EQUIWORD: A software application for the automatic creation of truly equivalent word lists

OLAF LAHL and REINHARD PIETROWSKY
University of Düsseldorf, Düsseldorf, Germany

Word lists are most commonly used in the investigation of human memory. To prevent transfer effects, repeated measures of memory for words require multiple lists of different words. Yet, the psycholinguistic properties of all word lists employed should match as closely as possible to avoid confounding with the independent variable(s) in question. Although comprehensive databases for word norms exist, to our knowledge no tool is available that automates the creation of such equivalent word lists. Instead, matching different lists is often accomplished *prima facie*. We have therefore developed a Windows program called EQUIWORD that completely automates the creation of word lists that are truly parallel with respect to a wide range of attributes. EQUIWORD takes psycholinguistic databases of different formats as input and computes several coefficients of distance for every possible word pairing. Program output consists of a list of all word pairs sorted according to their distance. On that basis, creating equivalent word lists is simply done by selecting the pairs with the lowest distance coefficients.

Lists of single words are the most commonly used stimulus material when studying verbal memory. In a typical experimental setting, subjects learn a set of words during one or more study trials either by intention (explicit learning) or incidentally through a dummy task such as rating the words on their pleasantness or counting their letters (implicit learning). At the end of the retention interval (which may last only a few seconds or minutes in the case of immediate testing), memory performance is probed either by explicit recall or recognition or by implicit tests such as word fragment completion.

Subjects' abilities to learn, retain, and retrieve a specific word largely depend on a variety of semantic and lexical features. The most important characteristics that influence these abilities are Paivio's scales of imagery and meaningfulness (Christian, Bickley, Tarka, & Clayton, 1978; Paivio, 1969, 1976); Osgood's scales of evaluation, potency, and activity (Bradley & Baddeley, 1990; Contini & Whissell, 1992; Eysenck, 1976; Howard-Voyer & Whissell, 1994; Lamarche, Campbell, Matheson, & Whissell, 1993; Levonian, 1972; Osgood & Suci, 1955); word length (Lovatt & Avons, 2001); and frequency of use (Gregg, 1976). Even when these attributes are not part of the investigators' research object, there are at least two frequent occasions when the semantic and lexical *equivalence* of different word lists is essential: when using repeated measures or optimizing the sensitivity of measurements.

Because of possible transfer or carryover effects (Hindmarch, 1980; Krauth, 2000), repeated measures of

memory as part of either the experimental settings (see, e.g., Coen, Kinsella, Lambe, Kenny, & Darragh, 1990; Hirshman, Passannante, & Henzler, 1999; Peters & Levin, 1979; Plihal & Born, 1997; Soetens, Casaer, D'Hooge, & Hueting, 1995) or clinical and diagnostic testing (e.g., Crawford, Stewart, & Moore, 1989; Randt, Brown, & Osborne, 1980; Wechsler & Stone, 1973) require two or more lists containing different words. Yet, the words must match as closely as possible with respect to all psycholinguistic attributes that might affect recall—that is, the lists must be parallel. Otherwise, with a large discrepancy in one or more critical variables, any differences between treatments cannot be attributed unambiguously to the experimental manipulation(s).

Even when only one word list is employed for a single measure of recall, a low intralist variance of item difficulty is highly desirable for obtaining sensitivity to the experimental manipulation(s). The sensitivity of a recall measurement might suffer when the individual items of a list differ considerably in one or more characteristics that govern how easily they may be learned or memorized. In that case, easy words would be remembered by almost all subjects, and difficult words only by a few, regardless of the experimental treatments that were presumed to influence recall.

Although numerous pools of word norms for the relevant dimensions exist (G. D. A. Brown, 1984; W. P. Brown & Ure, 1969; Friendly, Franklin, Hoffman, & Rubin, 1982; Gilhooly & Logie, 1980; Kerr & Johnson, 1991; Kučera & Francis, 1967; Paivio, Yuille, & Madigan, 1968; Thorndike & Lorge, 1944; Toggia & Battig, 1978), manually parallelizing words becomes a daunting and virtually infeasible task as soon as more than two variables are to be included. As Cutler (1981) points out, "At this point it is already clear to the experimenter that the task is probably impossible" (p. 68).

