STATISTICAL SEMANTICS: HOW CAN A COMPUTER USE WHAT PEOPLE NAME THINGS TO GUESS WHAT THINGS PEOPLE MEAN WHEN THEY NAME THINGS?

George W. Furnas, Louis M. Gomez,
Thomas K. Landauer, Susan T. Dumais

Bell Laboratories, Murray Hill, N.J.

The descriptors or categories assigned to entries in an information system form the basis of most retrieval mechanisms (e.g., menu or key word). These descriptors are the primary means of communication between system designers and end users. In this paper we analyze some of the factors which influence this communication link. Our goal is to uncover some psychological principles that will help us to understand naming and describing behavior and thus improve the communication between designers and users.

In traditional communication (e.g., conversation) the communicator can accommodate to different listeners, both by shifting perspective and by attending to explicit feedback from the listener. In describing items in a data base, however, system designers are at a disadvantage in that they do not usually get explicit, immediate, and continuous feedback from users. Knowing how people describe common objects and shift their descriptions for audiences of different levels of sophistication may help designers build systems whose information is accessible to the widest possible audience.

We will discuss selected results of four psychological studies in which subjects were asked to describe or name objects such that other people would be able to easily find or identify the target object in a collection of other objects.

The description task was carried out under one of two general types of instructions. People were told to describe objects using normal connected discourse, or they were told to describe objects with key words. Key word descriptions were single or multiple content word responses to target objects. In all cases, subjects were instructed to attempt to convey the most salient aspects of each target. The four studies were as follows.

(1) Forty-eight secretarial and high school students (with typing, but no computer experience) were given a sample manuscript with author's corrections. They were asked to prepare a typed list of instructions for someone else who was actually going to make the changes but did not have the author's marks. This technique allowed us to observe the natural or spontaneous names for common text editing operations used by non-programmers. Here we were especially interested in consistency of language use. Do different people use the same lexical units to describe the same text editing operation? Does the same person consistently give the same name to the same operation? In addition, we also studied how the lexical items used to specify text alterations varied as a function of the size (i.e., blanks, characters, words, lines, paragraphs) and type (i.e., insert, delete, replace, move, transpose) of text unit being changed.

(2) Three hundred thirty-seven college students gave short statements to specify verbal objects. They were given a list of common items like "Newsweek", "Empire State Building", and "motorcycle" and asked to describe each so that another student or (in other cases) a hypothetical computer would respond with the target word. The purpose of this study was to understand how people use language to get other people to think of some specific object. In addition, we wanted to explore the possibility that target specification language would change when it was directed toward a computer.

(3) In the third study, twenty-four subjects used a key word description technique to describe a collection of cooking recipes. Subjects were asked to nominate index entries for a collection of recipes. Expert, intermediate, and novice level cooks participated. Half of each skill group was instructed to tailor their index words to be particularly appropriate for use by expert cooks; the other half was asked to make their descriptors appropriate for novice users. These results help us to understand how people with different levels of expertise describe objects and shift their descriptions for different

audiences. Experts in the appropriate content area normally provide information access keys for all potential audiences. This experiment provides an initial test of the general utility of this approach to information indexing.

(4) In the fourth study, thirty subjects generated several single word or short phrase descriptors which categorized a target object into successively higher (superordinate) categorical levels. Subjects were instructed to complete "ISA" sentences (e.g., An apple is a fruit.) such that with each completion the last named category (e.g., fruit) was categorized under a higher superordinate (e.g., A fruit is a food.) These data can be used to construct menu retrieval systems based on graph structures corresponding to user conceptions of the relations among objects. Most objects can be categorized in many ways. Many menu systems, however, only allow their users one way to reach an object. The approach taken in this experiment could suggest ways to build menu structures with several psychologically salient routes to each object.

These studies investigated a wide range of psychological problems in the general area of human computer interaction. The results of all four studies, however, have led us to some general insights into human naming and describing behavior.

