A Natural Sentences Approach to the Computer Coding of Party Manifestos*

by Leonard Ray

Department of Political Science
Binghamton University
P.O. Box 6000
Binghamton NY 13902-6000
USA
lray@binghamton.edu

*Paper prepared for presentation at the 1999 ECPR Joint Workshops
Mannheim Germany.
This paper describes a technique for the computer aided content analysis of political party manifestos. While conceived of in the context of the problem of the analysis of party manifestos, this technique should be applicable to any area where large quantities of machine readable text are to be coded for thematic content. The method described could be used to speed the coding of new manifestos using the existing Comparative Manifesto Project (CMP) coding scheme, and to provide relatively rapid recodes of the existing stock of machine readable party manifestos using new or modified coding schemes. Such recoding could be desirable either to trace the emergence and evolution of novel issues not contained in the current coding scheme, or to extend the set of issues for which directional, as well as saliency measures are available. Applied to other types of political texts such as parliamentary speeches, proposed legislation, campaign materials and media reports, this technique would allow for a comparison of manifesto content with the thematic emphases of other actors, and the analysis of intra as well as inter-party politics.

This research is intended to develop a technique for computerized content analysis. The goal is not to use computers as an aid in the management of information generated through manual coding, but rather to develop algorithms which permit computer coding of texts without having each text read and coded by human researchers. Manual coding is not, however, entirely dispensed with. Manual coding of a subset of text(s) is employed to provide the “seed” information needed to generate dictionaries of words related to specific issues. One of the objectives of this research project is to determine the amount of human coding which is required in order to obtain reasonably accurate computer coding of virgin texts. For this reason, the existing stock of content coded party manifestos are an invaluable resource.

The research project is in the earliest of stages. This paper describes the approach I intend to follow, demonstrates the feasibility of this technique, and indicates the problems and choices which will arise in the application of this method. The reader is invited to suggest any improvements to this method, or indicate any difficulties that I may have overlooked.

Content Analysis

A well common reference on content analysis provides the following definition for this technique. "Content analysis is a research method that uses a set of procedures to make valid inferences from text" (Weber 1990). The term "content analysis" covers a wide range of research methods, as the broad definition presented above indicates. Content analysis, however performed, is essentially a data reduction technique (Weber 1990, 15). A text provides a wealth of nuance and detail. The task of content analysis is to reduce this mountain of information to a much smaller, more manageable, set of indicators relevant to the theoretical concerns of the researcher. This is generally accomplished through the assignment of values on a set of variables to text units (words, sentences, paragraphs, or entire documents.) The distinction sometimes made between quantitative and qualitative content analysis essentially turns on whether the analysis treats the document as a whole, or breaks it into subunits for analysis, (Holsti 1969, 5).

Content analysis requires the identification of units to be coded, and the elaboration of a coding scheme that is to be applied to these units. The codes are supplied by the researcher as a function of the specific theoretical goals of the research enterprise. The categories for this analysis will consist of political issues mentioned by parties. This analysis will employ coding at multiple levels, using the association between specific words and issue categories as a basis for categorizing sentences. The categorization of sentences is then used to produce summary statistics for entire party manifestos.

Content Analysis and Party Positions

Researchers have tried several different approaches to the analysis of party positions using content analysis. Some researchers have attempted to characterize manifestos as a whole, while others have focused on coding smaller elements such as sentences or words, or have analyzed only selected subsets of manifestos.
One approach has been to attempt to code entire documents for their content. The Party Change Project led by Kenneth Janda has employed an expert reader technique in which experts are asked to evaluate the positions taken by political parties on nineteen issues based upon a reading of those sections of the party manifesto relevant to each issue (Janda et al., 1995). This produces a set of ratings of the policy content of the manifesto as a whole. No attempt is made to assign codes to specific sections of the manifesto, and then to aggregate across them. Naturally, this approach involves only the consideration of those sections of the manifesto relevant to the issues under study.

