ABSTRACT

Title of Dissertation: TAG CLOUDS: HOW FORMAT AND CATEGORICAL STRUCTURE AFFECT CATEGORIZATION JUDGMENT

Anna Walkyria Rivadeneira Cortez,
Doctor of Philosophy, 2008

Dissertation Directed By: Professor Kent L. Norman
Department of Psychology

This paper examines how category judgments are influenced by categorical structure and the formatting of *tag clouds*. Despite the enormous research on categorization, little research has been directed at investigating whether one person can recognize another's categorical structure. A novel approach to measure similarity and categorical structure is proposed. This approach involves the use of latent semantic analyses to compute semantic distances between category exemplars. The empirical domain will be tag clouds, a new development in social computing that provides a particularly useful paradigm for investigating how people identify the categorical structures of others. Three experiments examine how categorical structure and different formatting styles used in tag clouds might affect categorization. Findings reveal that categorization judgments are influenced by categorical structure and tighter structures result in higher accuracy. Format variables such as font size and sorting order were also found to influence accuracy. Future experimental directions are detailed.

TAG CLOUDS: HOW FORMAT AND CATEGORICAL STRUCTURE AFFECT CATEGORIZATION JUDGMENT

By

Anna Walkyria Rivadeneira Cortez

Dissertation submitted to the Faculty of the Graduate School of the
University of Maryland, College Park, in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
2008

Advisory Committee:
Professor Kent L. Norman, Chair
Professor Benjamin B. Bederson
Professor Michael R. Dougherty
Professor Paul J. Hanges
Professor Thomas S. Wallsten

Acknowledgements

The success of this dissertation would not have been possible without the guidance and support of Kent Norman. Michael Dougherty – who was generous with his time and resources -- was instrumental in its completion. I also thank Benjamin Bederson, Paul Hanges, and Thomas Wallsten for providing thoughtful feedback. The support of my colleagues at the Collaborative User Experience Research Group at the IBM Watson Research Center fostered the initial inspiration for this dissertation.

Table of Contents

List of Tables

List of Figures

Chapter 1: Introduction

Cognitive and developmental psychology research shows that humans exhibit the ability to form categories and concepts early in their development (Daehler, Lonardo, & Bukatko, 1979; Mervis, & Crisafi, 1982; Smith, 1981). Categorization theory has undergone a theoretic progression from a classical view where members of a category share necessary and sufficient attributes to a probabilistic view where members of a category share a certain level of overall similarity (Smith & Medin, 1981; Medin, 1989; Komatsu, 1992). This shift originated with experiments that showed that variables such as typicality or structure influenced category judgment in a manner that the classical view could not explain. Despite the enormous amount of research on categorization, little research has been directed at investigating whether one person can recognize another's categorical structure. The issue is important at both theoretical and applied levels. At the theoretical level, this research will suggest a novel approach to measure similarity and categorical structure and how this metric can predict categorization judgments. At a practical level, this research can be used in addressing social issues regarding how people categorize the world around them. It can also facilitate the growth of what is now called "social computing," in which the usage of software and technology facilitates social interaction and communication.

Probabilistic models of categorization theory are founded on the use of similarity as a measure of category membership. The use of this metric has raised many objections. It has been argued that similarity is too flexible and that there is no consensus on its definition. I am proposing an approach to measure similarity which I believe is capable of withstanding such objections. This approach is quite practical

and adaptable. Similarity between words can be computed by means of a latent semantic analysis (LSA) on a text corpus. It is practical as a consequence of the wide availability of software packages that perform these analyses. It is adaptable inasmuch that the corpus can be in any language and in any domain[1].

While my research is aimed at broadening theory in general, the application domain will be tag clouds. Tag clouds are a new development in social computing that provide a particularly useful paradigm for investigating how people identify the categorical structures of others and form mental representations of other people's interests and expertise. Social and collaborative software sites use a terminology for book marking called "tagging". Tagging is the process by which a user assigns metadata to a document in the form of keywords – *tags*. This mechanism allows a user to organize content for future navigation, filtering or search (Golder & Huberman, 2006). Information is categorized by "tags" or keywords and can be visualized using "tag clouds". Tag clouds use attributes of the text – such as size, weight, or color – to represent features of the associated terms. For example, the prevalence of a term in the set could be represented by its size. Figure 1 presents an example tag cloud for information represented in this paper. Note that the terms or tags in the figure are concepts prevalent in this paper. Tag clouds are increasingly common on social software sites as links to the tagged websites and can serve as tables of contents. Tag clouds can represent the terms assigned by a single person. Just as a table of contents can give a reader the gist of what a book is about, tag clouds can provide an impression of that person and his or her interests and expertise.

[1] For example, in order to measure similarity in the medical domain, one could use the collection of medical journals as the input corpus.

Figure 1

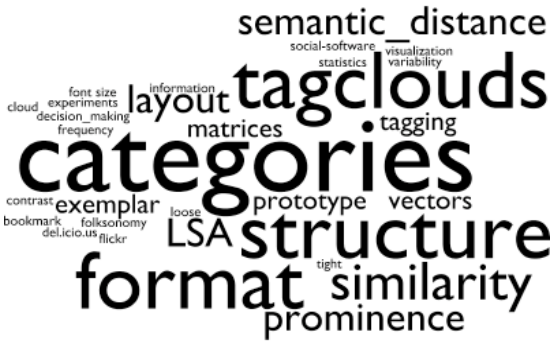

Figure 1. Hypothetical tag cloud representing keywords related to this paper.

This paper will examine how category judgments are influenced by categorical structure and tag cloud formatting. The empirical component of this research is twofold: study people's ability to verify a given category in a tag cloud and to discover (retrieve) one or several categories in a tag cloud. The experimental paradigm will incorporate categorization theory in addition to the various dimensions used to construct a tag cloud into the manipulations. I will propose a new measure to assess the variability of a structure and how it can predict category retrieval and verification. I will be using Posner and Keele's (1968) terminology, in which *tight* structures are categories with high degrees of within-category associations and *loose* structures are categories with low degrees of within-category associations. As an example, a tag cloud may contain categories that have either tight or loose structures. The category "music" could be a tight category, it could be composed of websites related to and tagged with: "guitar", "concert", "jazz", "songs", etc. The category "vacations" could be a loose category, it could be composed of websites related to and tagged with: "hotel", "weather", "airport", "Paris", etc. Categories with loose structure might be harder to identify. Boundaries for these types of categories may be

fuzzier than for ones with a tighter structure, resulting in misclassification. I will examine how different formatting styles used in tag cloud visualizations might affect category retrieval and verification. More specifically, I will investigate tag cloud layout and prominence.

In the next section, I will review the classical and probabilistic views of categorization and address the objections used against the use of similarity as a construct to determine category membership. I will introduce the use of LSA to compute a measure for similarity and categorical structure and end the section by explaining how the experimental paradigm I am employing can be used to investigate category research. I will then present three experiments that examine how manipulations of categorical structure and format affect judgments of category membership. I will conclude by discussing the use of semantic distance in categorization theory and provide a set of guidelines on how to visually present tags so that the information they represent can be accurately transmitted.

Categorization

The process of categorization involves regarding different entities as members of an equivalence class. It is assumed that members of the same class share one or more unobserved properties and equivalence classes can be systematically sorted into hierarchical levels (Mervis & Rosch, 1981). Categorization is a process that is conducive to organization of knowledge by enabling the formation of taxonomies when the levels are related to each other by class inclusion (Sloutsky, 2003). Categorization therefore is not equivalent to mere grouping of entities, but rather may offer a coherent structure that can be used by people to make inferences.