Correspondence concerning this article or requests to obtain the software should be addressed to O. Lahl, Heinrich-Heine-Universität, Klinische Psychologie, Universitätsstrasse 1 (Geb. 23.03), 40225 Düsseldorf, Germany (e-mail: olaf.lahl@uni-duesseldorf.de).

Although this statement holds for handmade lists, the mathematical formulation of the problem is straightforward. A single word, measured for n different attributes, can be considered as a vector in an n -dimensional space. Hence, the overall dissimilarity between two words is the mathematical distance $d(\mathbf{x}_k, \mathbf{x}_j)$ of the two vectors $\mathbf{x}_k$ and $\mathbf{x}_j$ that represent them. Several coefficients for the distance of two vectors are commonly used in psychological applications, and we will give an overview of them in the Appendix. For a list of m words, there are $(m^2 - m)/2$ pairs of different words. Once the pairwise distances are computed, the task of creating equivalent word lists is simply accomplished by picking up those pairs with the lowest (or the highest) coefficients of dissimilarity (or resemblance). Selection criteria are not restricted, though, to a preference for high proximities. It is also possible to create lists of words that have maximum distance, or any other degree of proximity that is of interest to the researcher.

We have created a software program called EQUIWORD that accepts databases of word norms as input, computes the distance for each word pair, and saves the sorted list of pairs to a file. EQUIWORD implements a set of five common distance measures and two methods of data standardization that are discussed in the Appendix. The program is capable of reading three different types of input files: The first file type adheres to the format used by the Medical Research Council (MRC) Psycholinguistic Database (Coltheart, 1981; Wilson, 1988). The MRC dictionary file is available on the Internet (ota.ahds.ac.uk/texts/1054.html) and contains a set of 150,837 English words measured for 26 psycholinguistic attributes. However, the entire set of attributes is not available for every word, and EQUIWORD is restricted to the 14 attributes that are of a quantitative nature. These attributes are number of letters, number of phonemes, number of syllables, Kučera and Francis (1967) written frequency, Kučera and Francis number of categories, Kučera and Francis number of samples, Thorndike–Lorge (1944) written frequency, G. D. A. Brown (1984) verbal frequency, familiarity, concreteness, imagery, mean Colorado meaningfulness (Toglia & Battig, 1978), mean Paivio meaningfulness (unpublished, but based on Paivio et al., 1968), and age of acquisition. For a detailed description of these psycholinguistic properties, see Wilson (1988). EQUIWORD provides the possibility to include either the entire set or any possible subset of these attributes in the calculation of distance coefficients.

In addition to MRC database files, EQUIWORD accepts Microsoft Excel spreadsheet files and comma separated value files. With these file types, the user can specify his or her own set of words and word attributes. As an example, Osgood's scales are not part of the MRC database, but they are important for learning and remembering, as outlined above. Provided that the user has access to ratings on these scales for a sample of words, he can easily create a database containing these ratings and use it with

EQUIWORD. The use of user-defined database files is described in detail in the following section.

Program Operation

After starting the program, MRC database files (file extension .dct), Microsoft Excel files¹ (.xls), or comma separated value files (.csv) can be loaded by selecting the "File Open" command. When working with an MRC database file, a subset of words can be selected by choosing one of four grammatical types—nouns, adjectives, verbs, and adverbs—and any combination of the 14 quantitative attributes the MRC database provides. Note that for the computation of distance coefficients, EQUIWORD includes only the intersection of words with valid entries in all selected attributes.

Excel and comma separated value files offer the opportunity to load arbitrary collections of words (or even other objects) together with a set of ratings/attributes provided by the user. The first line of the file is the header and must specify the attributes' names. All following lines begin with a single word and contain its attribute values in the following columns. In the case of comma separated value files, each attribute name in the header and each attribute value must be separated by a semicolon. See Listing 1 for two illustrative examples of how to format the data properly. Leaving an empty cell or field between two semicolons indicates a missing value for that attribute. As with MRC database files, words with missing values in at least one attribute will be discarded in the computation of distance.