The most striking result from the verbal production data was the great diversity in people's descriptions of even the most common objects. The average likelihood of any two people using the same main content word in their descriptions of the same object ranged from about .07 to .18 for the variety of stimulus domains studied. These values have important implications for computer systems which depend on lexical agreement between users and system designers. Information systems like this require that two people, namely a user and a designer, to agree before the system can be used.

Many current information systems require users to learn a special vocabulary to retrieve desired information objects. Furthermore, it is often the case that each information object can be retrieved by only one word. This approach can work well if the vocabularies are small or the users highly trained. Unfortunately, vast numbers of users will soon be trying to use large information retrieval systems, without extensive training. They will not know the right vocabulary, even for things they know the system can do. Their only recourse is to guess or give up. The probabilities just mentioned suggest that if they guess, they will fail to hit upon the word some particular other person, i.e., the designer, chose 80% - 90% of the time. It is worth noting that providing system documentation is not an alternative solution, since the same problem arises in finding appropriate entry points into the documentation. What terms should the user look up in the index? Documentation may in fact be one of the most common and frustrating places for users to encounter this obstacle.

What can be done? Using data from our experiments, several different approaches were explored to give users a better chance of being understood by a computer system. It was found that roughly a 2:1 improvement over the worst case (i.e., when the system stores only one word that the user is expected to match) resulted if each object (e.g., a command or an information item) was given the most frequently used name observed in the user population. Allowing the system several names for each or allowing the user several independent guesses increased the chance of success almost in proportion to the number of names or guesses (for the first few of each).

The best approach, however, involves taking a different perspective on the problem. In the analyses just mentioned the focus was on what the system brings to the interaction. We began with the system's set of objects and tried to find the best name for each object (an orientation likely to ignore to the human diversity in terminology). The alternative approach focuses on what the user brings to the interaction. It begins with all the words that people have been observed to use. The goal in this general effort is never to allow the information retrieval system to let a user's word go by without at least hazarding a guess as to their intent based on the data collected. In theory the best way to make a single guess given a word the user says is to choose the most common object that the word has meant in the past. Our data suggest that this general strategy can succeed upwards of 40% - 60% of the time, even when the user and the computer are restricted to one attempt each. Additional improvements are to be expected if either participant is allowed a few guesses. The data need not be collected all ahead of time. It is possible to accumulate these data continuously during use, adapting in real time to the vocabulary patterns of the user community.

The lack of language agreement between designer and user arises from the diversity of language use and the imprecision of its application. Diversity of language use limits successful communication between

designer and user for at least two reasons. First, there are many ways to say the same thing about an object. Variations of this sort range from simple lexical and grammatical differences to deeper semantic synonymy. Second, there are many different things to say about an object, each focusing on different aspects.

There are several extant analytic methods by which to tease out the contributions of these two sources of variety in people's descriptions. One approach is to measure the extent to which any two terms are applied to the same set of objects. A second method explores the "meanings" of the terms, in the vernacular sense. Terms may be thought of as similar not only if they are used to refer to the same things but also if the same semantic predicates are applied to them. Thus, we have asked people to classify terms with regard to yet higher level (more abstract) categories. Evidence for synonymy emerges to the extent the terms are put into the same abstract categories.

Both sorts of data, degree of common reference and degree of co-classification, can be used as measures of proximity between terms. The proximity data can then be analyzed by a variety of techniques (e.g. multidimensional scaling and factor analysis) to reveal a latent structure in which terms showing synonymous behavior appear close together, and those acting more independently appear far apart.

Imprecision in language usage can result in limited success in guessing the user's intention, even when the system recognizes the user's input word. It results from the fact that people use the same words for many different objects. One reason is that some objects are very similar to one another, so the same words are appropriate to them. Thus, as a set of objects get more similar, this imprecise reference problem becomes more severe. We have confirmed this conjecture. Pairs of recipes and common objects independently judged to be similar tend to have the same names attributed to them.

We hope the practical implication of all these analyses is clear. Given the alternative ways in which objects can accurately be described, designers of information systems should not accept their own descriptions as optimal, or even sufficient. The methods and results we describe will, hopefully, suggest ways in which systems can use data from people to accommodate the diversity of people's natural descriptions.