The theory of categorization has its origins in philosophy with Aristotle and in experimental psychology with Hull's 1920 monograph on concept attainment (as cited by Smith & Medin, 1981). This first inception into categorization theory is referred to as the classical view (Smith & Medin, 1981; Medin, 1989). It states that categories are determinately created by necessary and sufficient conditions (attributes) for membership. There are three main problems with this view: First, a comprehensive and exhaustive list of attributes is generally impossible to define. Research that asked experts to give a complete list of attributes that define a category showed there was considerable disagreement as to what exactly those necessary and characteristics attributes were (Murphy & Wright, 1984; Tanaka & Taylor, 1991). Second, category members are nonequivalent. As per the classical view, if judgments of category membership are based on a list of attributes that denote the specific category, then any member should be cognitively equivalent to any other member of the same category (Mervis & Rosch, 1981). However, members of categories exhibit degrees of variability amongst themselves. For example a monkey is more easily classified as a mammal than a whale. Berlin and Kay (1969) obtained experimental data from several languages using native speakers. They extracted the basic color terms of a language and then mapped these terms to a chart of fully saturated color chips. They found that the number of color terms and boundaries of color categories vary widely across cultures and languages. However, they were able to discover a very limited and universal set of color terms in all languages that they studied. Third, there are unclear cases of category membership. Research has shown that there are certain cases when a stimulus is difficult to assign to a specific category. Basic-level

categories are easier to identify than sub- or super-ordinate categories. Children are quite capable of identifying basic-level categories, while developmental differences occur once tasks involving sub- or super-ordinate categories are used (Rosch, Mervis, Gray, Johnson & Boyes-Braem, 1976).

This evidence suggested that categories have ill defined or fuzzy definitions leading to a new perspective of categorization theory. The probabilistic view states that judgments of category membership are based on family resemblance. There are several models developed within this view: prototype models, exemplar models and decision-bound models.

In prototype models (Posner & Keele, 1968; Reed, 1972; Rosch & Mervis, 1975), category membership is assessed by the similarity between the probe and an ideal element – the prototype -- that represents the category. In prototype theory, similarity is assessed by the representation of an additive combination of cues. Reed (1972) suggested that category membership is assigned by calculating the distance[2] from the prototype. Participants were asked to assign faces into categories that were defined by sets of exemplars and not by logical rules. The results showed that the distance from the prototype best predicted the data. Evidence favorable of prototype models have shown in learning studies that a prototypic stimulus pattern that is new may be correctly categorized on a subsequent test with probability as high or higher than old patterns (Posner & Keele, 1968). There are several problems with prototype

[2] Distance is not a physical distance, but a mathematical representation that satisfies the following axioms:

(i) Minimality: $\delta(a,b) \geq \delta(a,a) = 0$

(ii) Symmetry: $\delta(a,b) \geq \delta(b,a)$

(iii) The triangle Inequality: $\delta(a,b) + \delta(b,c) \geq \delta(a,c)$

.

models. First, prototype theory states that the classification of a stimulus will remain constant in different contexts. However, there is evidence that the surrounding context can affect judgments of similarity. One of the many demonstrations of these context effects is an experiment in which the subjects were presented with four countries that naturally formed two clusters (Tversky, 1977). Properties that are useful for categorization exert greater influence on similarity judgments. When one of the countries was substituted for another, the clustering of the countries changed. An example of this sort would be the quadruple formed by "Iran, Israel, Syria and England". Iran and Syria are judged similar based on religion (Muslim countries), while Israel and England are judged to be similar (non-Muslim countries). In a different context, where only one country of the quadruple is changed the similarity between two stimuli is considered differently. For example, if Iran is substituted with France, Israel and England are no longer judged similar but rather new natural categories are formed (France and England – European countries, Israel and Syria – Middle Eastern countries). This in turn had an effect on judged similarity of Israel and England where they were judged to be more similar when presented in the first group than in the second group. Second, prototype theory does not consider that the size, distribution or variance of the category will affect how similarity judgments are made (Medin & Shaffer, 1978; Homa & Cultice, 1984). A very variable category is supposed to be represented exclusively by its prototype. As an example, in the bird category, a robin could be described as the prototype. An ostrich would be a stimulus that belongs to the bird category but is quite different than a robin. Third, prototype theory does not do well with fuzzy set theory. Osherson and Smith (1981) describe

how prototype theory fails to demonstrate several of the consequences implied by Zadeh's fuzzy set theory (Zadeh, 1965). Critics have argued that the mental computations that would be needed to form prototypes are quite difficult (Smith & Medin, 1981). Estes (1986) proposed a general array model for categorization. Within the framework of this model the author stated that if the exemplars of a category can be represented in memory in terms of features, then the computations required to generate a prototype are no more difficult than those needed to estimate feature probabilities.

Exemplar models predict that classifications are made based on the means of examples within a category. Category membership is determined by the retrieval of exemplar information, where retrieval is a global match between the stimulus and memory representation (Smith & Medin, 1981). In the exemplar theory, similarity is assessed by a multiplicative combination of cues. This multiplicative representation gives a heavier weight than prototype theory to the absence of necessary cues (Medin & Shaffer, 1978). Exemplar models fair better than prototype models. They are context sensitive and allow predictions on partial information. The context model developed by Medin and Shaffer uses multiple binary dimensions. Exemplars in this model contain information about the dimensions and the context of the stimulus. Nosofsky (1986) developed an exemplar model with multiple continuous dimensions.

Decision-bound models assume that people perceive category membership with some degree of error. These models are based on Ashby and Maddox's (1993) multidimensional version of signal detection theory called general recognition theory (GRT). An exemplar in the context model is represented as a point in a

multidimensional space, while an exemplar in GRT is represented as a multivariate distribution. A category is a probabilistic mixture of multivariate distributions. A perceptual space in GRT is delimited by a boundary; persons assign a stimulus to a category depending in which region the stimulus falls. The shape of this boundary changes according to different assumptions. For example, responses with no noise and categories with a normal distribution set the shape of the boundary as a quadratic function and a deterministic model is predicted. Ashby and Maddox listed different assumptions that described when GRT and other category models predict the same behavior. If the distribution is logistic and similarity is a weighted distance in Euclidean space, GRT is a version of a probabilistic prototype model. The most general type of GRT models is a special case of weighted additive exemplar models.

I believe probabilistic views should be considered for theories of categorization. It is clear that categories have fuzzy boundaries and that instances of a class may vary in their degree of features associated with the class. Category membership should be considered a function of intra-class similarity.

Similarity

The development of categorization is based on perceptual and attentional mechanisms capable of detecting similarities in the environment. Similarity can operate as the premise by which people classify objects, form concepts, and make generalizations (Tversky, 1977). The construct of similarity is used because of its central role in categorization theory. One of the Gestalt principles of perceptual organization states that similar things will tend to be grouped together. Goldstone (1994a) stated that similarity is an indirect instrument used by psychologists to

examine both the structure of mental entities and the processes that operate on these entities. Two important theories of categorization -- prototype theory and exemplar theory -- assume that people categorize based on the similarity between the object to be categorized and the categories' reference class (Goldstone, 1994b). As per prototype theories, an item is classified in reference class A and not B if it is more *similar* to A's best representation for its class (A's prototype) than it is to B's (Posner & Keele, 1968; Reed, 1972; Rosch & Mervis, 1975). As per exemplar theories, an item is classified in reference class A and not B if it is more *similar* to all items that belong to class A than it is to those that belong to class B (Medin & Schaffer, 1978; Nosofsky, 1986). Among the several assumptions associated with Medin and Schaffer's context model, three deal specifically with similarity:

> […]
> 2. The probability of classifying exemplar i into category j is an increasing function of the similarity of exemplar i to stored category j exemplars and a decreasing function of the similarity of exemplar i to stored exemplars associated with alternative categories.
> […]
> 4. The similarity of two cues along a dimension can be represented by a similarity parameter whose value can range between 0 and 1.
> 5. The various cue dimensions compromising stimuli in some context are combined in an interactive, specifically multiplicative, manner to determine the overall similarity of two stimuli. The interactive rule has the potential to represent the effects of necessary features without the theory committing itself to the idea that category membership is defined in terms of singly necessary and jointly sufficient features. (p. 211-212)

The use of similarity in cognition has been criticized. Similarity needs to be specified by attributes, attribute relations, and higher order relations. The use of

similarity as the source of categorization predicts quite well how items will be categorized. However, justifying a specific classification scheme based on similarity is relatively difficult. Proponents of theory-based categorization believe that it is deliberative and capable of such justifications. Proponents of similarity-based categorization believe that similarity is not limited to sensory properties and it does not require sophisticated knowledge such as that required for theory-based categorization (Goldstone, 1994b).

