Words as Actors: A Method for Doing Semantic Network Analysis

Michael Schnegg and H. Russell Bernard

Field Methods 1996; 8; 7
DOI: 10.1177/1525822X960080020601

typos, spelling errors, or erroneous marks since the program reads them as different concepts. Also, CATPAC does not use a concept dictionary, so if concepts of interest have synonyms, the computer treats each synonym as a different concept.

Nor can CATPAC differentiate among different uses of the same word/concept. In the analysis of the presidential debates, "president" was clearly a concept of interest. The term, however, could be used in various contexts: "I want to be president because . . ." or "Mr. President, how do you feel about . . ." or "President Bush has been in the White House for . . ." and so on.

Different uses of single words, therefore, must be differentiated in the text by creating unambiguous word concepts for the program to recognize. Every time the word "president" was used as a title, a new word was created: "presidentbush." This cleaned up part of the problem, but "Mr. President" was a reference used for addressing President Bush as well, so that was also changed to "presidentbush." This enabled discrimination between the use of "president" as a title versus its use as a role in government ("President Bush" versus "I want to be president").

Changing "Mr. President" to "presidentbush," however, may have lost a potential element of analysis. Perhaps referring to "Mister" Bush rather than to "President" Bush indicated different levels of respect. With the adjustment of the data, there is no way to test this possibility.

The same shortcoming is apparent if you are trying to capture negative and positive connotations of the same concept. Some words carry negative or positive valence in and of themselves. Others need qualifiers to project such meaning. Finally, concepts that require two words to create meaning might be adjusted ("New York" would be changed to "newyork"). Although adjustments can be made in the data, one should be wary of changing too much of the syntax. Will the data still contain the original meaning?

Strictly speaking, these limitations are really a demand that the user exercise judgment in setting the parameters and in interpreting the results of the analysis. With proper caution and judgment, CATPAC is a convenient set of tools for analyzing and interpreting textual data.

For more information contact: Terra Research & Computing, 261 East Maple, Birmingham, MI 48009. Phone 810-258-9657. DOS version: $795 (professional), $350 (academic); Windows version $495 (professional) and $250 (academic).

Note

1. The cluster analysis is Johnson's (1967) diametric method, which is hierarchical. Nonhierarchical cluster analysis is an option and is useful for arranging categories that are not mutually exclusive (see Woelfel 1996).

References

Claffey, G. 1996. Customer feedback: Using content analysis to reduce uncertainty in a changing environment. Paper: annual meeting of the International Communication Association, Chicago.

Doerfel, M. L. 1994. The 1992 presidential debates: A new approach to content analysis. Paper: annual meeting of the Speech Communication Association, New Orleans.

Freeman, C. A. and G. A. Barnett 1994. An Alternative Approach to Using Interpretative Theory to Examine Corporate Messages and Organizational Culture. In *Organizational Communication: Emerging Perspectives IV*, L. Thayer and G. A. Barnett, eds. Norwood, NJ: Ablex.

Jang, H. and G. A. Barnett 1994. Cultural Differences in Organizational Communication: A Semantic Network Analysis. *Bulletin de Methodologie Sociologique 44*:31–59.

Johnson, S. C. 1967. Hierarchical Clustering Schemes. *Psychometrika 32*:241–253.

Terra Research and Computing 1994. *The GALILEO computer program.*

Woelfel, J. 1990. *GALILEO Manual.* Troy, NY: Terra Research & Computing.

Woelfel, J. 1996. Attitudes on Nonhierarchical Clusters in Neural Networks. In *Progress in Communication Science: Advances in Persuasion and Attitude Change*, Vol. 13, F. J. Boster and G. A. Barnett, eds.. Norwood, NJ: Ablex.

Words as Actors: A Method for Doing Semantic Network Analysis

Michael Schnegg
University of Cologne
michael.schnegg@uni-koeln.de
and H. Russell Bernard
University of Florida
ufruss@nervm.nerdc.ufl.edu

Introduction

A lot of attention has been paid recently to the development of tools and concepts that allow computer-based analysis of texts. Late-model text management programs help you code text, and many programs let you assign relations between words or embedded codes in a text or set of texts. Some even let you visualize relations among words, codes, and concepts. (Consult Weitzman and Miles, 1995, for a review of these features in text management software.)

All text management software lets you reduce the information in texts. In this paper we describe a method, in use among some students of communication, for analyzing texts as sets of network nodes (see Jang and Barnett 1994; Zegler 1994; and Kleinnijenhuis and Ridder for examples).

The aim of network analysis is to examine relations among a set of actors (people, institutions, countries), not just their covarying characteristics, for a deeper understanding of how actors influence one another. A set of companies, for example,

might have people in common on their boards of directors. This "interlocking directorate," as it's called, would influence how decisions are made in those companies. Rules of exogamy and endogamy set up situations that can also be looked at as interlocking directorates.