The Options menu offers choices for the type of data transformation (z standardization, range standardization, or none), for the sort order of program output (sort ascending or descending by distance coefficient), as well as for the font type used for display. Finally, selecting one of the available distance coefficients from the Compute menu starts its computation. After finishing the calculation, the tabulated file output appears in the program window. Depending on the user's selection, all word pairs are listed in ascending or descending order on the basis of the magnitude of their distance. Navigation to a specific word pair is possible using the Find Pair dialog box that can be opened from the Edit menu. Selecting the "Save" command from the File menu saves the output to a text file for further processing by spreadsheet programs or text editors.

EQUIWORD also supports a second mode of operation that enables the user to compute the available distance coefficients all at once. To adopt simultaneous computation, one can select the option "All Coefficients at Once" from the View menu. Pressing the F5 key or selecting "All" from the Compute menu initiates calculation. The program output appears in a spreadsheet with one column for each coefficient. Clicking a column header initiates sorting the entire output by the corresponding coefficient, in ascending order. Clicking the same column again toggles the sort order.

A

Noun	Frequency	Imagery	Concreteness	Meaningfulness	Number of letters
Umbrella	8	592	606	676	8
Reaction	124	395	312	484	8
...					

B

Noun;Frequency;Imagery;Concreteness;Meaningfulness;Number of letters

Umbrella;8;592;606;676;8

Reaction;124;395;312;484;8

...

Listing 1. Illustrative examples of how to format user-specific data files in (A) Microsoft Excel format and (B) comma-separated value format.

Discussion

With the software program presented in this article, we hope to facilitate the creation of word lists that are truly equivalent in the mathematical sense of the word, thereby reducing potential sources of confounding in experiments that rely on psycholinguistic material. The value of our software for researchers is rising as large word pools like the MRC database increasingly become available in electronic form and as an increasing number of psycholinguistic properties are included in the ratings. As an example, inclusion of ratings for Osgood's scales of evaluation, potency, and activity (Osgood & Suci, 1955) is highly desirable. Ironically, these ratings are at least partially available for a pool of 650 words (W. P. Brown & Ure, 1969), but not in electronic form.

We have already developed a test version of a software module for merging databases containing different word pools and ratings, and we look forward to implementing it in future versions of our program, as soon as more word pools become available in computer-readable formats.

System Requirements and Program Availability

EQUIWORD is a Windows program that requires Microsoft .NET Framework 2.0 or higher, which can easily be downloaded from msdn.microsoft.com/netframework/downloads/updates. The EQUIWORD setup program checks the target machine for the presence of the .NET Framework and installs it if necessary. EQUIWORD itself requires about 4 MB of hard drive space. The program has been tested on a Pentium 4 processor with 1.7 GHz and 1 GB of RAM, and we recommend using a similar amount of memory for handling large word sets. Memory consumption primarily varies with the number of words to be included in the computation. Therefore, if memory runs low when working with MRC database files, it is a good

idea to restrict the number of words by selecting more attributes. Since only the intersection of words with valid entries in all selected attributes is included, this will lower the number of words and thereby the required amount of memory.²

EQUIWORD is free of charge for noncommercial use and can be downloaded via FTP from <ftp://ftp.sleep.lab.psych.uni-duesseldorf.de/EquiWord> or <ftp://ftp.uni-duesseldorf.de/pub/psycho/lahl/EquiWord>. Users planning to use EQUIWORD for commercial purposes (e.g., research in the psychopharmacological industry) are required to contact the first author. They will receive a CD-ROM containing the software for a charge of US\$500. Authors who use EQUIWORD for their publication are expected to cite this reference and the program version used.

REFERENCES

- BRADLEY, B. P., & BADDELEY, A. D. (1990). Emotional factors in forgetting. *Psychological Medicine*, **20**, 351-355.
- BROWN, G. D. A. (1984). A frequency count of 190,000 words in the *London-Lund Corpus of English Conversation*. *Behavior Research Methods, Instruments, & Computers*, **16**, 502-532.
- BROWN, W. P., & URE, D. M. J. (1969). Five rated characteristics of 650 word association stimuli. *British Journal of Psychology*, **60**, 233-249.
- CHRISTIAN, J., BICKLEY, W., TARKA, M., & CLAYTON, K. (1978). Measures of free recall of 900 English nouns: Correlations with imagery, concreteness, meaningfulness, and frequency. *Memory & Cognition*, **6**, 379-390.
- COEN, R. F., KINSELLA, A., LAMBE, R., KENNY, M., & DARRAGH, A. (1990). Creating equivalent word lists for the Buschke Selective Reminding Test. *Human Psychopharmacology*, **5**, 47-51.
- COLTHEART, M. (1981). The MRC Psycholinguistic Database. *Quarterly Journal of Experimental Psychology*, **33A**, 497-505.
- CONTINI, L., & WHISELL, C. (1992). Memory disadvantages for CVC associates of emotional words. *Perceptual & Motor Skills*, **75**, 427-431.