Rips (1989) presented an experiment that showed that classification based on similarity may differ from classification based on rules. Subjects were presented with objects that had three dimensions where two of them did not vary and one was quite variable. An example of such tasks was to categorize a silver 3'-diameter circle and the possible categories were pizzas and quarters. Participants classified the object in the pizza class (variable category) but judged the objects to be more similar to the quarters class (fixed category). Rips (1989) showed a judgment dissociation where categorization decisions favored one category while similarity decisions favored another. Smith and Sloman (1994) tried to replicate this finding and concluded that similarity-based categorization is performed in an automatic manner, while rule-based classification is performed in a deliberative manner. The replication worked for trials that used a verbal protocol and that did not hint of time pressure. The authors suggested that participants employed rule-based classification when they felt encouraged to explain their reasons for categorization. When there was no verbal protocol, participants employed similarity-based classification. Goldstone (1994b) uses Smith and Sloman's (1994) study as evidence against the claim that

categorization requires sophisticated processes. In the study, participants were aware of the correct categorization rule but still relied on similarity in order to make categorization judgments. It may be premature to support one view over the other. Similarity may well have sufficient power to ground many categorizations.

Most similarity-based models assume that similarity can be specified as a distance metric in a multidimensional space. According to these models, any object can be represented by its coordinates in a similarity space. The closer the two objects are in this space, the more similar they are. Various multidimensional scaling procedures have been developed on the basis of this idea (e.g. Torgerson, 1952; Shepard, 1987). These methods use a matrix of pairwise distances between the objects to represent the objects in a space defined by a limited number of dimensions. Tversky (1977) has demonstrated that this metric assumption is sometimes violated. The metric assumption requires that the following three axioms be satisfied: minimality, symmetry and triangle inequality. The minimality axiom states that an object is most similar to itself than to other objects. However, there are occasions when an object is considered more similar to other objects than to itself. The symmetry axiom states that if *a* is similar to *b*, then *b* is similar to *a*. This directionality of the comparison plays an important role in the symmetry assumption. Similarity may depend on whether the object is the subject or referent of the comparison. Usually less prominent objects are considered more similar to more prominent objects than vice versa. Mervis and Rosch (1981) mention the violation of this axiom but use the term "representativeness" instead of "prominence" (p.97). Another example of a violation of the symmetry axiom can be found in a

psycholinguistic study by Whitten, Suter and Frank (1979), which was intended to investigate synonymy norms. A strong directional effect was found as perceived synonymy was significantly affected by encoding order of the rated noun pairs. The triangle inequality axiom states that if *a* is similar to *b* and *b* is similar to *c*, then *a* should be similar to *c*. For example, a comparison between Jamaica and Cuba would result in considering the pair of countries similar based on their location. A comparison between Cuba and China would result in considering this new pair of countries similar based on their politics. However, it would be difficult to justify, based on the triangle inequality, that Jamaica and China are similar. Tversky proposed a feature matching model[3] that does not need a metric assumption. The matching is a function of the similarities and dissimilarities of two objects.

Goodman (1972) argued that similarity is too flexible and vague, that it requires a frame of reference. Tversky and Kahneman (1996) have argued that it is not necessary to define similarity because it can be assessed experimentally. The assessment of similarity can follow the methodology used to measure psychophysical qualities such as loudness, which are defined experimentally in terms of respondents' judgments. Medin, Goldstone and Gentner (1991) agreed with the view of similarity as a flexible construct; but argued that the similarity comparison process could systematically fix similarity by setting constraints on similarity. Alignment in the comparison process could provide a reference for similarity judgments. Goldstone's (1994a) study argued, by means of empirical data and computational modeling, for the inclusion of structural alignment in a theory of similarity. Similarity judgments

[3] Because the feature matching function is a contrasting function, this model is also known as the contrast model.

require alignment of the pairs of the compared scenes. These individual alignments are influenced by the overall pattern of other emerging alignments. The comparison process may be able to predict the directionality of similarity judgments, with the subject of the comparison assumed to be the more salient item (Tversky, 1977). If directionality is stated in the instructions -- as in "*a* is similar to *b*" -- the properties of *b* (the subject of the comparison) are given more weight than those of *a* (the referent). Ambiguous features could be clarified during a comparison process (Medin et al., 1991); participants assign more weight to common features and less weight to distinctive features in similarity judgments (Gati & Tversky, 1984; Tversky, 1977). Constraints on similarity could come from the context of comparison. Certain contexts increase the diagnostic value of particular features and affect the judgments of similarity more than in other contexts. Another study performed by Tversky (1977) demonstrated that the surrounding context affected the judgments of similarity between two objects[4]. Medin et al. (1993) performed a study that investigated ambiguity and context-specific features. Participants were asked to list common features between two stimuli. In one condition, stimulus B was compared with stimulus A alone and in another condition, B was compared with stimulus C alone. The authors proposed that activated properties of one entity in a comparison would be evaluated as candidate properties of the other entity. Stimulus B was construed so that its properties would be ambiguous, one property could be construed as exclusive of A and incompatible with C and another property as exclusive of C and incompatible with A. The results showed that participants interpreted B's ambiguous properties depending on the context of the comparison. The fact that similarity judgments can

[4] A more elaborate summary of this study was previously presented in this paper. See p. 15

vary does not mean that similarity is unreliable as Goodman (1972) suggests. Medin et al. (1993) argue that entities participating in a comparison process jointly constrain one another and determine the outcome of a similarity judgment. Their paper -- "Respects for similarity" -- list the following statements that reiterate their position:

1. Similarity comparisons involve mutually constraining property instantiation and interpretations.
2. Similarity comparisons are informative and may be directional.
3. The respects associated with similarity assessments are influenced by the comparison context.
4. Similarity comparisons involve alignment driven by global constraint satisfaction.
5. The contribution of a match to similarity comparisons depends on the overall pattern of correspondences between entities. (p. 272)

I believe that two distinct processes can be used to categorize common objects: rule-based and similarity-based categorization. The former process is applied under conditions that require elaborate judgment and the latter process is a heuristic type approach. However, this should not diminish the relevance of the construct of similarity. On the contrary, this construct should be further investigated because of this tendency to use similarity in certain situations. Reliable measures of similarity should be constructed. Further, it has been proposed that similarity is a construct not exclusive to categorization. Other areas of cognition might be influenced by similarity.

People make judgments of the likelihood that an object belongs to a class by assessing the similarity between that object and an exemplar from that class (Kahneman and Tversky, 1972). In likelihood judgments, similarity appears to perform the function of a heuristic. This representativeness heuristic can lead to

departures from normative expectations that subjective probabilities are predicted to obey. The representativeness heuristic can produce biased results, because the factors that affect similarity do not necessarily affect likelihood. Consequences of the use of the representativeness heuristic are: (a) Biases in considering the effect of sample size. People assess the likelihood of a sample result by its similarity to the corresponding parameter disregarding sampling theory (Tversky & Kahneman, 1974). (b) Misconceptions of chance. People judge the sequence of coin tosses H-T-H-T-T-H to be more likely than H-H-H-T-T-T because the former appears more random (Tversky & Kahneman, 1974). (c) Insensitivity to prior probability of outcomes. In problems such as the ones that present sets composed of different proportions of lawyers and engineers, when subjects are asked to predict a person's occupation based on a description of the person, the person is assigned to the occupation for which the match between personal description and occupation stereotype is obtained disregarding proportion information (Kahneman & Tversky, 1973). (d) Overestimation of concurring events. This bias is commonly known as the conjunction fallacy that occurs in Linda-like problems, where people believe that the likelihood of that two specific conditions is greater than one general one because the two conditions appear more representative of Linda's description, even though it is mathematically less likely (Tversky & Kahneman, 1983).

Medin, Goldstone and Markman (1995) have suggested that similarity judgments and decision making share component processes:

(a) *Weighing of dimensions.* Tversky's seminal paper "Intransitivity of preferences" (1969) suggested that when choosing among multidimensional options,

people evaluate and compare options in reference to a single dimension. People do not integrate multiple dimensions when making a choice. A rational decision rule that selects the option with the highest attribute value can yield intransitive preferences when more than one dimension is relevant to the decision maker. One of the paradigms used by Tversky provided participants with a choice between two options that varied in the amount and the probability of winning some or no money. Small differences that showed a decrease in monetary reward with an increase in probability resulted in favoring the choice with the higher reward. However, when this difference was bigger, participants favored the choice with the higher probability resulting in intransitivity of choice. A study performed by Goldstone and Medin (as cited by Medin et al., 1995) found intransitive similarity judgments. Participants were asked to select which of two alternatives was most similar to a standard. One of the available strategies was to base selection on the largest dimensional difference. In one case the strategy led participants to choose a stimulus based on the color dimension; in another case, based on the size dimension; and in a third case, based on the angle dimension. This strategy induced intransitivity in similarity. Both of these examples of decision making and similarity judgments display value-specific dimension weighing in which small differences between choices in a specific dimension have a smaller effect than large differences on the same dimension.