The actors in the kind of study we introduce here are words—words that cooccur in texts produced by different informants. We won't be looking for clusters of people who hang out together, but clusters of words that cooccur a lot. Words that are linked across many texts tell us something about shared ideas and shared meanings across texts. In other words, treating words as nodes in a network analysis is a way of finding themes in a set of texts.

To do this you need three things: (1) a set of texts; (2) a program to count the frequencies of words in texts; and (3) a program for analyzing network data. The data we use here for illustration are 21 texts written by students and professionals in Germany about why they went into anthropology. The texts range from half a page to two pages. To count words, we use WORDS2 (reviewed by Bernard in *CAM* 7[3]). Procedures for doing multidimensional scaling (MDS), cluster analysis, and factor analysis are available in ANTHROPAC 4.9 (Borgatti 1994) as well as in all the major statistical packages. We use UCINET IV (Borgatti, Everett, and Freeman 1992) for our analyses because it has all the usual matrix procedures (MDS, factor analysis, etc.) and some special-purpose network measures that we find useful for semantic analysis.

How to Prepare the Data

Information about relationships among elements (in our case, words) is usually stored in a matrix. The matrix we want to build will contain information about the linkage of words in the texts. The links will be defined as the number of cooccurrences of any two words within the same text. Suppose we have the following three texts:

Text A	Text B	Text C
child	music	interest
interest	exotic	exotic
exotic	interest	book
book		

We want to create a matrix that stores the information on how often any two words in these texts cooccur. The words "interest" and "exotic" are mentioned in all three texts. We can define the relationship between those two words as being of strength 3. The pair of words "exotic" and "book" cooccur in two out of the three texts. Their relationship is of strength 2. Other words (like "book" and "music") never occur together. We define their relationship as being of strength 0—that is, there is no relationship between those two words.

The matrix that stores this information looks like this:

	child	interest	exotic	book	music
child	1	1	1	1	0
interest	1	3	3	2	1
exotic	1	3	3	2	1
book	1	2	2	2	0
music	0	1	1	0	1

It's easy to create this kind of matrix when we have just three texts, each containing no more than four words. How can we create a matrix like this when we have 50 texts (like interviews) with hundreds (or thousands) of words each?

There are four steps. (1) Count the word frequencies in *all the texts put together*. (2) Use this word count to generate a "stopfile" (a list of words you want to exclude from later analysis). (3) Using the stopfile, count the frequency of words in *each text individually*. (4) Generate a list that indicates which words belong to which text. This list can then be analyzed with programs like ANTHROPAC or UCINET.

The First Two Steps

We had asked 21 people to write about why they decided to go into anthropology. WORDS2 is a program that reads an ASCII file, counts and sorts the words in the file, and creates a list of the words in descending, ascending, or alphabetic order. We created a single ASCII file of the 21 individual texts and selected the descending order option in WORDS2. Here's the beginning of the output we got:

```
226 and
.
.
61 anthropology
```

In any set of texts, many words with high frequency will be nonsubstantive words like "and." We want to exclude those words from further analysis. We also want to exclude words (no matter how substantive) that occur very few times. (Words that occur once, for example, can only be related to words in the text where they occur.) We decided to exclude words mentioned only twice in the whole corpus so that we could concentrate on stronger ties (that is, ties above strength 2).

We browsed through the output (the list in descending order of use of the words in the whole text corpus) and used a word processor macro to copy those words of no further interest into a separate file. This is the so-called stopfile.

The Next Two Steps

Next, we ran WORDS2 on each of the 21 text files. This produced 21 lists of words. Each list has the same shape as the list above and shows how many times each word of substantive interest occurs in each text. For example, the output from WORDS2 on Julia's text, excluding all the words in the stopfile, looked like this:

```
5 interest
4 travel
4 study
.
.
etc.
```

We combined these 21 files into a single file showing which words were used how many times by each informant. Along the way we inserted each informant's name in front of each line of data, so that the file we wound up with looked like this:

```
Julia           interest        5
Julia           travel          4
Julia           exotic          1
.
.
Michael         interest        2
.
.
Sandra          book            1
.
etc.
```

This is easily done with a macro in any word processor (like WordPerfect or Word).

Finally, we added the following information to the top of the combined data file:

```
DL NR=21, NC=261
FORMAT=EDGELIST2
LABELS EMBEDDED
DATA:
```

DL stands for "data language" (all ANTHROPAC and UCINET files are imported with a data language statement). NC stands for "number of columns" and NR stands for "number of rows." In our data set, we had 21 informants (hence 21 rows of data) and 261 words of interest (hence 261 columns). The statements FORMAT = EDGELIST and LABELS EMBEDDED are crucial here. Placed at the top of the combined data file, ANTHROPAC and UCINET read the data as follows: Julia mentioned the word "interest" five times, the word "travel" four times and so on.