- CRAWFORD, J. R., STEWART, L. E., & MOORE, J. W. (1989). Demonstration of savings on the AVLT and development of a parallel form. *Journal of Clinical & Experimental Neuropsychology*, **11**, 975-981.
- CUTLER, A. (1981). Making up materials is a confounded nuisance: or, Will we be able to run any psycholinguistic experiments at all in 1990? *Cognition*, **10**, 65-70.
- EYSENCK, M. W. (1976). Arousal, learning, and memory. *Psychological Bulletin*, **83**, 389-404.
- FRIENDLY, M., FRANKLIN, P. E., HOFFMAN, D., & RUBIN, D. C. (1982). The Toronto Word Pool: Norms for imagery, concreteness, orthographic variables, and grammatical usage for 1,080 words. *Behavior Research Methods & Instrumentation*, **14**, 375-399.
- GILHOOLY, K. J., & LOGIE, R. H. (1980). Age-of-acquisition, imagery, concreteness, familiarity, and ambiguity measures for 1,944 words. *Behavior Research Methods & Instrumentation*, **12**, 395-427.
- GREGG, V. (1976). Word frequency, recognition and recall. In J. Brown (Ed.), *Recall and recognition* (pp. 183-216). London: Wiley.
- HINDMARCH, I. (1980). Psychomotor function and psychoactive drugs. *British Journal of Clinical Pharmacology*, **10**, 189-209.
- HIRSHMAN, E., PASSANNANTE, A., & HENZLER, A. (1999). The effect of midazolam on implicit memory tests. *Brain & Cognition*, **41**, 351-364.
- HOWARD-VOYER, C., & WHISELL, C. (1994). Emotional descriptors of words as predictors of recall in a paired-associate task. *Perceptual & Motor Skills*, **78**, 1187-1191.
- KERR, N. H., & JOHNSON, T. H. (1991). Word norms for blind and sighted subjects: Familiarity, concreteness, meaningfulness, imageability, imagery modality, and word associations. *Behavior Research Methods, Instruments, & Computers*, **23**, 461-485.
- KRAUTH, J. (2000). *Experimental design: A handbook and dictionary for medical and behavioral research*. Amsterdam: Elsevier.
- KUČERA, H., & FRANCIS, W. N. (1967). *Computational analysis of present-day American English*. Providence, RI: Brown University Press.
- LAMARCHE, S., CAMPBELL, M., MATHESON, P., & WHISELL, C. (1993). A memory advantage for serial lists composed of active words. *Perceptual & Motor Skills*, **77**, 748-750.
- LEVONIAN, E. (1972). Retention over time in relation to arousal during learning: An explanation of discrepant results. *Acta Psychologica*, **36**, 290-321.
- LOVATT, P., & AVONS, S. (2001). Re-evaluating the word length effect. In J. Andrade (Ed.), *Working memory in perspective* (pp. 199-218). Hove, U.K.: Psychology Press.
- MAHALANOBIS, P. C. (1936). On the generalized distance in statistics. *Proceedings of the National Institute of Sciences of India*, **2**, 49-55.
- OSGOOD, C. E., & SUCI, G. J. (1955). Factor analysis of meaning. *Journal of Experimental Psychology*, **50**, 325-338.
- PAIVIO, A. (1969). Mental imagery in associative learning and memory. *Journal of Experimental Psychology*, **76**, 241-263.
- PAIVIO, A. (1976). Imagery in recall and recognition. In J. Brown (Ed.), *Recall and recognition* (pp. 103-129). London: Wiley.
- PAIVIO, A., YUILLE, J. C., & MADIGAN, S. A. (1968). Concreteness, imagery, and meaningfulness values for 925 nouns. *Journal of Experimental Psychology Monographs*, **76**(1, Pt. 2).
- PENROSE, L. S. (1953). Distance, size and shape. *Annals of Eugenics*, **18**, 337-343.
- PETERS, B. H., & LEVIN, H. S. (1979). Effects of physostigmine and lecithin on memory in Alzheimer disease. *Annals of Neurology*, **6**, 219-221.
- PLIHAL, W., & BORN, J. (1997). Effects of early and late nocturnal sleep on declarative and procedural memory. *Journal of Cognitive Neuroscience*, **9**, 534-547.
- RANDT, C. T., BROWN, E. R., & OSBORNE, D. P. (1980). A memory test for longitudinal measurement of mild to moderate deficits. *Clinical Neuropsychology*, **2**, 184-194.
- ROMESBURG, C. H. (1984). *Cluster analysis for researchers*. Belmont, CA: Lifetime Learning.
- SOETENS, E., CASAER, S., D'HOOGHE, R., & HUETING, J. E. (1995). Effect of amphetamine on long-term retention of verbal material. *Psychopharmacology*, **119**, 155-162.
- THORNDIKE, E. L., & LORGE, I. (1944). *The teacher's word book of 30,000 words*. New York: Teachers College Press, Columbia University.
- TOGLIA, M. P., & BATTIG, W. F. (1978). *Handbook of semantic word norms*. Hillsdale, NJ: Erlbaum.
- WECHSLER, D., & STONE, C. P. (1973). *Wechsler Memory Scale*. New York: Psychological Corporation.
- WILSON, M. (1988). MRC Psycholinguistic Database: Machine-usable dictionary, version 2.00. *Behavior Research Methods, Instruments, & Computers*, **20**, 6-10.