(b) *Common Scale.* Luce and Raiffa (1957) proposed that preference among gambles might be mapped into a numerical utility function. Examples of framing effects have shown that preferences are not always converted into a single scale (Kahneman & Tversky, 1979). A gain of $100 is not perceived equally as a loss of

$100. Losses appear larger than gains. In similarity judgments, prototype and exemplar theory use featural overlap as a common metric. Similarity between two objects increases as function of the number of features they share and decreases as a function of mismatching features. However, different methods for obtaining judgments produce different values. In similarity judgments, participants assign more weight to common features and less weight to distinctive features. In dissimilarity judgments, more weight is give to distinctive features (Gati & Tversky, 1984; Tversky, 1977). Theories for both decision making and similarity judgments have proposed the use of a common scale for comparisons. These last two examples suggest that this is not necessarily so.

(c) *Reference Points and Asymmetry*. A study by Lowenstein (as cited by Medin et al. 1995) found that reference points determine the value of a purchased good. People requested a higher compensation when they agreed to delay the reception of a purchased good than what they offered to pay for a rush delivery of said good, resulting in asymmetries in judgment. Similarity comparisons can also produce asymmetries. The subject of the comparison often appears to be more salient than the referent and its features are given more weight (Tversky, 1977). The referent is judged more similar to the subject than vice versa. For example, people rate the similarity of China to North Korea to be less than the similarity of North Korea to China. The parallels described in (a), (b) and (c) suggest a correspondence between similarity judgments and decision making.

Dimensionality in Categorization Judgments

Research in psychology is based on the assumption that stimuli are perceived and judged in a dimensionally organized fashion (Krantz & Tversky, 1975). A category can be represented as a collection of points in multidimensional space, where each point represents a category member. The multidimensional space represents one dimension for each dimension of similarity among the category members. A categorization judgment is made by processing a subset of the most salient dimensions. In fact, a series of categorization studies performed by Medin, Wattenmaker and Hampton (1987) suggest that people prefer to use a subset containing only one dimension.

An example of unidimensional categorization is the experiment in which the subjects were presented with four countries that naturally formed two clusters (Tversky, 1977). Context effects increase the salience of one dimension over another. Panel A in Figure 2 shows how the countries are grouped based on the religion dimension. Panel B shows how the countries are grouped based on the location dimension.

Figure 2

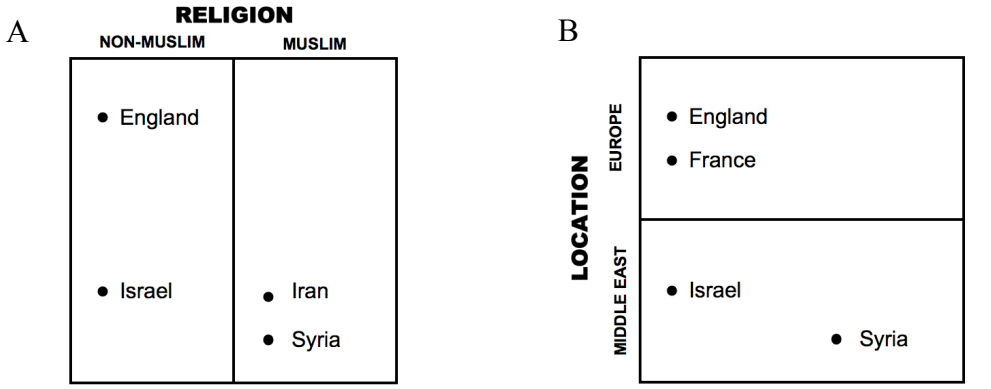

Figure 2. Dimensionality in Categorization Judgments: Panel A presents natural groupings based on the religion dimension. Panel B presents natural groupings based on the location dimension.

Categorical Structure

Rosch et al. (1976) have argued that categories are formed to communicate the contingency structure of attributes in the real world. Categorical structure is a measure of within- and between-category association.

Research performed by McCloskey and Glucksberg (1978) suggest that natural categories do not have clear boundaries that separate category members from non-members. Participants were give exemplar-category name pairs that varied in typicality and were asked to verify category membership. Participants were consistent within sessions and amongst themselves for highly typical objects (chair-furniture) and for unrelated objects (cucumber-furniture), but not for intermediately typical items (bookends-furniture).

Homa and Cultice (1984) performed several studies in which categorical structure was varied. A prototype was created by randomly assigning and connecting dots within a grid with a line. Moving each dot as per a previously designed statistical rule produced members of the same category. The degree of categorical structure was determined by how far each dot was moved. Their results showed that correct classification of novel exemplars is strongly and negatively correlated with degree of distortion of the exemplar from their respective prototype. When feedback was given regarding the correctness of classification, categories consisting of low distortions were learned faster than those consisting of large distortions. The authors concluded that highly structured material should be rapidly learned. A study by Posner, Goldsmith and Welton (as cited by Posner & Keele, 1968) had similar findings. As

the variability amongst instances of a category increased, the rate at which a category was learned decreased.

There are two facets of categorical structure: (a) the relation of members of a category to each other, and (b) their relation to items outside the category. This study is interested in the former. I view categorization as a process that is not solely influenced by the individual instances that belong to a category but by the properties these instances share within their reference class. Different categories can vary in their organization of exemplars. The degree of within-category associations has been proposed as a theoretical property of categories (Joelson & Herrmann, 1978). Posner and Keele (1968) referred to tight concepts as those with low variability and loose concepts as those with high variability.

Figure 3

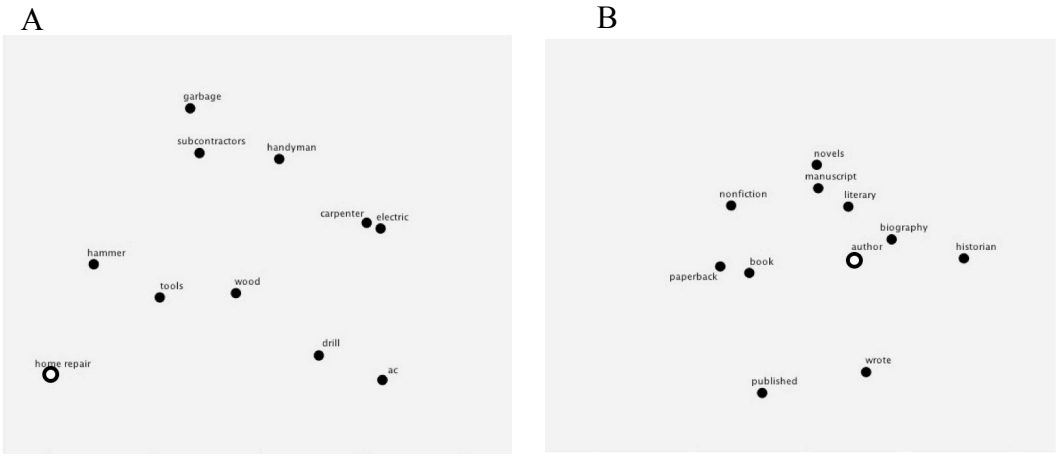

Figure 3. Examples of loose and tight categories represented in semantic space: Panel A presents a loose category ("home repair"). Panel B presents a tight category ("author").

Imagine all members of a category represented in a semantic space (See Figure 3). Categories with tight structures will have items that are semantically close to each other. A tight structure is one that has high inter-item similarities. Categories with loose structures will have items that are semantically far from each other. A

loose structure is one that has low inter-item similarities. It is foreseeable that the degree of within-category associations could influence categorization judgments. Before discussing this hypothesis, I will propose a new approach to measure categorical structure.

Semantic distance

Rips, Shoben & Smith (1975) performed a series of experiments exploring semantic distance. Following a multidimensional scaling analysis of ratings for semantic distance, they concluded that this metric could be represented as Euclidean distance in a semantic space. Additionally, semantic distance was able to predict response time in a categorization task and choices in an analogy task.