In an attempt to determine whether parties elected to government honor their electoral promises, Rallings (1987) studied the explicit pledges contained within party manifestos or speeches from the throne (in Canada and the UK). These pledges were then compared with subsequent government behavior to determine whether pledges are, in fact, honored. This analysis was restricted to those parts of the manifestos or speeches which clearly commit the party to a specific course of action. Much of the information contained in the texts is not used by this method, but this information is presumed to be relatively unimportant. (Similar analyses have been carried out in the American context. See David, 1971; Pomper and Lederman, 1980.)

The analyses mentioned above all focus on specific selections of manifesto texts. These sections of text are viewed as more meaningful, significant, or relevant than the remainder of the text. Other approaches have considered all of the parts of these texts as equally meaningful. These approaches generally involve dividing the platform or manifesto into subunits, and then assigning codes to as many of these units as possible. Overall evaluations of the content of the text are then derived from an aggregation across these units. The most common approach has involved assigning categories to words. Word counts for each category then provide the data from which measures of the characteristics of the text as a whole are derived.

Namenworth (1969) analyzed the theme of wealth in American party platforms from 1844 to 1964. This analysis was based upon word counts of words associated with the concept of "wealth." The "dictionary" of words which connotes wealth was drawn from a preexisting content analysis dictionary developed by Harold Lasswell. The frequency of such references as a percentage of all references was used as a summary measure of the progressiveness of each party. In a later analysis, Namenwirth and Weber (1987) use the full set of value categories in Lasswell's typology to analyze American party platforms, and "Speeches from the Throne" in the UK. By counting the frequency of use of words associated with each of the 69 content categories, they produce a summary measure of the emphasis in the text on each value. These data are then used to demonstrate the cyclical nature of cultural change.

Joel Paddock (1998) has used the paragraph as the basic unit of meaning in his content analysis of US state party platforms. Each paragraph was assigned a score from 1 to 5 according to how "liberal" or "conservative" the content of the paragraph was. This technique does suppose some homogeneity in the thematic content of paragraphs. Aggregating across paragraphs, he arrived at a summary score for the ideological position of the state party. The validity of this measure of party position is then checked by comparing it with a survey of party chairs.

The Comparative Manifesto Project (CMP) uses the sentence as the basic unit of analysis. The CMP project is an ongoing effort to produce comparable quantitative data on the content of postwar party manifestos from a large number of democratic nations. Begun under the auspices of an ECPR working group, this project is now supported by the Research Unit on Institutions and Social Change at the Science Center - Berlin (See Klingemann, Hofferbert, and Budge 1994 for details.) The CMP approach consists of assigning each sentence (or quasi sentence) to one of 56 policy themes. Manifestos are then characterized by the percentage of all coded sentences in each of these policy categories.

The CMP coding scheme was adopted with a specific theory of political competition in mind. According to saliency theory, parties compete not by taking different positions on a common set of issues, but rather by
stressing different sets of issues. It was thus more important to know how much emphasis parties place on each issue rather than their "position" on each issue. As a result, most of the issue codes do not have a directional component, although a few directional categories were included as a check on saliency theory. (See Budge 1987) It can be argued that the lack of directional tags makes the manifesto data an inappropriate tool for the estimation of party positions. Alternately, saliency theory suggests that the overwhelming majority of tags would be positive, and thus their inclusion would not be particularly illuminating, and the additional information is not worth the cost of obtaining it. If an automated procedure for assigning tags could be deployed, then the cost of assigning directional tags would be low enough to make the exercise worthwhile even if most tags turn out to be positive.

**Computer Aided Content Analysis**

Given their ability to process large volumes of information with relative speed, it is not surprising that computers were used for the analysis of text very shortly after their introduction into the social sciences. Computers soon turned out to be far more useful for data management and analysis than for the actual coding of texts. While a computer can infallibly recognize recurrences of a string of characters, associating that string with a specific meaning (assigning it to a category) is less straightforward. Attempts to implement computerized text analysis reveal the extreme complexity of the process of reading, an act which most academics take for granted. The comprehension of a written text requires the evaluation of an extraordinarily large amount of information.