This produces a matrix that looks like this:

	interest	book	exotic	travel	. . .
Julia					
Michael					
Sandra					
.					
.					
.					

There are, of course, 21 rows (one for each informant or text) and the list of words, or labels for the columns in the matrix, extends out to 261 items.

Cleaning the Data

"Cleaning" here means collapsing columns that contain words which contain basically then same content. Singular and plural forms of nouns, for example, and various tense forms of verbs might be combined into single columns. ANTHROPAC and UCINET have a procedure called COLLAPSE. (To collapse column 12 "travel" with column 78 "traveling," for example, you would simply issue the command: col 12, 78.) You can combine as many words as you want and form as many new categories as you like. UCINET and ANTHROPAC will label the new columns "B1," "B2," "B3," etc. You can rename those columns using the spreadsheet built into the programs. We eventually got the number of columns down to 141 theme words.

Converting from 2-Mode to 1-Mode

The matrix we've created is called a 2-mode data set because the elements in the rows (informants) are different from the elements in the columns (words). To explore the relationships among the informants or among the words, we need to convert the data into 1-mode form—either an informant-by-informant matrix or a word-by-word matrix. In this case, the cells in an informant-by-informant matrix would tell us how many words any two people used in common. The higher the entries in the cells, the more words those two people used in common. This is a measure of how similar any pair of texts is.

The cells in a word-by-word matrix tell us in how many texts any two words cooccur —that is, how strongly two words are associated with one another. In our analysis we concentrated on this kind of analysis. To create this matrix, we dichotomized our data (there is a procedure for doing this in ANTHROPAC and UCINET) so that column entries were either 1 or 0 (a word was either used or not used in a text). Then we converted the 2-mode data set into a 1-mode matrix (using the procedure called AFFILIATION in UCINET and CONVERT in ANTHROPAC).

This gave us a matrix with 141 rows and columns. The matrix is symmetric, so that the relationship of word A to word B is equal to the relationship word B has to word A. The entries in the cells of the matrix tell us the relationship between all 9870 (141*140/2) pairs of words where the relationship is defined as common occurrence in different texts.

Reducing the Number of Words

This is, of course, an enormous amount of information. In fact, for our purposes, it's too much information. We gave three coders the 21 texts, told them the question we had asked our informants, and told them to pick out the words most closely associated with the main theme of the research: reasons for going into anthropology. The three coders came up with 49 words and had an intercoder reliability of 79%.[1]

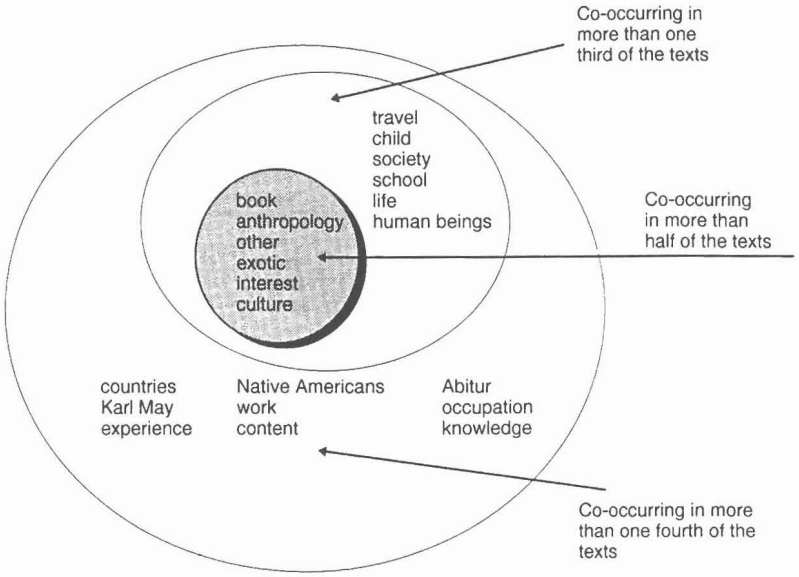
Some Results

Which word indicates the core of motivations of our informants to study anthropology? We address this with the concept of "components" in graph theory. A component is a maximal subgroup within a graph, in which a path of length *n* exists between any two nodes. All the words in the 49-by-49 matrix form a connected graph, so we can look for subgroups that are connected by a path of some specified length. This procedure is available in UCINET.

First, we dichotomize the data by defining a cutoff value. That is, we say that a relationship exists between two words if they cooccur in a given number of texts. We begin with a very strong criterion and loosen the criterion to see how that effects the formation of components (see Figure 1).