The semantic similarity[5] for two words can be assessed by analyzing the set of documents in which these words occur and assigning a metric based on their semantic content. Word similarity measures are computational means for calculating the association strength between terms. They can be obtained in two forms: (a) By performing a relationship analysis of a thesaurus (or an ontology). Miller (1995) developed a method that quantifies similarity relationships based on information from the manually crafted WordNet thesaurus. The thesaurus is represented as a hierarchy and its terms (words) are represented as nodes. Similarity is the minimal distance between the term nodes. In theory, this method can be used with any ontology. (b) Or by analyzing co-occurrence statistics in a text corpus. One line of work in the information retrieval literature considers two words as similar if they occur often in

[5] Throughout the paper I will use the terms similarity and semantic similarity interchangeably.

the same documents (Widdows & Dorow, 2002; Dorow & Widdows, 2003). These techniques are based on statistics, information retrieval and computational linguistics.

The use of ontologies such as WordNet to investigate semantic similarity has been criticized due to its manual nature. Domains are rapidly changing (i.e. technology) and the means required to recognize and classify new terms is resource intensive. Researchers would need to update their ontologies constantly to reflect current usage of the English language. The use of co-occurrence statistics in a text corpus bypasses this problem. There is no need of human resources to recognize and classify new terms. Once the appropriate corpus is obtained, a computer program can analyze association strength between terms even when they are new. I will be using this last approach to calculate a word similarity measure for all stimuli used in this study. Particularly, I will be using the Infomap software provided by the Computational Semantics Lab from Stanford University (n.d. a). This software builds a multidimensional space -- called WORDSPACE -- for a text corpus (Computational Semantics Lab from Stanford University, n.d. b). Terms are represented as word vectors that encode information about how the word is distributed over the corpus. Each word vector represents a list of coordinates that point towards a specific location in a multidimensional vector space. A term document matrix can be created, where rows represent terms, columns represent documents and cells specify how many times a term occurred in a particular document. A problem usually encountered when building such matrices is that similar words are seldom used in the same document. The Infomap software instead builds a co-occurrence matrix as in Figure 4, where rows represent special content bearing terms, columns represent terms and cells

specify how often regular terms co-occurred with content-terms (within a pre-established window). To differentiate content-terms with regular-terms, Infomap selects the 1,000 most frequent words in a corpus as content-terms, excluding stopwords. A stopword is a frequently used word, such as "a" or "the", that is filtered out prior to the processing of natural language data. Stopwords are usually not used in search engine queries nor indexed in online documents.

Figure 4

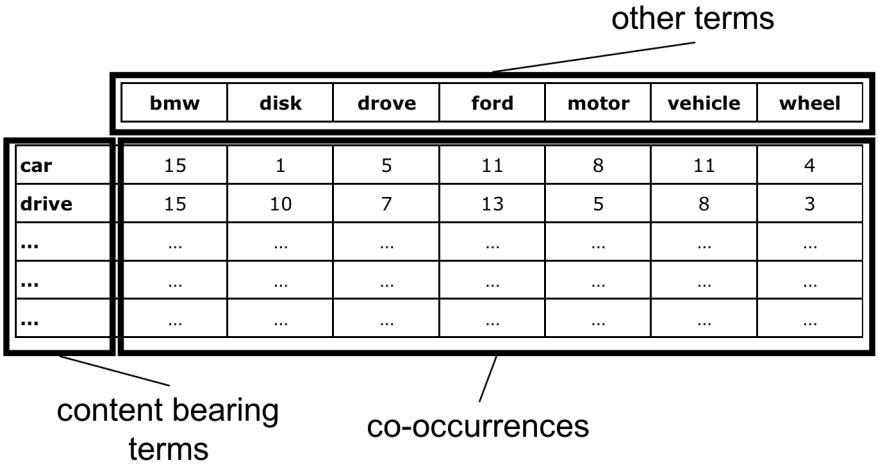

Figure 4. Hypothetical example of a Co-Occurrence Matrix, based on the second figure presented in "Infomap algorithm description." (Computational Semantics Lab from Stanford University, n.d. b).

Note that the dimensionality of the WORDSPACE at the moment is quite high and has at least 1,000 coordinates. The Infomap software is able to reduce the number of dimensions by means of latent semantic analysis[6] (LSA). The reduced number of dimensions used for this study is 100[7]. If two dimensions have an equivalent context, LSA combines these two axes into a single "latent" axis. This scaling method permits words with similar meaning to have similar vector representations even though they

[6] Latent semantic analysis is also called latent semantic indexing or singular value decomposition.
[7] This is also the default number of dimensions for the Infomap software. A different number of dimensions can be calculated by performing additional singular value decomposition analyses. It would require licensing a different software: SVDPACKC (Retrieved February 5, 2008, from http://www.netlib.org/svdpack/index.html)

may have never co-occurred in the same document. As a result, a more accurate representation of the relationship between words is created (Steyvers, Griffiths and Dennis, 2006). LSA is one of several methods used to analyze document collections. For example, the Hyperspace Analog to Language (HAL) model uses word vectors with coordinates that represent weighted co-occurrence values between words (Lund & Burgess, 1996). LSA uses word vectors with coordinates that represent co-occurrence between words and the documents they occur in. The assignment of coordinates in this manner implies that HAL does not use documents as boundaries and LSA does. Another example is probabilistic topic models, which assume that a document is composed of a collection of topics. These models represent words using topics. The topics can be identified manually or by assigning the topic label to the word with the highest probability[8]. LSA represents words as vectors in a multi-dimensional space. Both LSA and probabilistic topic models use words and documents and are considered to be "similar in spirit" (Steyvers et al., 2006, p. 331).

Word similarity is obtained by calculating the cosine of the angle between two word vectors as specified in equation 1:

$$\cos(a,b) = \frac{a \cdot b}{\|a\|\|b\|} \qquad [1]$$

One word, A, is represented by vector a with coordinates $[a_1,a_2,\cdots,a_n]$; a second word, B, is represented by vector b with coordinates $[b_1,b_2,\cdots,b_n]$. The cosine similarity is obtained by dividing the scalar product by the norms of vectors a and b. The scalar product between vectors a and b is calculated by the sum of

[8] For example, the word *play* is represented with topic *077* – music – in a specific document and is represented with topic *082* – literature – in another document (Steyvers et al., 2006, Figure 1, p. 330).

products of their coordinates ($a \cdot b = a_1 b_1 + a_2 b_2 + \cdots + a_n b_n$). The norm of a vector is calculated by obtaining the square root of the sum of its squared coordinates ($\|a\| = \sqrt{a_1^2 + a_2^2 + \cdots + a_n^2}$). The use of the cosine of the angle between two word vectors indicates that it is the direction – not the length -- of the vectors that is relevant to calculate word similarity. Most cosines between words are positive, though small negative values are common. The study will use stimuli with word similarities between [0, 1], where 1 signifies high similarity and 0 signifies little similarity.

Family resemblance is a construct that denotes the extent to which category members share attributes with other category members. Rosch and Mervis (1975) showed that items that have the highest family resemblance also have the fewest attributes in common with members of related contrast categories. This is not the case with the similarity measure used in this study. An item can have high similarity correlations with all members in its category but there are no restrictions as to whether that same item should have low similarity correlations with members of other categories.

Research has demonstrated that words that are semantically similar usually occur with similar distributions and in similar contexts (Miller and Charles, as cited in Widdows & Dorow, 2002) leading to similar word vectors. In order to create the stimuli for each category, a list of words needs to be extracted using the notion of semantic similarity. The Infomap software uses an incremental algorithm for extracting categories of similar words as specified by Widdows and Dorow (2002):

Let A be a set of nodes and let $N(A)$, the neighbors of A, be the nodes which are linked to any $a \in A$. (So $N(A) = \bigcup_{a \in A} N(A)$.)

The best new node is taken to be the node $b \in N(A) \setminus A$ with the highest proportion of links to $N(A)$. More precisely, for each $u \in N(A) \setminus A$, let the affinity between u and A be given by the ratio

$$\frac{|N(u) \cap N(A)|}{|N(u)|}$$

The best new node $b \in N(A) \setminus A$ is the node which maximizes this affinity score. (p. 1095)

This algorithm can be explained in five steps. First, a seed word is fed to the algorithm. Second, it starts counting the co-occurrence of words and seed words within the corpus. Third, it calculates the affinity score upon these counts to select new seed words. Fourth, steps 2 and 3 are iterated n times. Fifth, it uses the affinity score to rank words for category membership. For this study, a list of seed words was created and manually fed into the algorithm. This was done to control the types of categories used as stimuli (e.g. types of categories are occupation, location, hobby or sport).

