

Communication & Science

JOURNAL

2011 Oct 10

Chen, H., Evans, C., Battleson, B., Zubrow, E., Woelfel, J. (2011). Procedures for the precise analysis of massive textual datasets. *Communication & Science Journal*, 2011Oct10. (received 2011Jul23)



A RAH Press Publication

<http://www.comSciJ.com>

**PROCEDURES FOR THE PRECISE ANALYSIS OF MASSIVE
TEXTUAL DATASETS**

by

Hao Chen, Carolyn Evans, Brenda Battleson, Ezra Zubrow, Joseph Woelfel

9 October 2011

(Paper originally presented at Sunbelt XXVIII: International Sunbelt
Social Network Conference, St. Pete Beach, FL.)

ABSTRACT

Unsupervised Artificial Neural Networks have been used in the analysis of text. In general, they provide richer, deeper and more finely detailed clusters than co-occurrence models because of their ability to consider indirect connections among words. Since the number of possible indirect connections increases exponentially with increases in size of the network, this advantage should be greatly amplified in very large datasets. In this study over 4,500 world news articles about disability were gathered and analyzed using a large artificial neural network running on the Center for Computational Research (CCR)'s supercomputing cluster at the State University of New York at Buffalo. Increasing the size of the artificial neural network allowed more connections between concepts to be discovered. This lead to a network that was better trained and results that were more detailed and informative, showing more depth.

1 INTRODUCTION

In an age of information explosion, data digitization is an inevitable decision for most companies, institutes, government agencies, and individuals for archiving purpose. Collecting this digitized data on a voluminous scale with the help of computer programs is now possible, although previously it was impractical, if not impossible, to do so manually. Another task which is difficult without computers is to analyze the data. "From its current position as a research tool, the future presented herein as one in which computer content analysis becomes a crucial agent in taming an ever-increasing torrent of information, and an important aspect of many commercial operations" (West, 2001). Together with the advance in the computational power available, this might be the motive force behind "computer content analysis has undergone something of a renaissance" (West, 2001).

Diefenbach (2001) gave a definition on the role of the computer in content analysis:

There are many advantages to using computer support in content analysis. First, computers can do menial tasks, such as repetitive counting and sorting, and thereby liberating researchers for more theoretical and creative tasks. Second, not only can a computer do all the counting, but it can do so with perfect reliability. That is, computer-produced results will be the same every time the data

are counted or otherwise examined. This means that when using units of analysis, such as words or other symbols, the computer can generate perfectly reliable frequency counts. The computer cannot, however, differentiate the senses of homographs (i.e., words with different meanings that are spelled the same) without explicit disambiguation instructions nor can it analyze units larger than words, such as phrases or themes, or make evaluations, without explicit and detailed instructions. A computer will do exactly what you tell it to do, but it can only do exactly what you tell it to do. There is no room for a computer to make judgments or interpretations unless it is given explicit subroutines explaining how to make a “judgment.” (pp. 14–15)

Another way content analysis is often done is by preparing extensive transcripts, hand-coding passages according to predefined categories, and manually counting the categorical instances to determine the most common themes of a given work (Walberg et al., 2001). Drawbacks of this approach are: subjectivity in formulating the set of categories, lack of reliability in coding or categorizing passages, and laborious effort in coding and analyzing the frequencies of the categories (Walberg et al., 2001). As the above definition shows, computer content analysis helps to fix all these problems; it can greatly reduce the effort in coding and data analyses, it can assure reliability by reporting the same results for each run, and it can do so without subjective interference of researchers.

The above definition stated incorrectly, however, that computer programs cannot analyze units larger than words, such as phrases or themes. Some programs do have the ability to find out the pattern of words which can then be formed into word phrases or themes. For example, Danowski (1982, 1993, 2001) mapped word co-occurrences in a sliding window of text into a matrix, then partitioned the stream of word pairs into a network of words by applying clustering method(s) on the co-occurrence matrix. Woelfel (1993) implemented an artificial network layer on top of this co-occurrence approach to find the strongest linkages among frequent words in the text (Walberg, 2001). The latter approach has a few advantages over the former. When using co-occurrence the window¹ slides through the text and only words appearing inside the window at the same time are counted as *co-occurring* (so only direct connections between words are counted). Using a neural network, however, will also allow indirect connections to be incorporated into the word connection weights. This will reveal more hidden patterns of words in the text, hence it will provide a more precise connection strength matrix. As a result, the clusters produced from the matrix are more clearly defined.

Katmandu, a computer program currently being developed, boosts its analyzing power by allowing more unique words to be counted as nodes in the artificial neural network. The present study reports what this change brings to the field of computer content analysis and why it is important.