NOTES

1. Loading Excel files requires that the Microsoft Excel application be installed on the target machine. EQUIWORD checks for this during program startup and offers the option of opening Excel files in the File Open dialog box only if it finds an intact installation of Microsoft Excel.
2. On the other hand, 1 GB of RAM or more may be required if the number of possible word pairs is very large. This might happen on some rare occasions, such as when the entire MRC database (mrc2.dct) is used and the set of variables to be included for computation is available for a great number of words. An extreme example would be the matching of all nouns for their word length as the sole variable. The word length is of course available for the entire set of 22,061 nouns, resulting in $(22,061^2 - 22,061)/2 = 243,332,830$ word pairs. Since EQUIWORD requires 12 bytes for the internal representation of each pair, this would result in a memory consumption of about 2.7 GB.

APPENDIX

There are many ways to quantify the distance between objects, and the decision of which one to use depends mainly on the specific object of the research. EQUIWORD implements a set of five common measures that we will discuss briefly. For the following formulas, we assume a data matrix $\mathbf{X}$ with words (objects) arranged in rows $i = 1, \dots, m$ and attributes (object properties) arranged in columns $j = 1, \dots, n$:

$$\mathbf{X} = \begin{pmatrix} x_{11} & x_{12} & \dots & x_{1j} & \dots & x_{1n} \\ x_{21} & x_{22} & \dots & x_{2j} & \dots & x_{2n} \\ \vdots & \vdots & & \vdots & \vdots & \vdots \\ x_{i1} & x_{i2} & \dots & x_{ij} & \dots & x_{in} \\ \vdots & \vdots & & \vdots & \vdots & \vdots \\ x_{m1} & x_{m2} & \dots & x_{mj} & \dots & x_{mn} \end{pmatrix}. \quad (1)$$

Minkowski Metric

The general Minkowski or L_r metric $d_r(\mathbf{x}_k, \mathbf{x}_l)$ is defined as

$$d_r(\mathbf{x}_k, \mathbf{x}_l) \equiv \sqrt[r]{\sum_{j=1}^n |x_{kj} - x_{lj}|^r}, \quad (2)$$

where the Minkowski constant r is a positive integer. The special case of $r = 1$ is called the *city block metric*:

$$d_1(\mathbf{x}_k, \mathbf{x}_l) = \sum_{j=1}^n |x_{kj} - x_{lj}|. \quad (3)$$

The term *city block metric* refers to the fact that this is the distance a pedestrian would have to cover on a grid of city streets that is confined to horizontal and vertical directions only. An airplane, on the other hand, would take the diagonal direction or airline distance, which is the familiar Euclidean distance with $r = 2$. The Euclidean distance has a maximum size of the city block distance (triangle inequality).