The co-occurrence statistics approach is also used for document retrieval systems. Scatter/Gather (Pirolli, Schank, Hearst & Diehl, 1996) is a cluster-based browsing system for document collections. It uses a measure of inter-document similarity to cluster documents. Documents are represented as vectors, with each vector coordinate associated with a unique content word (previously defined in the document collection). The similarity of two documents is computed by the cosine of

the angle between the two vectors representing each document. They call this the cosine measure or normalized correlation. As a side note, the Scatter/Gather interface has a similar purpose as that of tag clouds. They both provide an interactive method to support the browsing of a text collection by means of a summary of the content of the said text collection.

Table 1

A

	home repair	tools	hammer	wood	subcontractors	handyman	drill	garbage	ac	electric	carpenter
home repair	1.000	0.674	0.642	0.636	0.634	0.612	0.609	0.608	0.591	0.585	0.557
tools	0.674	1.000	0.756	0.780	0.784	0.702	0.749	0.726	0.730	0.738	0.725
hammer	0.642	0.756	1.000	0.755	0.705	0.698	0.715	0.743	0.705	0.668	0.716
wood	0.636	0.780	0.755	1.000	0.727	0.797	0.770	0.697	0.706	0.712	0.789
subcontractors	0.634	0.784	0.705	0.727	1.000	0.716	0.702	0.755	0.664	0.771	0.770
handyman	0.612	0.702	0.698	0.797	0.716	1.000	0.711	0.767	0.720	0.757	0.767
drill	0.609	0.749	0.715	0.770	0.702	0.711	1.000	0.729	0.761	0.762	0.741
garbage	0.608	0.726	0.743	0.697	0.755	0.767	0.729	1.000	0.640	0.736	0.689
ac	0.591	0.730	0.705	0.706	0.664	0.720	0.761	0.640	1.000	0.782	0.738
electric	0.585	0.738	0.668	0.712	0.771	0.757	0.762	0.736	0.782	1.000	0.719
carpenter	0.557	0.725	0.716	0.789	0.770	0.767	0.741	0.689	0.738	0.719	1.000

V M

B

	author	book	literary	novels	paperback	published	biography	nonfiction	historian	manuscript	wrote
author	1.000	0.942	0.928	0.928	0.926	0.918	0.917	0.914	0.908	0.907	0.906
book	0.942	1.000	0.935	0.932	0.975	0.928	0.915	0.951	0.841	0.933	0.928
literary	0.928	0.935	1.000	0.951	0.918	0.900	0.934	0.920	0.877	0.925	0.908
novels	0.928	0.932	0.951	1.000	0.932	0.891	0.931	0.931	0.857	0.921	0.903
paperback	0.926	0.975	0.918	0.932	1.000	0.918	0.912	0.955	0.827	0.929	0.907
published	0.918	0.928	0.900	0.891	0.918	1.000	0.898	0.911	0.822	0.889	0.910
biography	0.917	0.915	0.934	0.931	0.912	0.898	1.000	0.929	0.889	0.912	0.912
nonfiction	0.914	0.951	0.920	0.931	0.955	0.911	0.929	1.000	0.825	0.916	0.882
historian	0.908	0.841	0.877	0.857	0.827	0.822	0.889	0.825	1.000	0.871	0.846
manuscript	0.907	0.933	0.925	0.921	0.929	0.889	0.912	0.916	0.871	1.000	0.886
wrote	0.906	0.928	0.908	0.903	0.907	0.910	0.912	0.882	0.846	0.886	1.000

V M

Table 1. Table 1-A (top panel) presents word-to-word similarities for the category "home repair". Table 1-B (bottom panel) presents word-to-word similarities for the category "author". The first column in each table represents the similarity vector for this category (this vector is highlighted with Boundary V). Each entire table represents the similarity matrix for this category (the matrix is highlighted with Boundary M).

The organization of exemplars within a category can be represented by means of a similarity vector or by a similarity matrix. For each category, semantic distances between exemplars and the category label can be computed. These distances can be

arranged into a similarity vector with words as coordinates. For each category a matrix of pairwise similarities can be computed with cells within the matrix representing word-to-word semantic distances as obtained through LSA. Table 1 presents examples of vectors and matrices for two categories. A measure of central tendency for both vectors and matrices can be calculated. Categorical structure will be operationalized by this measure.

Figure 5

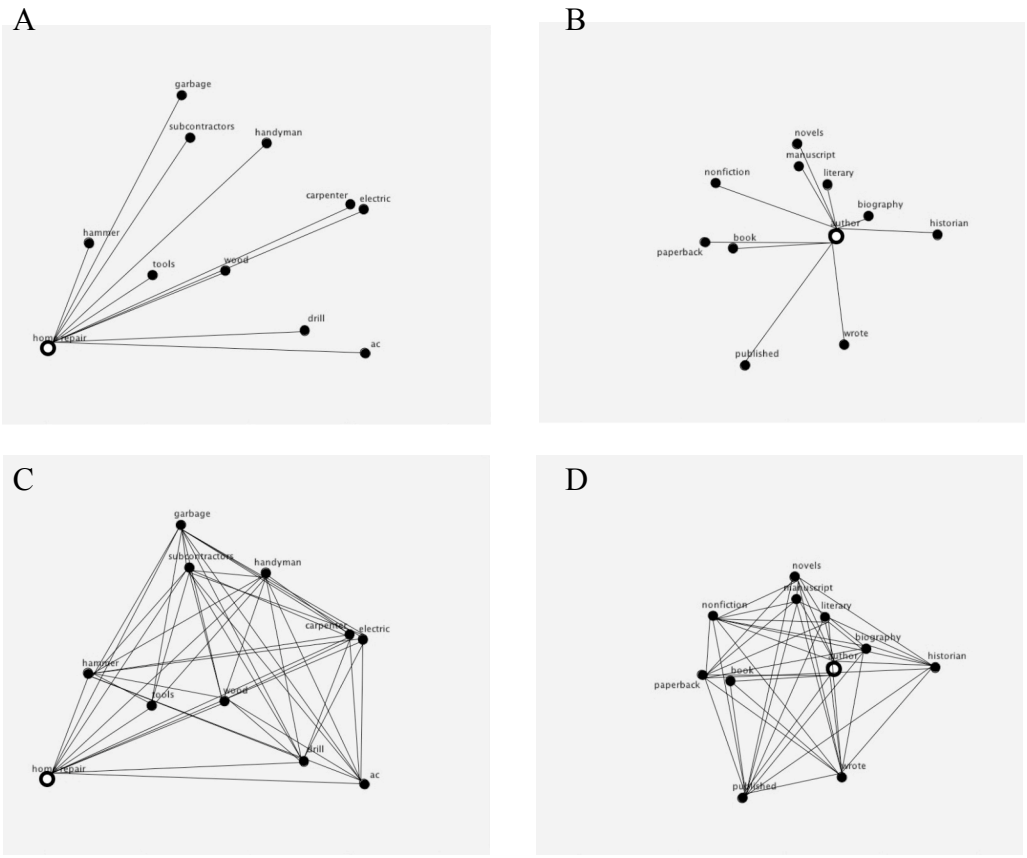

Figure 5. Pictorial depictions of information represented by a similarity vector and matrix. A similarity vector represents the semantic distances between the category label and the category members, as seen in Panels A and B. Panel A presents a loose category (home repair). Panel B presents a tight category (author). A similarity matrix represents the semantic distances within all members of a category (including its label), as seen in Panels C and D. Panel C presents a loose category (home repair). Panel D presents a tight category (author).

A category's similarity vector and matrix are two different representations of categorical structure. A vector represents structure by computing the relationship between members and the category label. A matrix represents structure by computing the relationship within all members of the category (including its label). The difference between these two methods can be pictorially illustrated (See Figure 5). One could claim that a matrix's mean of pairwise similarities is a stronger measure of categorical structure because it takes into account all elements of a class and all relationships between these elements. One could also claim that a vector's mean of similarities is a better measure of categorical structure because of its simpler computation and that people may not exhaustively compute all inter-item similarities. It is apparent, based on their mathematical definition, that these two measures are highly correlated. This paper will use both measures in all related analyses in order to compare them.