The task in computerized content analysis is to develop some algorithm(s) which, when applied by the computer, assign elements of a text to the correct content categories. One of the greatest difficulties is that the meaning of a text is not only a function of the individual words that compose the text. There are complex contingencies, where the meaning of a word will be altered by the presence or absence of a second word. The order of the words in a text can also change its meaning dramatically. Context raises a number of impediments to computerized content analysis.

One of the problems of context dependence is homography (See Kelly and Stone 1975). A given ordered set of alphabetic characters may have multiple meanings or senses. For instance, the character string “lead” could refer to the act of giving direction or impulsion to an organization or society, or it could refer to a heavy ductile metal. Homography is, however, not the only obstacle to the computer coding of text. Even if the technical meaning of a term in is quite clear, its "meaning" in terms of the categories used for a particular content analysis may be unclear. (For instance, is the term "interest rates" used as part of a condemnation of the single European currency? Or is it part of a criticism of inflationary government spending?) These two problems do not exhaust all of the difficulties of context dependence. The problem of word order remains even if both of the sources of ambiguity mentioned above have been addressed. If content coding a Labour Party manifesto, we would surely desire an algorithm which could distinguish between “The Labour Party opposes the Conservative Party proposal to join EMU” and "The Conservative Party opposes the Labour Party proposal to join EMU.”

Any computer analysis of text will solve, finesse, or ignore these issues of context dependence. The approach described in this paper finesse the problem of the ambiguity of individual words by assigning words to multiple categories of meaning. It finesse the issue of context dependence by coding entire sentences rather than individual words. This is particularly important in the assignment of valance to party positions. The approach offers no resolution of the problem of word order.

**Computer Coding in Political Science**

A number of research projects have been conducted using computerized coding of text. Most of the research surveyed below deals with the analysis of political parties. However, one of the most productive applications of computer coding has been the analysis of media reports in international relations.
Content analysis has long been used to create data sets tracking the nature and timing of interactions between nations. Two of the most commonly used sources of event data, COPDAB (Azar 1982) and WEIS (McClelland 1976) were both compiled through the manual coding of newspaper articles. Unfortunately, the high cost of human coders has stymied both of these programs. The Kansas Event Data System (KEDS) (Schrodt and Gerner 1994) is a set of computer programs written by a team of researchers at the University of Kansas. KEDS is used to extract event data directly from one line summaries of Reuters news stories. By automating the process of gathering event data, KEDS dramatically lowered the cost of this type of dataset, and allows for almost real time monitoring of interstate relations.

The very specific task and specialized text to which KEDS is applied may partially account for its success. Even within this specialized application, the task of automatically compiling an event dataset is not simple. The computer is first required to parse the summary sentences and extract the subject, verb and object of the sentence. Then the subject and object (the actor, and the nation being acted upon) are assigned codes from a 400 word dictionary of international actors. This dictionary is continually updated as diplomatic and political personnel change. Finally, the verb is coded into one of the WEIS action categories using a dictionary of verbal phrases. In addition to coding, the dictionary of verbs is used to reverse the subject and object of the sentence when the verb is in the passive case. (So in the case of "IRAQ WAS ATTACKED BY US WARPLANES" the program recognizes "WAS ATTACKED" as a passive construction, and the US is coded as the subject of the action, Iraq as the object.) The verb and actor dictionaries were constructed through a long process of trial and error ("training") during which human coders supervised the automated scoring of events and intervened to edit or add dictionary entries when the program miscoded an event. The homography problem described above is resolved on a word by word basis as human researchers edit the verb dictionary to clarify any ambiguities in word meaning which are revealed during the "training" process. For example, "accept" receives one code when followed by "diplomatic credentials," and receives another when followed by "formulation" or "invitation" (Gerner, Schrodt, Francisco, and Weddle 1994, 8). The KEDS program assigns the same code that a human researcher would over 80% of the time. This is comparable to the intercoder reliability between human coders compiling events data (Schrod and Gerner 1994). While an impressively successful solution to the problem of generating event data, KEDS is tailored to the identification of the subjects and objects of sentences. Because of this, it is not suited to the analysis of the thematic content of party manifestos.