$$d_2(\mathbf{x}_k, \mathbf{x}_l) = \sqrt{\sum_{j=1}^n (x_{kj} - x_{lj})^2}. \quad (4)$$

The city block metric applies the same weight to all differences. With increasing r , large differences contribute more to the magnitude of d_r , which also means that the measure becomes more sensitive to outliers. In the case of $r \rightarrow \infty$, only the largest difference determines the distance. This is called the *Chebychev distance*:

$$d_\infty(\mathbf{x}_k, \mathbf{x}_l) = \max_j |x_{kj} - x_{lj}|. \quad (5)$$

Mahalanobis Distance

The Euclidean distance tends to an overestimation of highly correlated variables. The *Mahalanobis distance* (Mahalanobis, 1936) is a generalization of the Euclidean distance that compensates for any of these intercorrelations by weighting the Euclidean distance with the inverse covariance matrix of the data, $\mathbf{C}^{-1}$. It should therefore be applied in favor of the Euclidean distance when two or more variables of the data set are correlated to a considerable degree.

$$d_m(\mathbf{x}_k, \mathbf{x}_l) \equiv \sqrt{(\mathbf{x}_k - \mathbf{x}_l)^T \mathbf{C}^{-1} (\mathbf{x}_k - \mathbf{x}_l)}. \quad (6)$$

In the case of uncorrelated variables, $\mathbf{C}$ and $\mathbf{C}^{-1}$ become diagonal matrices and the Mahalanobis distance is proportional to the Euclidean distance.

Coefficient of Shape Difference

The distance measures discussed so far will change their value if one vector is subjected to a contraction or dilation (multiplication by scalar $a \neq 1$) or to an additive displacement (addition of vector $\mathbf{c}$ with all elements set to a constant $c \neq 0$) (Romesburg, 1984):

$$d(\mathbf{x}_k, \mathbf{x}_l) \neq d(a\mathbf{x}_k + \mathbf{c}, \mathbf{x}_l), \quad (7)$$

with

$$\mathbf{c} = c \begin{pmatrix} 1 \\ 1 \\ \vdots \\ 1 \end{pmatrix}. \quad (8)$$

In contrast, the coefficient of shape difference introduced by Penrose (1953) ignores additive displacements. It is defined as

APPENDIX (Continued)

$$\begin{aligned}
d_p(\mathbf{x}_k, \mathbf{x}_l) &\equiv \sqrt{\frac{1}{n} \sum_{j=1}^n (x_{kj} - x_{lj})^2 - \left(\frac{1}{n} \sum_{j=1}^n (x_{kj} - x_{lj}) \right)^2} \\
&= \sqrt{\frac{1}{n} d_2^2(\mathbf{x}_k, \mathbf{x}_l) - \left(\frac{1}{n} \sum_{j=1}^n (x_{kj} - x_{lj}) \right)^2}. \tag{9}
\end{aligned}$$

From Equation 9, it can be seen that the first term is the squared average Euclidean distance. The second term denotes the squared average distance due to additive displacements and is subtracted from the first term. Consequently, two vectors that only differ by an additive constant c yield a Penrose distance of 0:

$$\begin{aligned}
d_p(\mathbf{x}_k, \mathbf{x}_k + \mathbf{c}) &= \sqrt{\frac{1}{n} \sum_{j=1}^n (x_{kj} - [x_{kj} + c])^2 - \left(\frac{1}{n} \sum_{j=1}^n (x_{kj} - [x_{kj} + c]) \right)^2} \\
&= \sqrt{\frac{1}{n} nc^2 - \left(-\frac{1}{n} nc \right)^2} = 0. \tag{10}
\end{aligned}$$