It has been argued that similarity is too flexible. Similarity *is* flexible. Two items may have different degrees of similarity depending on the context in which they are compared. For example, items such as "cotton" and "drip" could be judged to be more similar if they are compared in a medical context. The use of LSA allows researchers to make provision for such flexibility. The semantic distance for this pair is .68 in a general text corpus such as the New York Times. The semantic distance is .75 in a medical text corpus such as the MEDLINE database[9]. It has been argued that

[9] The New York Times corpus spans from 1994 through 1996, it was obtained from the North American News Text Corpus published by the Linguistic Data Consortium with approximately 143M words and 370K documents (Widdows, 2003). The MEDLINE database is a collection of 270 medical journals spanning from 1987 through 1991. This collection is also known as the Ohsumed corpus of medical documents, which contains approximately 40M words and 230Kdocuments (Hersh, Buckley, Leone, & Hickman, 1994; Widdows, 2003).

there is no consensus on what similarity is. The use of semantic distance as measured by LSA permits a narrow definition for similarity. Similarity between two items is defined by the usage of those items in a predetermined language and, if necessary, a predetermined domain.

The role that dimensionality played in categorization was previously mentioned; only a subset of dimensions may be attended to in order to group similar objects. This selective process relies on the assumption that stimulus dimensions are separable (Garner, 1978). Dimensions of real-world entities are not always independent of one another. The importance of scaling solutions is they suggest that the multiple dimensions that describe an entity may be reduced to a single semantic distance, regardless of their interconnectedness while preserving as much as possible the covariation structure of words and documents.

Categorization, Tagging and Tag Clouds

Social book marking tools permit individuals to create metadata for websites they encounter online. This activity is called *tagging*. Tagging-based systems enable users to assign keywords and/or insert their own explanations to web resources in order to organize these resources (Halvey & Keane, 2007). Tagging could be considered as a categorizing mechanism. Once a webpage that requires bookmarking is encountered, the person compares the content of the page to different concepts and chooses one (or more) of these categories as tags. Individuals are able to organize and display a document collection with meaningful labels by using tags. Tagging is a more flexible and convenient method of categorizing for several reasons: (a) the person can assign multiple tags to a webpage, thus the webpage or document can

belong to more than one category (this overcomes the limitation of traditional hierarchically organized folders that most browsers offer), (b) tags can be suggested based on the semantic content of the webpage, and (c) if the specific webpage has been previously tagged by others, the system can also suggest what tags are related. Tagging is also a more complex method of categorization because a person needs to consider additional elements when assigning a tag. If a webpage is assigned a generic tag, akin to a basic-level category, the documents contained in this tag can be quite numerous. If this is the case, then the webpage may be difficult to find in the future. The person also needs to consider the architecture supported by the tagging site. Humans start by learning basic categories, for example, dogs. As they develop, they learn sub- and super-ordinate categories. For example, dachshunds and beagles are types of dogs and dogs are types of mammals. The cognitive system is able to build taxonomies naturally. A computer system is not and will require user effort to make the changes and re-organize the tags in a taxonomy.

Tag clouds are text-based visual depictions of content tags that belong to a person's or group's bookmarks. Tag importance – or frequency of occurrence – is usually emphasized by the use of font size, although factors such as order, color and boldness have been also used to denote importance. Trends in social software have increased the popularity of tag clouds. Websites that employ tag clouds provide users with a content overview of its document repository. For example, the photo sharing site – flickr.com (Marlow, Naaman, Boyd and Davis, 2006) – allows users to upload personal images to make them publicly accessible. Uploaders can organize pictures by tagging them. Flickr uses a tag cloud to visualize the most popular tags and

provide users an overview of the type of images it contains. Another example is the social bookmarking website del.icio.us. Users organize bookmarked websites by tagging them. Del.icio.us also uses tag clouds to provide overviews of the type of websites it provides links to[10]. In social bookmarking sites, individual tags are clickable and link to subsets of repository content. For example, clicking on a tag for "Seattle" will link to a subset of tags and websites related to the "Seattle" tag. Tag clouds function as both summaries of the information they represent and as means of topical browsing. Tagging systems offer two different navigation mechanisms of a document repository. First, a user can click on the name of other users in order to see their bookmarks and can get a sense of the topics of interest and/or expertise of a particular user. A tag cloud can provide a meaningful reflection of the topics of general interest of the tag cloud owner (Millen, Feinberg & Kerr, 2005). Second, a user can click a particular tag to see all bookmarks that share that common tag and can browse this new set of bookmarks to search for new sources. By browsing specific people and tags, users can find people that share common interests and new relevant websites (Golder & Huberman, 2006). A case study of enterprise-wide social bookmarking performed by Millen, Feinberg and Kerr (2005) found that during the initial usage period of their company's social bookmarking system, 42 percent of its 300 users created bookmarks and 57 percent navigated to an original document tagged by others. As a note of caution, the social nature of tagging systems has the potential of creating confusion during browsing and missing sources of interest.

[10] Following are the direct links to flickr's and del.icio.us' tag cloud webpages. Retrieved on February 4, 2008, from:
http://www.flickr.com/photos/tags/
http://del.icio.us/tag

Research has shown that categories do not necessarily have clear boundaries (McCloskey & Glucksberg, 1978), add to the mix the fuzziness of linguistic boundaries and the tagging system could result in a collection of idiosyncratic personal categories in addition to basic categories (Golder & Huberman, 2006).

Tag clouds are prime candidates for stimuli in categorization research. The elements (tags) that compose these visualizations are a byproduct of categorization processes. Despite the increasing popularity of tag clouds, there have been few experimental studies evaluating their effectiveness (Rivadeneira, Gruen, Muller & Millen, 2007). An additional goal of this paper is to provide guidelines on how to visually present tags so that the information they represent can be accurately transmitted.

Questions of Interest, Predictions and Hypotheses

I am arguing that the measure of categorical structure presented in this study is a measurable construct based on the use of the English language. This proposition can be assessed by examining whether the measures obtained through a latent semantic analysis of an English corpus translate unto actual judgments. To test this idea, I have designed three experiments aimed at exploring judgments of category membership. The stimuli used in all experiments share the same framework: tag clouds. In Experiment 1, participants perform category retrieval tasks after each stimulus is presented. A test of memory recognition follows each categorization judgment. Experiment 2 simultaneously presents a stimulus and asks participants to perform a category verification task. Confidence judgments are collected for each categorization judgment. Experiment 3 simultaneously presents a stimulus and asks participants to

perform a category retrieval task. These experiments will test how categorical structure and different formats affect judgments of category membership.

Categorical Structure

To the extent that categorization is based on categorical structure, it should be more specifically affected by manipulations of categorical structure. Differences in observed accuracy of categorization for classes with loose structure and those with tight structures would foretell a categorization process rooted on categorical structure. I hypothesize that there will be a correspondence between the degree of categorical structure and judgments of category membership.

Format

Salience of particular dimensions can influence selective attention resulting in changes in the degree of judged similarity between two items (Medin & Shaffer, 1978). Research on attention has found an effect of reading direction. English-speaking participants show a left-to-right bias and Arabic-speaking participants show a right-to-left bias (Spalek & Hammad, 2005). Such biases suggest that the layout of items on a screen may result in increasing the attention of particular items over others. Research has suggested that prominence is a variable that affects judgments of similarity (Tversky, 1977; Mervis & Rosch, 1981). Font size has been studied as a variable that affects the performance of signal words in warnings. A study by Adams and Edworthy (1995) found an increasing linear relationship between font size and perceived urgency. Words with larger fonts may be considered as more prominent than words with smaller fonts. It has been argued that categorization is based on

similarity and consequently it should be sensitive to factors that influence similarity, such as prominence and selective attention. Different font sizes and layouts are hypothesized to produce differences in judgments of category membership.

Chapter 2: Experiment 1

In Experiment 1, participants perform category retrieval tasks after each stimulus is presented. A test of memory recognition follows each categorization judgment.

Methods

Participants

University of Maryland undergraduate students (n=17; 6 males and 11 females) and IBM employees (n=13; 10 males and 3 females) participated in Experiment 1. Employees volunteered and students received course extra credit. All subjects had normal or corrected-to-normal vision. Participants were run individually in single sessions lasting approximately 30 minutes.

Materials

Materials included 52 tag clouds that varied among some dimensions of format and were presented in PC-based equipment using MediaLab research software (Jarvis, 2006).