¹The window size can be adjusted.

1.1 History of Development

As indicated above, CATPAC/Katmandu is an artificial neural network text analysis tool which performs neural network, cluster, and perceptual space analyses on qualitative input. Using an unsupervised self-organizing artificial neural network as the core, CATPAC/Katmandu has a number of advantages (more precise results and clearer linkages between concepts) when compared to other text analysis programs based only on a co-occurrence model.

CATPAC, a precursor of Katmandu, has been widely-adopted in the field of content analysis. The initial development of CATPAC dated back to the late 70s. In 1985, Danowski's *moving window* was implemented (The program's codename at that time was Newton). Then in 1993, artificial neural network was introduced and the program was renamed as CATPAC. In 2007, the total limit of unique words was expanded, and the program was renamed as Katmandu.

Previously, the Windows version of CATPAC allowed as many as 300 unique words for its analysis. The updated version, Katmandu, expanded this limit to 1,300. Although it is not a significant improvement numerically, this expansion has an immense impact on the result; it also shows that expanding the limit of unique words (increasing the word limit also increases the size of the artificial neural network at the back-end) achieves greater precision in the text analysis.

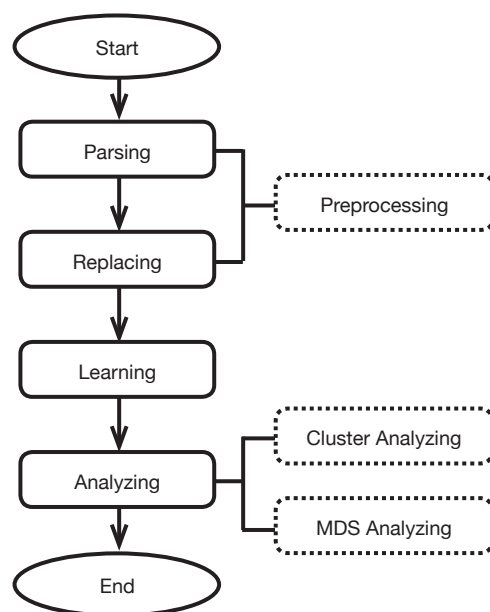


Figure 1.1: The Mechanism of CATPAC/Katmandu

1.2 Mechanism of CATPAC/Katmandu

Workflow

The workflow of CATPAC/Katmandu is demonstrated by Figure 1.1.

First, data are collected in a plain text file. There is no special formatting or predefined dictionary needed.

Second, the text is considered as a data stream and the whole text is parsed into single words (*Parsing*). At this stage, the frequency of each unique word is calculated. Then if the words in the excluded list are activated, those words on the list will be omitted in the forthcoming process.

These two steps are called *Preprocessing*.

At this stage, a list of words ordered by their frequencies in the text can be obtained. As stated by Diefenbach (2001), “Since English-language usage produces far fewer higher- than lower-frequency words, a list of the dominant words in a document can be a meaningful tool for analysis, but simple frequency counts are limited by the issue of context”. CATPAC/Katmandu will also try to find out how words are related to each other in the text. It would be ideal to include each unique word (as unique *concept*) for the followup process; however, due to limitation of computing power, only parts of words can be included for the final analysis (the less frequent words will be omitted). As mentioned before, the Windows version of CATPAC allows 300 unique words (the top 300 most frequent words).

Next, together with Danowski’s “moving window” method, CATPAC will “learn” the text with its artificial neural network engine. From the point of view of co-occurrence approach, when a window is sliding through the text, words coexist inside the window at the same time are counted as co-occurred pairs. From the point of view of neural network approach, unique words are treated as “neurons” and when they coexist inside a same window, they are “activated” (it is worthy to mention CATPAC can do both co-occurrence and neural network analysis). CATPAC uses a learning rule called “Hebbian Learning” to guide the learning (this will be discussed in more detail in the next subsection). Basically, the

neuron will remember the link with another neuron when they coexist (are “activated”) in the previous window and it will pass that connection on to the next window. Hence, indirect connection between words are enabled for analyzing which is not possible in the co-occurrence model.

Finally, in order to present the data in an accessible way, CATPAC performs a hierarchical cluster analysis on the connection strength matrix collected in the previous step. One can also choose to perform a multidimensional scaling analysis on the same matrix to produce mapping of the words. With the help of a plotting program, the interrelation between words can be shown in a 2D/3D way to facilitate interpretation for researchers.