Namenwirth and Weber (1987) are concerned with the value content of party manifestos and speeches from the throne. This brings them closer to the substantive concern of the project described in this paper. They deploy a set of computer programs developed as part of the General Inquirer textual analysis package. Their approach builds upon the considerable effort that has gone into developing the General Inquirer software.

The General Inquirer attacks the problem of homography in a very direct fashion. The program contains a routine for "disambiguation" which applies a rigid set of rules to determine the meaning of an ambiguous word given the surrounding words. A number is then appended onto the word to indicate which meaning applies. Subsequent textual analysis is then based upon this new "word" ("lead1", or "lead2", replaces "lead."). This approach involves a painstaking analysis of all appearances of a given word. The word must be assigned to one of a set of specific meanings, and then a rule devised to classify similar occurrences of the word to that meaning. The rules are based upon the context of the sentence, looking specifically for the presence, absence, or location of other words. Fortunately, this painstaking work had been done by others (Kelly and Stone 1975; Weber 1990, 29-30).

Namenwirth and Weber disambigufy only the high frequency words in their texts. Once the text has been disambigufied, it is coded for thematic (value) content. This process consists of searching for each word in the text in a "dictionary" file that indicates the value category with which that particular word is associated. These researchers do not employ a specialized dictionary constructed for their purposes, but rather use the Lasswell Value Dictionary
(LVD) which was elaborated in the 1960's to operationalize Lasswell's theories of political communication. By employing such a general purpose dictionary, they are restricted in the categories into which content may be coded (although they do add a few categories.) However, they gain the ability to compare their results with those of other researchers using the same dictionary.

Each word is associated with at most one of the categories. (Some words are unclassifiable either because they are too rare to be useful, or too frequent to discriminate across categories.) For instance, the category WEALTH TRANSACTIONS is associated with words such as "auction, buy, loan, repaid." Namenwirth and Weber explicitly consider and reject the alternative of allowing words to be associated with more than one category. They argue that such an approach would require the elaboration of a system of weights indicating the strength of the association between each word and each category. Finally, by summing within the categories, and dividing by the total number of classified words, the relative frequencies for each category are obtained. These serve as summary indicators of the content of the text.

Laver and Garry (1998) have proposed a method for the computer coding of party manifestos. Their method is similar to that of Namenwirth and Weber in the use of a dictionary linking specific words to categories. However, the Laver and Garry dictionary is purpose built for the task of coding the policy positions of parties. The dictionary construction technique they describe is inductive, but with a heavy dose of researcher effort. They compute frequency counts of the words in manifestos written by ideologically distant parties, in this case, Labour and the Conservatives in 1992. From these word counts, they identify those words which seem to distinguish the two parties. These words, with their high discriminatory power, are then used in the creation of policy dictionaries.

Laver and Garry use these policy dictionaries to code "virgin" texts by obtaining a word count for the text to be coded. Each time a word from one of their policy dictionaries appears in a text, that score of that text for the policy area increases by one. The result is a count of the frequency with which words from each policy dictionary appear in the text. By dividing these frequencies by the total number of appearances of words from their dictionaries, they obtain relative frequencies that can be used to characterize the content of the text.

The method described below differs from the approach of Laver and Garry in three fundamental ways. Firstly, they create dictionaries of keywords associated with policy areas based upon a comparison of word counts in entire manifestos. Second, they use these keyword dictionaries to assign values for the document as a whole based upon word counts, not sentences. Third, they propose classifying policy emphases as positive, negative, or neutral based upon these same keyword dictionaries.

