. In order to maintain consistency, specific rules are also provided for transcription of incomplete or unclear words, contractions and colloquial forms, punctuation and other linguistic and paralinguistic features. The transcriber also uses special items to mark each change of speaker and to indicate who is speaking.

The RA Divisions. RA scores are assigned by judges on a continuous scale from 0 to 10. The final averaged RA score for each text was rounded to the nearest $1/8^{th}$ (.125), yielding 81 score divisions, counting both 0 and 10. The total number of tokens in the Combined data set, in each of these 81 divisions, was counted, and the 81 scores were then divided to form six categories which were as close to equal as possible in number of tokens. Table 3 describes these six *range categories*.

Table 4: The First Twenty Rows of the Main Matrix

Item	Range 1 Very Low	Range 2 Mod. Low	Range 3 Mildly Low	Range 4 Mildly High	Range 5 Mod. High	Range 6 Very High	Num. Of Tokens	Dictionary Weight
	Proportion of total tokens in range							
I	0.0563	0.0569	0.0491	0.0512	0.0451	0.0468	6642	-.75
and	0.0327	0.0382	0.0384	0.0461	0.0494	0.0573	5652	1.0
the	0.0247	0.0307	0.0286	0.0323	0.0343	0.0408	4135	1.0
to	0.0293	0.0293	0.0290	0.0306	0.0256	0.0271	3713	.25
it	0.0247	0.0245	0.0229	0.0206	0.0216	0.0196	2912	-.875
s	0.0247	0.0217	0.0236	0.0202	0.0233	0.0178	2852	-1.0
she	0.0216	0.0168	0.0205	0.0198	0.0220	0.0248	2718	1.0
that	0.0230	0.0229	0.0208	0.0180	0.0170	0.0165	2574	-.875
a	0.0166	0.0146	0.0196	0.0205	0.0227	0.0198	2457	.625
was	0.0097	0.0130	0.0143	0.0182	0.0204	0.0230	2120	1.0
t	0.0185	0.0198	0.0167	0.0170	0.0129	0.0120	2115	-.625
of	0.0167	0.0167	0.0164	0.0153	0.0157	0.0144	2066	-.625
in	0.0102	0.0110	0.0131	0.0130	0.0141	0.0151	1654	1.0
MM	0.0144	0.0161	0.0143	0.0111	0.0101	0.0066	1587	-.625
he	0.0107	0.0098	0.0120	0.0122	0.0151	0.0115	1542	.625
you	0.0109	0.0142	0.0102	0.0095	0.0077	0.0070	1298	-.625
her	0.0096	0.0073	0.0084	0.0100	0.0076	0.0116	1180	1.0
my	0.0063	0.0051	0.0069	0.0087	0.0125	0.0120	1104	.625
but	0.0095	0.0095	0.0098	0.0085	0.0069	0.0063	1102	NA
is	0.0100	0.0079	0.0076	0.0073	0.0064	0.0063	991	-1.0

The main matrix. Using the Combined data set, a first matrix with 5,542 rows and 8 columns was produced. In each row, the first column contains the item, and the next 6 columns contain the total number of tokens for that item in each RA range category. The main matrix (see Table 4) was then derived from the first matrix using the following procedures: First, any item for which the total number of tokens was less than 13 (approximately one in 10,000) was eliminated; this eliminated most specific content words and reduced the number of rows in the main matrix to 737. Second, since the total numbers of tokens in each RA range category were only approximately equal, the entries presented as numbers of tokens in the six range columns were changed to proportions, i.e., the number of tokens for a given item in each range category, divided by the total number of tokens in that range (given in the third column of Table 3.) The eighth column in the main matrix shows the number of tokens in the Combined data set for each item.

For illustrative purposes, the first twenty rows of the main matrix, arranged in order of frequency in the Combined data set, are shown in Table 4. The last column in Table 4 is the dictionary weight for these items in the WRAD; this will be explained below. We note that, of these twenty most frequent items in our sample of spoken language, eight are pronouns; three are prepositions; two are copulative verbs; two are conjunctions; two are articles; and there are the three special items "MM", "s" and "t".