Learning Rule

The learning rule used in CATPAC’s neural network engine is a “Hebbian Learning” rule. As Stent (1973) reviewed in his article, Hebb (1949) formulated his “neurophysiological postulate” of learning which states: “When an axon of cell A is near enough to excite a cell B and repeatedly and persistently takes part in firing it, some growth or metabolic change takes place in one or both cells such that A’s efficiency, as one of the cells firing B, is increased.”

As indicated in this rule:

1. If two neurons on either side of a connection are activated synchronously, then the strength of that synapse is selectively increased;

2. If two neurons on either side of a connection are activated asynchronously, then that synapse is selectively weakened or eliminated.

When this rule is applied in content analysis, unique words are considered as “neurons”. In Danowski’s (1982) “moving window” approach, the window becomes an environment in which the neurons can react. When a pair words appear in the window at each iteration when the window slides, they are activated synchronously and the connection between them is selectively strengthened. On the other hand when they do not appear in the window at each iteration, their connection is selectively weakened or deleted. In CATPAC, for the second situation (when two words are activated asynchronously), the connection between words will not be deleted, but instead, be weakened. In this way, CATPAC can “remember” the indirect connection between words. For example, word A and word B keep showing up in a pair in a sample text, when in the same text, word A and word C keep showing up in a pair too. However, B and C never show up together anywhere in the text. Theoretically, B and C are still related since they share the same connection to A. The connection between them might not be as strong as their respective connections to A, but it is still there.

To enable indirect connections between words helps to reveal hidden patterns, hence, more information can be found. Also, as the sliding window goes through the entire document, the neural network keeps training itself with the new information until the window reaches the last word of the text. New linkages are built and old ones are retrained (strength-

ened or weakened) to adjust to the “environment” (the whole text) in this process. As an end result, the cluster analysis performed on the connection strength matrix collected will show more precise patterns. Since the neural network model can find more connections between words (direct connections plus indirect connections) than the co-occurrence model, the dendogram created with the neural network model will have a deeper depth than the one created with the co-occurrence model. In other words, researchers can find more variances between different clusters.

Moreover, the parameters of neural network learning can be adjusted to fit different situations. Among these parameters are clamping, threshold, decay rate, and learning rate (Salisbury, 2001, pp. 73):

Clamping Choosing to have the program clamp the nodes instructs the program to have the concepts related to one another. Clamping the nodes allows the program to retain the concepts in its perceptual field and, therefore, constantly be aware of them. As a result, those concepts which are strongly related to other concepts emerge as such because the links between the nodes are strengthened. If clamping is not chosen, the concepts are not retained in the perceptual field...

Threshold Each neuron is either turned on or else receives inputs from other neurons with which it is connected. These

inputs are transformed by an activation function. CATPAC can use one of four transfer functions: a linear function varying between -1 and +1, a logistic function ranging between 0 and +1, a logistic function varying between -1 and +1, and a hyperbolic tangent function varying between -1 and +1 (Woelfel, 1991)...Once the inputs to any neuron have been transformed by the transfer function, they are summed, and if they exceed a given threshold value the nodes are activated. Otherwise, they are left inactive. Lowering the threshold makes it more likely for neurons to be activated. Raising the threshold makes activation less likely...

Decay Rate The decay rate function refers to the rate at which a node loses a proportion of its activation as a function of time due to restoring forces that return nodes to their resting levels. The decay rate specifies how quickly the neurons return to their rest condition after being activated. Raising the rate turns them off faster. Lowering the rate keeps them on longer...

Learning Rate When neurons behave similarly, the connections between them are strengthened. The learning rate determines how much they are strengthened each cycle.

CATPAC/Katmandu allows users to change all these parameters. Researchers can experiment with these settings to find out the optimal ones for their situations. For example, if results show too few clusters, lowering the learning rate can help find more clusters.

1.3 Human Interpretation

Although a neural network is very good at finding patterns or associational relations between words in a text, it cannot understand what it has found. It regards the whole text as a data stream and sees unique words as individual neurons (“concepts”). It is still the responsibility of the researcher to figure out the meaning of the clusters. To avoid bias during interpretation of clustered result, combining several researchers’ opinions and getting consensus are necessary.