A Natural Sentences Approach to Content Analysis of Manifestos

I propose the extraction of policy dictionaries from subsets of coded text. This would permit the formalization, and hopefully automation of the dictionary creation process. I then propose applying these policy dictionaries to sentences, or quasi sentences so that the assignment of meaning to ambiguous words can be aided by reference to the surrounding words. Finally, I intend to use these policy dictionaries only to assign sentences to policy themes. I suggest assigning a positive or negative valence to each policy reference using a multiplicative valence algorithm.

The reliance on sentences as the basic unit of meaning in a manifesto brings a number of advantages. The resulting content data will be directly comparable to the CMP data. This allows for rigorous testing of the validity of the approach using the CMP data as a benchmark. The problem of homophony and context dependence is finessed by borrowing information about the policy content of a word from the other words in the sentence. Because of the reliance upon sentences, a time consuming procedure such as disambiguation may not be necessary. Instead, the meanings of all of the words in a sentence are combined to produce a code for that sentence. In this approach, the ambiguity of individual words is recognized, and the meaning of a sentence is derived from the overlap of meaning of its constituent words.
Constructing the Policy Dictionaries

Any automated content analysis procedure requires the elaboration of a set of dictionaries which associate specific strings of characters (AKA: words) with the thematic content categories of interest to the researcher. Dictionary creation is generally described as a deductive process whereby dictionaries are created based upon researcher's expectations before the analysis is conducted. In practice, dictionary compilation is often an iterative procedure where word lists are modified during the course of the analysis. Laver and Garry propose a partially inductive method in which word counts of documents are used to identify potentially useful words. These words are then assigned to policy categories based upon the researcher's intuition.

I propose a formal and inductive method of dictionary creation which would make the creation of dictionaries more easily replicable, and reduces the role of researcher's intuitions or expectations. The method does not fully automate dictionary creation, because it does rely on the human coding of a certain number of "seed" texts.

The first step in this procedure would be the selection of a number of texts to be coded by hand. One of the goals of this project is to determine the quantity of human coding which is needed to obtain reasonably valid data. Each sentence in these "seed" texts will be assigned to a policy category by a human coder.

The second step is to sort all of the sentences by policy theme, and create separate files for each policy theme containing those sentences referring to that theme. Word counts are them performed on these files. These word counts become the basis for the policy dictionaries. The usefulness of the policy specific word counts could be enhanced by automated procedures for lemmatization (Holsti 1969). A raw word count will generate separate tallies for the same word if its form changes with its grammatical use. Thus "benefit" and "benefits" will appear as separate entries. Philip Stone (Stone et al. 1966, 89; Stone 1968, 25-29) describes a simple procedure for reducing such occurrences of a word to its root form. A second automated procedure which would improve the ability of the dictionaries to discriminate between policy areas is the deletion of commonly used words with little substantive content such as articles, forms of the verb to be, etc. A final step in improving the quality of these dictionaries is the deletion of very low frequency words. If a word appears very infrequently, then it is hazardous to assume that it is systematically associated with a given policy area. The deletion of words appearing fewer than four or five times would help to mitigate this source of error.

This procedure will generate a list of words for each policy area, and each word will be associated with a numeric value indicating its frequency in sentences referring to that policy area. Individual words may also appear in multiple policy dictionaries. In order to move to the next step, coding sentences in virgin texts, these frequencies must be transformed into weights. A number of sophisticated statistical techniques could be applied here. (For an early discussion of this problem, see Stone et al. 1966, 154. For a critical view on weighting techniques, see Namenwirth and Weber 1987, 39). However, it is not clear that a sophisticated weighting procedure would appreciably increase the validity of the results. I would opt instead for the simplest procedure, and transform the divide the raw frequencies of each word by the total number of occurrences of that word. The resulting proportion would correspond to the proportion of occurrences of the word which refer to a given policy area.