Table 5: First Step in Construction of Weights

C	Weight
1	-1.0
2	-.625
3	-.25
4	.25
5	.625
6	1.0

Constructing a single weighted dictionary depending on a parameter. For each item in the main matrix, one of the six proportions is necessarily larger than the other five; we call the category with the largest proportion the dominant category. We introduce the *weight parameter* P, which provides a measure of ***how much larger*** this largest proportion is than the other five. The assignment of a weight to each lexical item depends on the level of P that is chosen as the criterion of dominance. The following procedures are used to determine this assignment. First, the range category with the maximum proportion is identified for each generalized type. Call this maximum proportion M, and the range category in which it occurs C. We then compute P*M using the selected value of the parameter P. If the entries in the other five columns are all less than P*M, then the item will be included in the dictionary with the weight assigned to that range category according to the scheme given in Table 5; the weights range from -1 to +1, representing the six range categories varying from Very Low to Very High.

Table 6: Second Step in Construction of Weights

Column where M occurs	Column of next highest value	Weight
1	2	-0.875
2	1	-0.750
2	3	-0.500
3	2	-0.375
3	4	-0.125
4	3	0.125
4	5	0.375
5	4	0.500
5	6	0.750
6	5	0.875

If the first step fails for a given entry; that is, there are one or more of the five cells in this row other than C, where the value in that cell is greater than P*M, then we look at all such cells. If one of these cells (i.e., cells whose entry is greater than P*M) is **not** adjacent to C, then this item is viewed as having bimodal features and it is dropped from the dictionary. If there are two such cells, both adjacent to the cell containing M, one on either side, then this item is put in the dictionary, with the same weight as above. If there is only one such cell, and it is adjacent to the cell containing M, then this item is put in the dictionary, with weight as given in table 6.

Selecting the weighting parameter. The fifth data set, TS, which was not used for the selection of the dictionary items, or for computing the weights assigned to them, was then used to find the best value of the weight parameter, P. We can construct a weighted dictionary, WRAD(P), for each parameter value P, using the above procedure. For each WRAD(P), we can obtain a WRAD(P) score for each text of any of our data sets, as the sum of the dictionary weights of each of the items in the text that match an item in the dictionary. To obtain this score, each item in the text is compared with the items in the dictionary; if the item matches an item in the dictionary, then the WRAD(P) score for the text is increased (decreased) by the positive (negative) weight assigned to that item; if the text item does not match an item in the dictionary, the score for the

text is neither increased nor decreased. After all the words in the text have been compared, the total score for the text is the sum of these weights. As previously shown in Table 2, there is a fairly strong correlation between the RA scale scores and text length; this problem may be handled in several different ways. The method selected here was to correct the WRAD(P) scores for text length; that is, we introduce the mean WRAD(P) score (MWRAD(P)); this is the WRAD(P) score divided by the number of words in the segment. We report both the WRAD(P) and MWRAD(P) scores here. A full discussion of the linguistic and psychogical differences between these two measures, one corrected for text length and the other uncorrected, for this as well as for our other dictionaries, will be presented elsewhere (Bucci and Maskit, in preparation).

Using the procedure described above for computing the WRAD(P) and MWRAD(P) scores, we compare the relative validity of the WRAD(P) dictionaries for different values of P as follows. We compute, for each value of P, the (Pearson) correlation between the WRAD(P), or MWRAD(P), scores for the texts in a data set with the corresponding RA scale scores. We computed these correlations for the TS data set, which was not used to construct these dictionaries, using P values from P=.80 to P=.985, in steps of .005. We observed that the correlation increased from P=.80 to P=.975, and then decreased from P=.975 to P=.985; i.e., the maximum correlation with judges' scores occurred at P=.975. Our almost final dictionary was then WRAD(.975), which has 699 items.