1.4 Bigger the Size of the Artificial Neural Network Means What?

Katmandu, the current version of CATPAC, increases the limit of 300 unique words to 1,300. What does that mean for the analysis? As was discussed previously, due to computing power limitation, it is still not possible to include each unique word in a large document as a unique word for the final analysis. Increasing this limit makes the most recent

version closer to the ideal (the ideal being able to find out how each word is connected to other words), however, and this helps to understand what the content really means. Previously with the limit of 300, when the window was sliding through text, those words which were not as frequent as the top 300 were overlooked. Their relation to other words was also disregarded. By increasing this unique word limit more words can be considered in the analysis. The artificial neural network trained by the text will fit better with the real words network of the target. Since more unique words are counted as valid neurons, the size of some final clusters built on those words is expanded. In other words, some clusters might include more items than before when the limit is set as 300. This is very helpful for the researchers (“the interpreters”) when they are considering the meaning of the clusters. Moreover, since changing the limit for the unique words means changing the size of the connection matrix, the end product—clusters—will have a larger depth which will show more difference between clusters (compare the depths of clusters between the dendograms in Appendix 1 and Appendix 2).

2 METHOD

The current paper shows an application of Katmandu on a project about how disability is portrayed by online media. We developed a procedure for collecting news articles using Google Alerts¹. The tool queried Google News website with disability related keywords each day for over 6 months. The news articles found with this tool were then combined into a single plain text file (MasterText.txt). This file was then examined using Katmandu. No special formatting or pre-defined dictionary was needed. The file did, however, have to be encoded in ASCII (American Standard Code for Information Interchange) which meant only English alphabets are allowed. Hence, this project only concerns those news articles on disability in English from Google News website. For those who are interested in broadening this range to include other languages, they can consult another content analysis program called Wölpak which allows Unicode-encoded input (Woelfel et al., 2005; Evans et al., 2008).

In order to save time on processing data and to test the most powerful computing power available to us, the program Katmandu was moved to the Center for Computational Research (CCR) supercomputing center's cluster machines at University at Buffalo and all the analyses were done on them. The same text file was run multiple times with different numbers of unique words and the number of final clusters found were compared.

¹<http://www.google.com/alerts>

The default neural network learning parameters, as preset by Katmandu, were used for all runs so as not to bias the comparisons. These parameters were: Clamping = "YES", Threshold = .0, Decay Rate = .1, and Learning Rate = .05. Also, the sliding window size was set as 7 and the size of slide was set as 1 (each time the window slides one word).

The CPU time each run took was also recorded as a reference to how much computing power was consumed. Since increasing the size of the limit of unique words is a continuing task to apply the latest computer technology for enhancing the program's processing power, this kind of data will be valuable for program developers.

3 RESULTS

The size of the input file (MasterText.txt) is 14.1 megabytes, and the file contains 11,862,495 total words and 21,687 unique words. Table 1 is the table for each experiment with different unique word limits set using neural network method.

The counting of total clusters found is arbitrary. By drawing a horizontal line at the center of the dendrogram, whichever clusters stand above the center line is counted in total clusters found. Appendix One displays a sample dendrogram showing cluster analysis when the limit of unique words was set as 50. There were 11 clusters found above the arbitrarily drawn line. If the limit is increased to 100, 4 more clusters are found (15 in total, see Appendix Two). When the limit is boosted to 200, the total clusters found increases to 22. At the level of 300 (17 clusters found in

Table 3.1: Different number of clusters found when the limit of unique words is set differently (using neural network method).

Unique Words	Clusters Found	CPU Time
50	11	1.40
100	15	4.05
200	22	15.32
300	17	57.53
400	13	125.91
500	17	198.79
1,000	12	903.07
1,300	17	1949.67

Note: Time measured in seconds.

total) and 400 (13 clusters found in total), however, the clusters found are fewer than at the level of 200 (22 clusters found in total). This finding is understandable and predictable, because increasing the limit of unique words means more words are recognized by the artificial neural network when it is “learning” the text. The words assigned to different clusters before might be connected together after “new” words are added in, because such “new” words help to link those previously different clusters into a big single cluster. For example, at the level of 50, the word *DRUG* and the word *ATTENTION* do not belong to the same cluster and there are even 3 clusters between the clusters they belong to. At the level of 100, these two are side by side in the same cluster. When the limit of unique words is increased to 1,300 (currently the largest limit Katmandu allows), the number of total clusters found doesn’t vary much to previous findings (17 clusters found in total).

As shown in this section, counting of total clusters found in a dendrogram is an arbitrary task (the method this paper used is to draw an arbitrary line at the center of the dendrogram). It is an issue about the “degree” chosen by researchers. To think about it, every word is connected to each other word at some degree if the whole text is considered as a big box. From the dendrogram, every word is connected to each other at the bottom. If similar number of clusters are found when the limit of one run is lower and the limit of another run is higher, the difference between them is the size of some clusters (especially the major ones): smaller the size for

the lower limit and bigger for the higher. Researchers can either choose smaller size for simplified result or bigger size for informative result. It is worthy noting that if too few unique words are chosen it might affect the validity of the result. For this study, if 50 is chosen as the limit, only 11 clusters are found. As compared to 22 clusters found at the level of 200, there is a deficit of 11 clusters which means a lot of information is lost. On the other hand, it is acceptable to choose the limit between 100 and 1,300, because the numbers of total clusters found do not vary much in this range. It is recommended that researchers run multiple trials with different limits to find the optimal choice for a particular dataset.