Assigning Sentences to Policy Areas

Once a set of policy dictionaries listing words and the respective weights has been constructed it can be used to assign codes to new texts which have not been hand coded. Using the dictionary creation procedure described above, some words would be assigned to multiple categories, along with weights reflecting the strength of the association between a word and each issue category. In order to assign codes to sentences, the weights of each word in the sentence must be aggregated, and some formula applied to determine which category the sentence corresponds to.

The first step in the coding procedure is to separate the text to be coded into sentences. By breaking the text at exclamation marks, question
marks, and periods (periods not followed by numbers, to distinguish them from decimals) a file can be broken into natural sentences. Each word is then compared with the dictionary lists, and assigned a weight for each policy area. (If a word is absent from a policy dictionary, then it has a weight of zero.) The weights for each policy code are then summed over the words in the sentence. This provides, for each policy code, an indicator of the likelihood that the sentence refers to that policy area. The sentence is assigned to that policy code with the highest sum of weights. The result should be analogous to the raw CMP data: a set of sentences coded according to their policy content.

Valence: An Additive or Multiplicative Approach

As discussed above, the originators of the Comparative Manifesto Project coded manifestos for the salience of issues. According to saliency theory (Budge 1987) parties compete by emphasizing different issues rather than by taking contrasting positions on the same issues. While there are a number of issue areas coded for direction in the CMP data, the procedure I have described above would produce a pure saliency coding scheme, recognizing the policy area to which a sentence refers, but not the orientation of the party.

There are a number of possible approaches to the identification of the valence of a policy reference. Two will be mentioned here. The first approach involves creating separate policy dictionaries for positive and for negative references to a policy. I refer to this approach as an additive one. The additive approach is based upon the creation of two separate policy dictionaries for each policy area. One dictionary would list words associated with positive references to the policy area, and the other words associated with negative references. The advantage of this approach is its simplicity. Because it can be incorporated into the creation of policy dictionaries, it eliminates any extra steps in the coding procedure. However, there are practical and logical flaws to such a procedure.

A practical disadvantage of such a procedure involves the automated identification of words for inclusion into policy dictionaries. If a statistical technique is applied to determine which words have the greatest power to discriminate across categories, the importance of two types of words risk being downgraded. Any generic valence words which are often employed to indicate opposition or support ("advocates," "rejects," "espouses," "condemns,") may appear in a number of policy categories, and thus will be ranked as words with low discriminatory power. Likewise, words often associated with a specific policy area ("European Union") risk being downgraded if they fail to discriminate between positive and negative policy references. Another serious drawback to the additive approach is the inherently multiplicative nature of directional references. The additive approach would assign a positional value to a sentence based upon the sum of the words which compose the sentence. For example, imagine that a policy dictionary for negative references to the EU includes the words "European Union" "opposes" and "reduce." Applying this dictionary to the phrases "Our party opposes the European Union" or "We advocate the reduction of the powers of the European Union" would yield a negative, and thus correct coding. Applying the dictionary to the phrase "Our party opposes the reduction of the powers of the European Union" would also yield a negative valence code, one that is patently incorrect. I feel that the additive approach is unlikely to yield valid categorizations of sentences, and propose an alternative multiplicative approach where valence tags are assigned to sentences in a second procedure once policy tags have been assigned.

The anomalous outcome above results from the inherently multiplicative nature of directional statements. A negative statement "opposes" combined with another negative statement "reduced" yields a positive (or at least neutral) statement "opposes reductions". Past efforts at content analysis have recognized this aspect of valence by using multiplication as the function to aggregate across the codes in a sentence. (Osgood 1959, p47; Holsti 1969 p.124) I propose a multiplicative algorithm to assign valence tags to sentences. The first step in such a procedure is the identification of a set of "valence words" which generally express positive or negative orientations.¹ In the

¹ A valence dictionary could be built de novo, or could be based on the Stanford Political Dictionary, which contains sets of words associated
example above, the word "reduce" would be coded -1, as well as the word oppose. The valance of the sentence "Our party opposes the reduction of the powers of the European Union" would not be the sum of the two codes (-2) but rather the product (+1). Such an algorithm would permit the reconstitution of the valence categories of the CMP coding scheme, which affords a test of the validity of the results.