Final adjustments and comparisons. The items in this almost final dictionary were reviewed, and the following anomalies were found and corrected. Proper names including names of days of the week were removed. (Interestingly, the days of the week occurred with widely different weights in the WRAD; such anomalies need to be further explored.) Several numbers between one and ten appeared in the dictionary with positive but somewhat different weights. Since the modal weight for these numbers was +1, all integers from one to ten were put in the dictionary with weight +1. After the above modifications were carried out, we obtained the final WRAD with 696 items.

The final step in the process was to test the WRAD on the new data set, MPJ; that is, we obtained WRAD scores, which are uncorrected for text length, and the MWRAD scores, which are corrected for text length, for each text, and then computed the correlation with the RA scale scores.

Results

Description of the WRAD. The final WRAD is a list of 696 item types, of which 674 are ordinary words. Of the other 22, 12 are beginnings of contractions, such as "couldn" or "didn"; 7 are ends of contractions, such as "s" or "t"; two are artificial words used for disambiguation, "knowD" and "likeV" as described above. One item is the neutral sound often written as "mm", "um" or "hm"; all of these are written, following our transcription rules, as "MM".

Correlations with scale scores. Table 7 shows the correlations of both the WRAD scores, and the Mean WRAD scores, with the overall RA scores based on the scales as scored by judges, and the proportion of text covered by the dictionaries, for each of the seven data sets described above. These are compared to the corresponding correlations for the first generation Computerized RA dictionary, the CRA.

As the table indicates, with one exception (the Mean WRAD score for EKM compared with the Mean CRA score), the correlation of WRAD with the RA scales is higher than the RA-CRA correlation for both measures in every data set. This advantage holds for the three data sets (JSI, JST and EKM) that were also used to produce the CRA dictionary, as well as for the new data sets. The TSI data were not used to generate the word list, but only to find the best value of the parameter P, and the MPJ is a new data set used for test purposes only. The data indicate that the WRAD is robust across demographic groups and across a time span of several decades, and is generally robust across text type, with the exception of the WRAD/RA correlation for the MSC data set, the psychoanalytic sessions. We believe this data set was affected by being restricted to one patient-analyst dyad; the assessment of correlations for therapy text material will be expanded substantially in subsequent work.

As table 7 also shows, the coverage of the WRAD list is very high; the 696 types account for a total of between .83 and .87 of all tokens in the 6 data sets. This contrasts with coverage ranging from .50 to .56 for CRA. The greater coverage is made possible primarily by the weighting procedure; without weighting, as in the construction of CRA, only items associated with RA extremes could be included; the weighting procedures permit inclusion of mid-range items with appropriate weights.

Summary data for WRAD. The mean of the Mean WRAD (MWRAD) scores for the combined data set, which was used to build the WRAD, with a total of 763 texts, is -0.03, where the MWRAD scores of course lie between -1 and +1. This would be equivalent to a RA scale score of 5, the mid-point of the RA range. The maximum MWRAD score for the combined data set was .37; and the minimum was -.61; this minimum can be found in the therapy material. We can now place our excerpts in the context of these summary data. As shown in Table 8, the Hemingway excerpt gets an MWRAD score of 0.41, higher than the maximum of our combined set; the segment about the decanter has an MWRAD score of 0.374, which is the maximum of the MWRAD in the combined data set; the Russell excerpt a score of .15; and the psychiatric material a score of -.24.

Table 7: Correlation of RA scale scores with Computerized Procedures.

Data Set	WRAD/RA Correlation	MeanWRAD/RA Correlation	WRAD Coverage	CRA/RA Correlation	MeanCRA/RA Correlation	Coverage
JSI	0.61	0.60	0.85	0.47	.50	0.53
JST	0.58	0.54	0.87	0.37	0.33	0.52
EKM	0.62	0.51	0.83	0.55	0.53	0.51
MSC	0.23	0.49	0.87	0.11	0.39	0.56
Combined	0.60	0.54	0.86	0.40	0.44	0.53
TSI	0.57	0.57	0.86	0.30	0.39	0.51
MPJ	0.47	0.38	0.83	0.31	0.33	0.50