Since increasing the limit will increase the size of the neural network, the learning process will consume more computing power (take longer time to run). As shown in Table 1, at the level of 50 unique words, it only takes 1.40 seconds to process; at the level of 100, it takes 4.05 seconds; at the level of 500, it takes 198.79 seconds; at the level of 1,300, it takes 1949.67 seconds, a little bit over half an hour (all these time are measured as CPU time which means how long the program Katmandu takes to process the data). These numbers clearly illustrate that as the size of neural network increases, the computing power consumed increases almost exponentially. As tradeoff, however, researchers have a much more informative clustered report.

As a comparison, the same procedure is used on testing co-occurrence method. Table 2 is the table for each experiment with different unique

Table 3.2: Different number of clusters found when the limit of unique words is set differently (using co-occurrence method).

Unique Words	Clusters Found	CPU Time
50	8	0.33
100	13	0.63
200	13	1.92
300	19	9.58
400	20	19.75
500	21	32.24
1,000	21	145.38
1,300	25	314.96

Note: Time measured in seconds.

word limits set using co-occurrence method.

The most notable difference between the neural network method and the co-occurrence method is that when using co-occurrence method the number of clusters found steadily increases when the unique word limit is increased. Since co-occurrence cannot find the indirect relation between concepts, the clusters found are all about directly related concepts. The average size of these clusters is smaller than those found in neural network method. Since only direct relation will be found in the co-occurrence method, the relation between different clusters are distinctive. It is rarely seen that different clusters converge into bigger ones. The size of clusters is kept mostly the same from the top of the dendrogram to the place close to the bottom. This is exactly the reason why more clusters will be found if the unique word limit is increased. Also, the depth of the dendograms is much smaller than those found in neural network method (see Appendix

for more detail). Since the depth of dendograms is related to how many steps the hierarchical clustering method takes to calculate all the possible clusters, this difference between two methods shows that neural network method has a higher degree of precision on clustering.

4 DISCUSSION

As stated by Stevenson (2001), “Computers are still dumb clerks. Programs which parse a text file into its grammatical components or search for any level of meaning are still in their infancy.” Implementing an artificial neural network helps the program to learn better from the text and to find more precise word patterns, but human interpretation is still necessary to make sense out of these patterns. However, “dumb clerks are useful in content analysis because so much of the work involves searching for relevant material. We may still spend a lot of time in a content analysis project searching for a few needles in very large haystacks, so automating even that part of the job is no insignificant accomplishment” (Stevenson, 2001).

This paper describes progress on a content analysis tool implementing an artificial neural network as its core. Since neural networks allow indirect connection between words, more patterns in the text may be discovered than using the co-occurrence model. Moreover, the parameters of the neural network can be modified to make the network adapt better to the text. Hence, neural network “learning” will produce a more precise and informative result. Indeed, as the size of the neural network is increased, the complexity of the network is increased. As a direct result, the program is better able to learn from the target text. This saves researchers not only time and effort in relation to parsing text into predefined categories

manually (which may also introduce bias), but also produces a more precise data.

As shown in the experiment above, increasing the limit of unique words expands the size of some major clusters. This means human researchers will have more hints to construct the “meaning” (make a sentence or a situation) for the words inside a single cluster. However, it should be remembered that this is not necessarily advantageous when the size of a cluster becomes too large. One possible solution is to lower the learning rate of the neural network; more clusters will be found in this way (by subdividing large single clusters into smaller ones). This could, however, lead to another problem that is, there may be too many clusters to work with.

An alternative approach (Battleson et al., 2001) is a non-hierarchical clustering method. The clusters are then not calculated in a top-to-down or down-to-top methods. Rather, such a system uses keywords to explore the results, and it will find out what other words are related to the keywords and then form a cluster with them. Finding such clusters is totally nonlinear. This means a word can appear in different clusters depending how many other different words it is related to. For example, *mustang* could mean a brand for car, and it is related to the car cluster (which could have concepts like Volvo, BMW, or Benz). It could also mean a half-wild horse, and it is related to the horse cluster (which could have concepts like colt, bronco, or pony). The non-hierarchical clustering approach will

allow this one-to-many relation in the word-to-cluster relation. Different meanings of a word can be explored in this way. A tool, called Listiac, is developed to achieve this goal (Battleson et al., 2001).