Testing and Benchmarking

In order to assess the success of this enterprise, one must be able to measure the resulting data against some standard in order to test its validity. In addition, given the cost involved in human coding, it would be desirable to know with some specificity the payoff obtained from larger "seed" texts. One of the advantages of survey research over content analysis is the applicability of sampling theory to the problem of statistical inference. With computerized content analysis, uncertainty may be introduced both through the sampling of texts to analyze, and through the algorithms used to content code texts.

It seems reasonable to suppose that the validity of the procedure described above will depend upon the quantity of text used as "seed" text for the creation of the dictionaries. If very few texts are hand coded, then it is likely that important words will be omitted from the dictionaries. It is also possible that the "seed" texts use certain words in an idiosyncratic way, which then introduces error into the weights in the dictionaries. For any given amount of seed text, there will also be differences across policy categories in the frequency of reference to that category. It is probable that the policy dictionaries associated with infrequently used content categories will be less accurate.

The validity of the data obtained by this procedure could be tested by comparing computer coding of manifestos with the hand coding from the Comparative Manifesto Project. This would require that a certain number of randomly selected manifestos be omitted from the "seed" texts, and held as test texts. After a single application of the computer coding method, the overall validity of the resulting codes could be checked, along with the relationship between the validity of codes in a content category, and the frequency of that category in the "seed" text. These results would indicate the extent to which validity is a function of the volume of "seed" text.

Of course, this approach is misleading to the extent that valid codes may be easier to obtain for some policy areas than for others. This would be the case if certain policy areas were distinguished by a very specialized vocabulary while other areas used more general and common terms. Ideally one would control for this type of influence on the relationship between quantity of seed text and validity of estimates. An automated content analysis algorithm could be repeatedly applied to the CMP documents in a procedure akin to bootstrapping. By repeating the procedure several times, each time with different sets of seed texts, and different volumes of seed text, one could determine the amount of human coding required for a given level of validity.

This approach to testing could also help to establish the distribution of estimates over repeated iterations of the algorithm. With some notion of the mean and variance of the computer coded output, a rough rule of thumb regarding "significance" could be constructed. Given a rough idea of the variability of output for a given amount of seed text, rules of thumb for descriptive inference could be established. One could, for example, establish

with positive and negative affect. (Stone et al. 1966 p. 189) In the Stanford Political Dictionary, words are assigned a magnitude as well as a valance. Thus a strongly positive word such as "earnestly" may be assigned a +3, while a less strongly negative one such as "reduce" would receive a +2. If the word "oppose" was also considered moderately negative (coded -2), then the example sentence is coded +4 due to the conjunction of two moderately negative modifiers. This approach, while suggestive, could not be tested by comparison to the CMP data.
the conditions under which two computer coded texts could be said to differ in the frequency of references to a given policy area.

**Conclusion**

The technological advances in computing power optical character recognition, and data storage mean that social scientists now have at their disposal the technical means for very sophisticated types of automated content analysis. Unfortunately, in political science the theory and methodology of computer coding has not kept pace with the technology, and researchers often seem to invent procedures de novo which reflect the idiosyncrasies of their specific applications. The intention of this research project is to design, implement, and most importantly test a relatively general procedure for automating content analysis. The Comparative Manifesto Project has produced an ideal set of data which could be used to elaborate and test the procedure described above. These data permit the quantitative evaluation of the overall validity of this procedure, and an optimization of the procedure by determining the method which yields the greatest validity. The goal of content analysis is making inferences from texts. This research project should allow future scholars to apply a technique for content analysis which has known characteristics in terms of validity. As a result, they will be able to say with some confidence which inferences are supported by their data, and which may be artifacts of the procedure. If this goal can be attained, content analysis will become a much more useful tool not only for the study of political parties, but also for social science in general.
Bibliography