Cosine Coefficient

The cosine coefficient is a measure of resemblance. It gives the cosine of the angle θ between two vectors,

$$d_c(\mathbf{x}_k, \mathbf{x}_l) \equiv \cos \theta = \frac{\mathbf{x}_k \cdot \mathbf{x}_l}{\|\mathbf{x}_k\| \|\mathbf{x}_l\|} = \frac{\sum_{j=1}^n x_{kj} x_{lj}}{\sqrt{\sum_{j=1}^n x_{kj}^2 \sum_{j=1}^n x_{lj}^2}}, \tag{11}$$

thereby ranging from -1 (minimum similarity) to 1 (maximum similarity) and neglecting any differences in length (contractions or dilations due to multiplication by a scalar $a > 0$). A vector and its proportional transform will therefore obtain maximum similarity:

$$d_c(\mathbf{x}_k, a\mathbf{x}_k) = \frac{\mathbf{x}_k \cdot (a\mathbf{x}_k)}{\|\mathbf{x}_k\| \|a\mathbf{x}_k\|} = \frac{a(\mathbf{x}_k \cdot \mathbf{x}_k)}{|a| \|\mathbf{x}_k\| \|\mathbf{x}_k\|} = \frac{a \|\mathbf{x}_k\|^2}{|a| \|\mathbf{x}_k\|^2} = 1 \text{ for } a > 0. \tag{12}$$

Correlation Coefficient

The familiar Pearson product-moment correlation is given by

$$d_R(\mathbf{x}_k, \mathbf{x}_l) \equiv r(\mathbf{x}_k, \mathbf{x}_l) \equiv \frac{\sum_{j=1}^n (x_{kj} - \bar{x}_k)(x_{lj} - \bar{x}_l)}{\sqrt{\sum_{j=1}^n (x_{kj} - \bar{x}_k)^2 \sum_{j=1}^n (x_{lj} - \bar{x}_l)^2}}. \tag{13}$$

The correlation coefficient denotes the proportion of covariance of two vectors in relation to their total variance. Since any additive and proportional transformation of one vector affects both the numerator and denominator of Equation 13 alike, the correlation coefficient is invariant for such transformations. A vector and its proportional and/or additive transform will therefore obtain maximum similarity:

$$\begin{aligned}
d_R(\mathbf{x}_k, a\mathbf{x}_k + \mathbf{c}) &= \frac{\sum_{j=1}^n (x_{kj} - \bar{x}_k)(ax_{kj} + c - [a\bar{x}_k + c])}{\sqrt{\sum_{j=1}^n (x_{kj} - \bar{x}_k)^2 \sum_{j=1}^n (ax_{kj} + c - [a\bar{x}_k + c])^2}} \\
&= \frac{\sum_{j=1}^n (x_{kj} - \bar{x}_k)(a[x_{kj} - \bar{x}_k])}{\sqrt{\sum_{j=1}^n (x_{kj} - \bar{x}_k)^2 \sum_{j=1}^n (a[x_{kj} - \bar{x}_k])^2}} \\
&= \frac{a \sum_{j=1}^n (x_{kj} - \bar{x}_k)^2}{\sqrt{a^2 \left(\sum_{j=1}^n (x_{kj} - \bar{x}_k)^2 \right)^2}} = 1 \text{ for } a > 0. \tag{14}
\end{aligned}$$

APPENDIX (Continued)

Standardization of the Data Matrix

With the exception of the Mahalanobis distance, all coefficients discussed here are subject to bias if two or more attributes differ in their units of measurement. This is a result of the fact that attributes with smaller units carry more weight in the calculation, and therefore EQUIWORD offers two methods of data normalization. The common z standardization normalizes all attributes to the same unit (multiples of the standard deviation) before performing any calculation. (Note that the Mahalanobis distance always does an implicit z standardization.) Thus, the attribute values of each word are replaced by their corresponding z score, yielding the standardized data matrix $\mathbf{Z}$ with elements z_{ij} given by

$$z_{ij} \equiv \frac{x_{ij} - \bar{x}_{.j}}{s_{.j}}. \quad (15)$$

On the other hand, the following standardization uses the minimum and maximum value of an attribute to confine all values to a range between 0 and 1, thereby expressing them in units of proportions of the range:

$$p_{ij} \equiv \frac{x_{ij} - \min_{.j}}{\max_{.j} - \min_{.j}}. \quad (16)$$

(Manuscript received October 20, 2004;
revision accepted for publication January 27, 2005.)