Finally, for those researchers interested in analyzing text in languages other than English, Wölpak (Woelfel et al., 2005; Evans et al., 2008) is another content analysis tool which implements similar methods to those introduced in this paper. Wölpak allows Unicode-encoded text as input; this allows most of the world's writing systems as input.

BIBLIOGRAPHY

- Battleson, B., Chen, H., Evans, C., & Woelfel, J. K. (2008, January). *A non-hierarchical neural network approach for analyzing textual data*. Poster session presented at Sunbelt XXVIII: International Sunbelt Social Network Conference in St. Pete Beach, FL.
- Changeux, J. P., & Danchin, A. (1976). Selective stabilization of developing synapses as a mechanism for the specification of neural networks. *Nature*, 264, 705–712.
- Danowski, J. A. (1982). Computer-mediated communication: A network-based content analysis using a cbbs conference. In M. Burgoon (Ed.), *Communication yearbook* (Vol. 6, pp. 905–924). Beverly Hills, CA: Sage.
- Danowski, J. A. (1993). Network analysis of message content. In G. Barnett & W. Richards (Eds.), *Progress in communication sciences* (Vol. 12, pp. 197–222). Norwood, NJ: Ablex.
- Danowski, J. A., & Lind, R. A. (2001). Linking gender language in news about presidential candidates to gender gaps in polls: A time-series analysis of the 1996 campaign. In M. D. West (Ed.), *Progress in communication sciences* (Vol. 17, pp. 87–102). Westport, CT: Ablex Publishing.
- Diefenbach, D. L. (2001). Historical foundations of computer-assisted content analysis. In M. D. West (Ed.), *Progress in communication sciences* (Vol. 16, pp. 13–16). Westport, CT: Ablex Publishing.
- Doerfel, M. L., & Barnett, G. A. (1996). The use of catpac for textual analysis. *Cultural Anthropology Methods*, 8, 4–7.

- Evans, C., Chen, H., Battleson, B., & Woelfel, J. K. (2008, January). *Artificial neural networks for pattern recognition in multilingual text*. Poster session presented at Sunbelt XXVIII: International Sunbelt Social Network Conference in St. Pete Beach, FL.
- Hebb, D. O. (1949). *Organization of behavior*. New York: John Wiley & Sons.
- Salisbury, J. G. T. (2001). Using neural networks to assess corporate image. In M. D. West (Ed.), *Progress in communication sciences* (Vol. 17, pp. 65–85). Westport, CT: Ablex Publishing.
- Stent, G. S. (1973). A physiological mechanism for Hebb's postulate of learning. *Proceedings of the National Academy of Sciences of the U.S.A.*, 70, 997–1001.
- Stevenson, R. L. (2001). In praise of dumb clerks: Computer-assisted content analysis. In M. D. West (Ed.), *Progress in communication sciences* (Vol. 16, pp. 3–12). Westport, CT: Ablex Publishing.
- Walberg, H. J., Arian, G. W., Paik, S. J., & Miller, J. (2001). New methods of content analysis in education, evaluation, and psychology. In M. D. West (Ed.), *Progress in communication sciences* (Vol. 16, pp. 143–158). Westport, CT: Ablex Publishing.
- West, M. D. (2001). The future of computer content analysis: Trends, unexplored lands, and speculations. In M. D. West (Ed.), *Progress in communication sciences* (Vol. 16, pp. 159–175). Westport, CT: Ablex Publishing.
- Woelfel, J. (1991). *Cascaid user's manual*. New York.

Woelfel, J., & Stoyanoff, N. J. (1993, September). *Catpac: A neural network for qualitative analysis of text*. Paper presented at the Annual Meeting of the Australian Marketing Association, Melbourne, Australia.

Woelfel, J. K., Chen, H., Kim, J. H., Sharma, B., Woelfel, J., Cheong, P., et al. (2005, May). *Artificial neural networks for the analysis of text across cultures and languages*. Paper presented at the International Communication Association conference in Dresden, Germany.