In Figure 1, the words in the central circle cooccur in at least half the texts. In the next

Figure 1. Subgroups of a network of words



circle the strength criterion is relaxed to co-occurrence in a third of the texts. By this criterion, words like "child" and "human beings" join the inner circle. The motivations represented by these words are still fairly central for many informants, but not as central as are the shared motivations represented by "book," "other," and so on.

The last circle includes words that cooccur in more than a fourth of the texts. Beyond these circles are words that appear often in the texts—words like "sociology"—but that are not linked to other words in at least a fourth of the texts. These words identify more individual, less commonly shared reasons for the decision to go into anthropology.

An interesting finding in our research is that only one connected component exists. This indicates high consensus across informants. If there were really two quite different subgroups telling us totally different stories about their reasons for going into anthropology, we would expect to find at least two connected components.

We picked the network concept of a component to illustrate the usefulness of this kind of analysis. There are dozens of analyses that can be performed on the kind of 1-mode matrices we've described. We can define subgroups differently. We can measure centrality and power (of people and of words/concepts). The application of matrix analysis to texts is just beginning. We need to study carefully how large texts can be best segmented into smaller units for coding themes and how to understand the relationship among connected themes.

Note

1. The formula to measure the intercoder reliability for nominal codes is $C = 2C_{1,2}/C_1 + C_2$ where C is the intercoder reliability, $C_{1,2}$ the number of words both coders picked in common, C_1 the number of words coder 1 picked in total, and C_2 the number of words coder 2 picked in total (North et al. 1963:49).

References

Borgatti, S. 1994. ANTHROPAC 4.9. Columbia, SC. Analytic Technologies.

Borgatti, S., M. Everett, and L. Freeman. 1992. UCINET IV Version 1.0. Columbia, SC. Analytic Technologies.

Jang, H., and G. Barnett. 1994. Cultural Differences in Organizational Communication: A Semantic Network Analysis. Paper presented at the International Sunbelt Conference, New Orleans.

Kleinnijenhuis, J., and J. Ridder. 1995. Reasoning in Economic Discourse: An Application of the Network Approach to the Dutch Press. In *Text Analysis for the Social Sciences: Methods for Drawing Statistical Inferences from Texts and Transcripts*. Carl Roberts, ed. Hillsdale, NJ: Lawrence Erlbaum Associates.

North, R., O. Holsti, G. Zaninovich, and D. Zinnes. 1963. *Content Analysis. A Handbook with Applications for the Study of International Crisis*. Evanston, IL: Northwestern University Press.

Weitzman, E. A. and M. Miles. 1995. *Computer Programs for Qualitative Data Analysis*. Thousand Oaks, CA: Sage Publications.

Zelger, J. 1994. Cognitive Mapping of Social Structure by GABEK. Paper presented at the European Conference on Social Science Information Needs and Provisions in a Changing Europe. Berlin.

A Hard Day's Work: Measuring Women's Work in Anthropological Research[1]

Kirsten D. Senturia
Pomona College
dkfz34a@prodigy.com

Introduction

Women's physical activity during pregnancy is known to impact their health and that of their offspring (Chamberlain 1984; Saurel-Cubizolles and Kaminski 1986, 1987; Barnes et al. 1991). There are conflicting conclusions about the specific effects of women's work on pregnancy. Some studies cite certain economic and psychological factors as beneficial; others cite posture-related and energy-expenditure factors as detrimental (Alegre et al. 1984; Hytten 1984; Peters et al. 1984; Homer et al. 1990). In my study of women's lives during pregnancy in Albania, I was interested in work as a possible factor in women's health and pregnancy outcomes.

Previous projects on women's work during pregnancy used categorical definitions of work and/or energy-expenditure measurements to evaluate women's activity on the job (i.e., manual and service work vs. administrative and clerical work) (Saurel-Cubizolles and Kaminski 1987; Barnes et al. 1991). Of 25 studies reviewed, only 3 (Berkowitz et al. 1983, Launer et al. 1990, Barnes et al. 1991) accounted for the impact (such as spontaneous abortion and low birth weight) of housework on a woman's pregnancy. The others focused on the impact of waged work on pregnancy. (For the full bibliography, see Senturia 1995.)

While energy-expenditure measures are thorough in assessing output of energy during certain portions of the workday, they are difficult, expensive, and can fail to provide a holistic perspective of a woman's experiences on the job. They may provide correlations between energy output and physical pregnancy complications, but they can not tie in the psychosocial factors of work that affect women's pregnancies. Important factors such as household help and commuting must be considered in developing a complete picture of pregnant women's work lives.

In addition, few studies combine easily quantified survey questionnaires with ethnographic interviews that lend depth to quantitative results. The research reported here shows that certain qualitative issues like concern about unemployment and family economics emerge most clearly through life-history interviews. Whether these issues directly impact pregnancy outcome is less obvious, but their significance in women's *experiences* during pregnancy is quite clear.

The Measures

I developed several measures for use in this