Table 8: The value of small words

Text	Mean WRAD	Mean 18WRAD	WRAD Cover	18WRAD Cover
Hemingway	0.4097	0.355	0.675	0.413
decanter	0.374	0.208	0.842	0.389
Russell	0.154	0.113	0.575	0.26
depressed	-0.281	-0.017	0.888	0.3

Linguistic implications. The nature of the WRAD list is suggested by the items previously shown in the illustrative matrix in Table 4. Of the twenty most frequent items in our sample data set of spoken language, only one, “but” is bimodal, and not in the dictionary, and only one, the item “to”, has relatively low weight. The remaining 18 items account for approximately 35% of all items in our data set. In order to show the importance of these words, we formed a new small dictionary, the 18WRAD, consisting of just these 18 items.. The results of using both the full WRAD, and this 18WRAD on the four vastly different samples of language we used in our introduction appear in Table 8. The first column identifies the sample; the second column is the mean WRAD score for this segment; the third column is the mean 18WRAD score for this segment; the fourth column is coverage of the WRAD, and the fifth column is the coverage of the 18WRAD. We note that the 18WRAD shows the same relationship among the sample segments as the WRAD, but also shows lower variability and lower coverage. While we do not expect this small dictionary to be useful for making fine distinctions, it can be used for widely varying texts. We also note the difference in coverage between spoken and written language, and expect to explore this further.

Other languages. The RA scoring manual (Bucci et al. 1992) has been translated into Italian and Spanish, and substantial corpora of texts have been scored for RA in these languages. There are currently two projects underway, one in each language, involving teams of researchers, including native language speakers, in a project to construct a weighted referential activity dictionary in each of these languages. Preliminary indications suggest that, while spoken Italian and Spanish are linguistically similar, there are strong differences between these languages as spoken, and spoken (American) English. We will report on these similarities and differences in a future publication.

The mystery of simple words. In discussing the Hemingway paragraph quoted above, Joan Didion (1998) notes that it contains 126 words, of which 24 are "the" and 15 are "and"; and says that the arrangement of these simple words "remains as mysterious and thrilling to me now as it did when I first read them, at twelve or thirteen". We hope that our RA and WRAD studies will help to unravel at least some of these linguistic and psychological mysteries. The differential linguistic role of particular lexical items in producing vivid and evocative vs. abstract and general texts, and the psychological significance of these differences will be discussed in detail by Bucci and Maskit (in preparation).

References

Bucci, W. (1995) The power of the narrative; A multiple code account. In Pennebaker, J. (Ed.) *Emotion, Disclosure and Health*, Washington D.C.: American Psychological Association Books, pp. 93-122.

Bucci, W. (1997) *Psychoanalysis and Cognitive Science: A multiple code theory*. N.Y.: Guilford Press**.**

Bucci, W. (2002) Referential Activity (RA): Scales and computer procedures. In Fonagy et al., (Eds.) *An Open Door Review of Outcome Studies in Psychoanalysis;*

Second Edition. International Psychoanalytical Association

Bucci, W. and Kabasakalian, R. and the RA Research Group (1992). *Instructions for scoring Referential Activity (RA) in transcripts of spoken narrative texts*. Ulm, Germany; Ulmer Textbank.

Bucci, W. and Maskit, B. (In preparation) Linking words and things: Measurement and linguistic features of Referential Activity.

Didion, J. (1998) Last Words. *The New Yorker, November 9, pp. 74-80.*

Maskit, B., Bucci, W. and Roussos, A.J. (In preparation) Capturing the flow of verbal interaction – the Discourse Attributes Analysis Program.

Mergenthaler, E. (1985) *Textbank Systems: Computer science applied in the field of psychoanalysis.* Heidelberg & New York: Springer.

Mergenthaler, E. (1998) CM - the Cycles Model software. (Version 1.0) Universitat Ulm, Ulm, Germany.

Mergenthaler, E. and Bucci, W. (1999) Linking verbal and nonverbal representations: Computer analysis of Referential Activity. *British Journal of Medical Psychology, 72,* 339-354.