27

B SAMPLE DENDOGRAM 2: 100 UNIQUE WORDS (CREATED WITH NEURAL NETWORK METHOD)

DIAMETER METHOD

0.56153619E+00	C	T	F	F	L	L	R	A	N	N	T	Y	L	H	C	S	Y	M	D	A	D	H	A	P	P	F	W	C	W	L	N	H	T	P	W	S	N	P	T	G	H	S	T	P	S	E	M	O	M	
0.29886851E+00	H	W	I	A	I	E	E	C	E	E	H	E	O	E	O	C	E	E	R	T	E	Y	I	D	E	R	O	I	O	W	A	E	O	I	R	O	U	A	A	E	R	I	A	R	R	I	F	E	V	E
0.26038569E+00	I	O	R	M	F	A	S	C	E	W	I	A	N	L	N	H	A	T	U	T	F	P	S	U	R	O	O	L	M	W	S	W	M	M	O	R	P	T	R	A	O	G	Y	E	E	D	F	D	E	
0.12404771E+00	L	.	S	I	E	D	E	O	.	N	R	G	P	D	O	R	H	G	E	I	E	O	L	C	G	D	.	.	T	S	E	E	B	K	P	I	E	C	U	H	S	A	S	E	E	I	R	I		
0.11239886E+00	D	.	T	L	.	A	R	.	K	.	.	I	O	S	Y	.	N	C	R	T	E	L	.	O	O	N	H	P	E	.	T	C	.	C	C	.	C	C	
0.11113122E+00	R	.	Y	.	R	D	.	.	.	T	L	.	L	.	T	I	A	D	S	N	A	E	.	R	N	T	E	.	R	.	M	H	.	T	A	.	A			
0.11071443E+00	E	.	.	.	C	I	.	.	.	I	.	P	.	I	T	C	E	.	T	M	M	.	T	A	S	R	.	.	E	O	.	S	T	.	T	.	T			
0.79989977E-01	N	.	.	.	H	N	.	.	.	O	.	H	.	O	.	T	R	S	.	L	.	S	N	O	.	I	.	I	.	I				
0.79473488E-01	G	N	.	E	.	N	.	I		
0.61954848E-01	N	N		
0.51719833E-01	I	
0.26460744E-01	D	
0.13151607E-01	A	
0.11237700E-01	T	
0.73528765E-02	E	

[illegible]

[illegible]

33

C SAMPLE DENDOGRAM 3: 50 UNIQUE WORDS (CREATED WITH CO-OCCURRENCE METHOD)

DIAMETER METHOD

ADHSTPSACD	TYDPKRTTFDA	SDSNWUHRPT	CHPPGFM	RSSYP	SMIMS
TEIYYRRLD	HIWEREIIIR	DOTEINEEA	HEEEROE	IEICTER	TENEI
TFSPMEOE	HIAAURDC	MRUUNUWL	IASRAIL	ORUUTTH	UAODFDD
EIOEPABED	LG.RGCSKES	GLD.LVLEE	CLPP.UNH	AODRGD	IOIE
NCRRITLP	DN.S.E.Y	TST.Y.ET	ANHD.L	PDYLOI	.RECR.
TIDAOME	.RO..N...	.S...RHR	TE.E...L	LILE.A	ANAMA.
ITECMEM	.ES.T...	.S.CSR...	.PN.S	MTTAT.	
O.RITSNS	.NE..I.H	.S...H...	.SITTI		
N.I.T	.D...T...	.Y...N...	.NON.		
.V...	.I...	.I...	.I...	.SN.	
.T...	.T...	.D...	.D...		
.Y...	.Y...	.A...	.A...		
.T...	.T...	.T...	.T...		
.E...	.E...	.E...	.E...		
0.2900000E+03					
0.2620000E+03					
0.2080000E+03					
0.6800000E+02					
0.6400000E+02					
0.5000000E+02					
0.3200000E+02					
0.3100000E+02					
0.3000000E+02					
0.2200000E+02					
----- Arbitrary Center Line -----					
0.2000000E+02					
0.1800000E+02					
0.1600000E+02					
0.1400000E+02					

[illegible]

D SAMPLE DENDOGRAM 4: 100 UNIQUE WORDS (CREATED WITH CO-OCCURRENCE METHOD)

DIAMETER METHOD

0.29000000E+03	A	D	H	P	S	A	C	D	T	Y	M	P	T	F	A	D	S	N	W	U	P	T	M	R	P	S	S	S	D	T	C	B	T	R	M	F	L	P	A	N	H	M	H	C	D	H	S	E	O
0.26200000E+03	T	E	I	Y	R	L	D	H	I	W	E	E	I	I	D	O	T	E	I	N	A	E	E	I	R	T	T	A	R	A	H	R	R	E	E	A	I	R	M	A	E	E	E	O	I	I	F	V	
0.20800000E+03	T	F	S	P	O	E	H	I	A	O	A	D	R	M	R	U	N	U	W	L	I	R	A	T	T	O	U	Y	U	K	I	A	E	S	D	M	F	E	O	T	A	N	L	N	A	G	D	F	E
0.84000000E+02	E	I	O	E	B	E	D	L	G	.R	I	C	E	S	L	.D	.L	V	E	C	H	A	G	D	D	S	G	I	L	I	A	E	I	I	E	S	N	I	L	T	P	D	G	H	E	E	R		
0.72000000E+02	N	C	R	R	L	P	.D	N	.S	C	E	.T	.Y	.E	N	H	L	R	E	I	.S	N	D	N	T	A	C	L	.C	G	O	T	A	.I	N	E	.C		
0.68000000E+02	T	I	D	A	E	.R	O	.A	N	.S	.S	R	P	N	M	T	S	.E	C	L	.O	A	.I	S	.S
0.64000000E+02	I	T	E	C	M	.E	S	.T	.I	.S	.H	.S	.T	.E	.S	.N	H	.O	L	.O	I	.S	
0.64000000E+02	O	.R	T	S	.N	E	.I	.S	.H	.S	.T	.E	.S	.N	H	.O	L	.O	I	.S	
0.58000000E+02	N	.I	.S	.D	.O	.N	.Y	.N	.I	.S	.H	.S	.T	.E	.S	.N	H	.O	L	.O	I	.S	
0.58000000E+02	.I	.S	.D	.O	.N	.Y	.N	.I	.S	.H	.S	.T	.E	.S	.N	H	.O	L	.O	I	.S	
0.38000000E+02	.T	.S	.D	.O	.N	.Y	.N	.I	.S	.H	.S	.T	.E	.S	.N	H	.O	L	.O	I	.S	
0.32000000E+02	.Y	.S	.D	.O	.N	.Y	.N	.I	.S	.H	.S	.T	.E	.S	.N	H	.O	L	.O	I	.S	
0.31000000E+02	.T	.S	.D	.O	.N	.Y	.N	.I	.S	.H	.S	.T	.E	.S	.N	H	.O	L	.O	I	.S	
0.30000000E+02	.Y	.S	.D	.O	.N	.Y	.N	.I	.S	.H	.S	.T	.E	.S	.N	H	.O	L	.O	I	.S	
0.24000000E+02	.T	.S	.D	.O	.N	.Y	.N	.I	.S	.H	.S	.T	.E	.S	.N	H	.O	L	.O	I	.S	
0.22000000E+02	.Y	.S	.D	.O	.N	.Y	.N	.I	.S	.H	.S	.T	.E	.S	.N	H	.O	L	.O	I	.S	
	----- Arbitrary Center Line -----																																																
	

0.29000000E+03	L	S	W	M	T	K	N	G	A	P	Y	F	R	A	T	I	S	S	S	H	Y	L	O	A	L	S	C	W	L	D	N	O	F	P	D	F	H	C	R	H	L	P	D	
0.26200000E+03	O	Y	O	E	R	I	E	R	G	E	O	O	E	C	H	N	U	E	C	I	E	A	C	D	E	H	O	W	I	E	E	F	A	A	R	O	O	A	I	O	I	E	A	O
0.20800000E+03	N	M	R	D	E	D	W	O	E	O	U	U	S	C	I	F	P	P	H	G	A	S	T	D	A	I	M	W	K	P	E	T	M	T	U	O	M	R	C	L	S	R	Y	W
0.84000000E+02	G	P	K	I	A	S	S	U	.	P	N	N	E	O	N	O	P	T	O	H	R	T	O	E	D	R	.	E	R	D	E	I	I	G	D	E	E	K	L	A	.	.	.	
0.72000000E+02	.	T	.	C	T	.	P	.	L	G	D	A	R	K	R	O	E	O	.	.	B	R	.	E	.	L	E	N	L	E	.	.	Y	A		
0.68000000E+02	.	O	.	A	.	.	E	.	R	D	.	M	R	M	L	.	.	E	A	.	.	Y	S	.	I	N		
0.64000000E+02	.	M	.	T	.	.	.	C	I	.	A	T	B	.	.	.	R	L	.	.	S	.	E	T		
0.58000000E+02	.	S	.	I	.	.	.	H	N	.	T	.	E	.	.	.	L	.	.	I	.	S	S			
0.38000000E+02	O	.	.	.	E	G	.	I	.	R			
0.32000000E+02	R	.	O			
0.31000000E+02			
0.30000000E+02			
0.24000000E+02			
	----- Arbitrary Center Line -----																																											
0.22000000E+02			
0.20000000E+02			
0.18000000E+02			
0.16000000E+02			
0.14000000E+02			
0.12000000E+02			
0.11000000E+02			
0.10000000E+02			
0.80000000E+01			
0.60000000E+01			
0.40000000E+01			

0.20000000E+01
0.00000000E+00