Font Type

A study performed by Mansfield, Legge and Bane (1996) found a small advantage of a fixed-width, sans-serif font over a proportionally-spaced serif font for subjects with low vision. For subjects with normal vision, the differences were slighter, with the proportionally-spaced serif font having an advantage for reading speed. The experiments will control for people with normal vision. Additionally, the

study is concerned in controlling for reading comprehension rather than reading speed. The font that will be used in the tag cloud stimuli is "Gill Sans", which is a proportionally-spaced sans-serif font.

Contrast Polarity

Research in psychophysics for normal vision has found that contrast polarity has little effect on reading (Legge, Pelli, Rubin, & Schleske, 1985). Contrast polarity will not be manipulated; all stimuli will be black-on-white, black fonts on a white background.

Layout

Figure 6 presents the different layouts used in the experiment. A study performed by Vitu, Kapolua, Lancelin and Lavigne (2004) found a systematic bias of the eye behavior towards the center of the visual display. They proposed that this systematic deviation is resource efficient, as eye movements should be contained within the part of the visual configuration where stimuli are displayed. A layout was created with the largest word located towards the center of the tag cloud, the Spatial Layout (Feinberg's algorithm[11]). Research has shown that reading direction has an effect on word recognition. Rivadeneira et al. (2007) performed a memory study using spatial tag clouds where they investigated what effects the different locations within the tag cloud had on a free recall test. A quadrant effect was found, words located on the top-left quadrant were retrieved more frequently than other areas in the tag cloud. This effect is usually expected on stimuli that require westernized reading

[11] This algorithm is proprietary to IBM. I have been able to use it to create my stimuli thanks to my collaborators at the Collaborative User Experience group from IBM Research: Daniel Gruen, Michael Muller, David Miller and the algorithm creator, Jonathan Feinberg.

(left-to-right and top-to-bottom). Battista and Kalloniatis (2002) demonstrated that a reading direction effect is a consequence of attending to a particular area of visual space as part of the normal reading habit. Another effect of reading direction was shown in a study by Morikawa and McBeath (1992) where results provided strong evidence that reading habits can influence directionality in motion perception (participants used to Westernized-reading exhibited a bias to experience leftward movement with ambiguous motion stimuli). The Sequential Layouts facilitate left-to-right reading and the Single Column List Layout facilitates top-to-bottom reading. In addition, these types of layouts are among the most common types found in the industry. The Sequential Layout with Alphabetical Sorting is found on flickr (Marlow, Naaman, Boyd and Davis, 2006) and Josuha Schacther's del.icio.us (Golder & Huberman, 2006). The Sequential Layout with Frequency Sorting and the Single Column List with Frequency Sorting are features available to del.icio.us' users. The Spatial Layout is used in IBM's enterprise-wide bookmarking site.

Note that the Single Column List with Frequency Sorting has a scrollbar on the right. The initial monitors used for Experiment 1 were small and did not permit the presentation of this layout in a single screen. Participants needed to scroll through this tag cloud in order to see all words presented.

Figure 6

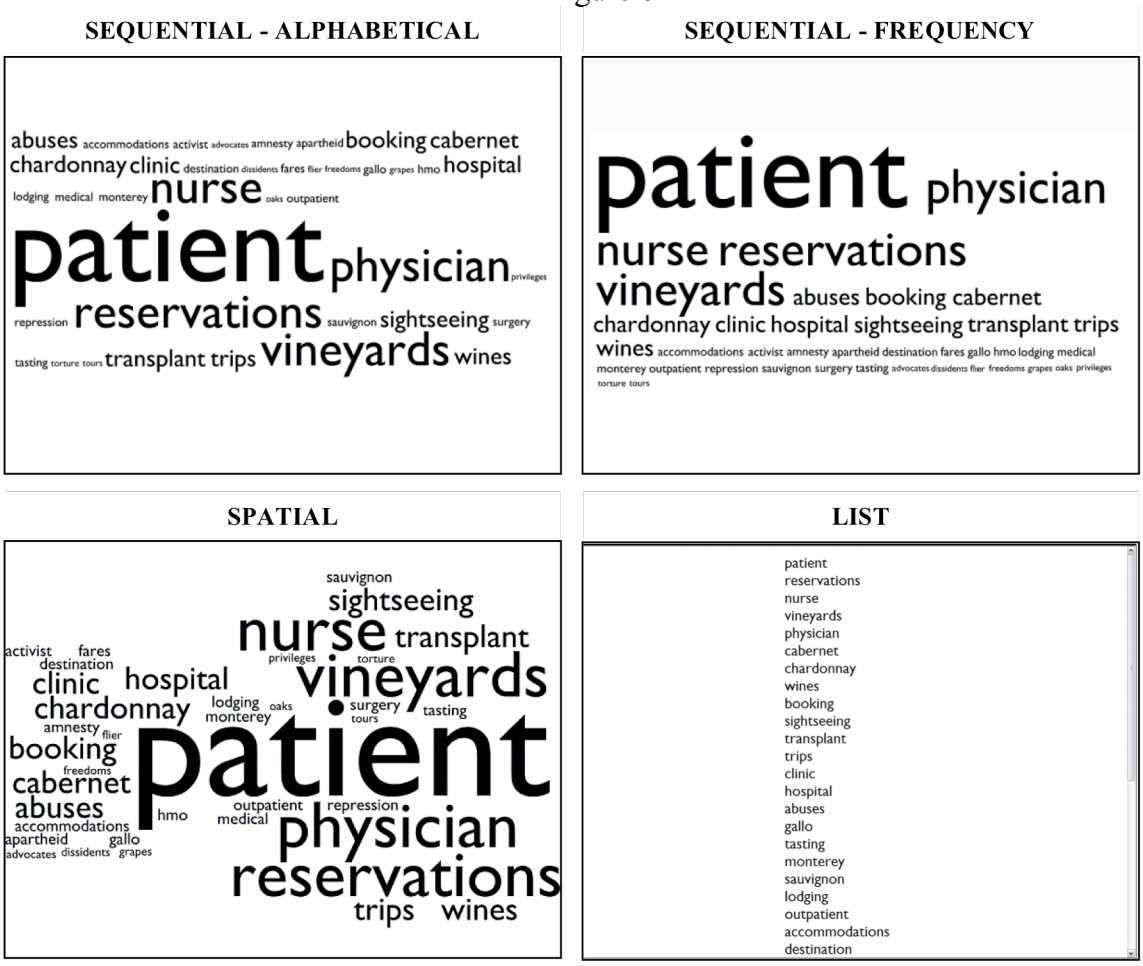

Figure 6. Figure 6 presents an example stimulus formatted based on the four different layouts used: Sequential with Alphabetical Sorting, Sequential with Frequency Sorting, Spatial Layout and Single Column List with Frequency Sorting. Note that there are four different categories present: doctor, winery, travel and human rights – in decreasing order of prominence.

Category, Words and Tag Clouds

Fifty-two categories and 764 words were obtained from the Information Mapping Project (Computational Semantics Lab from Stanford University, n.d. a). Categories are obtained by the distribution of co-occurrences between a word and some set of content-bearing terms. The document collection used for this study is the New York Times corpus spanning from 1994 through 1996, from the North American

News Text Corpus published by the Linguistic Data Consortium with approximately 143M words, 370K documents (Widdows, 2003).

Four categories appeared per tag cloud; one related to an occupation and the other three were either hobbies or travel locations. The category seed was not used as stimulus. Ten words per category were used for each tag cloud, for a total of 40 words. A tag-per-person analysis for repeat users of IBM's enterprise-wide social bookmarking site through April of 2006 reveals a mean of 39 tags/person (Rivadeneira et al., 2007). Each category was associated with four distractor words to be used in the memory recognition test. One of these distractors was semantically unrelated to the category, two were semantically related and one was the category seed[12]. Thirteen tag clouds were created in all four layouts for a total of fifty-two tag clouds.

Design and Procedure

Experiment 1 consisted of three phases: a presentation phase, a category retrieval phase, and a recognition phase. Initial instructions welcomed the participants and provided them with the definition of a tag cloud and a general example of one. A tag cloud was said to represent the general interests of a person who was named "the tag cloud owner". Subjects performed one practice trial and twelve experimental trials. Participants were informed that no data would be collected during the practice trial. After the practice trial, participants were given the opportunity to ask questions before the experimental trials started. Participants were further informed of the details of the experimental procedure. Each trial encompassed all three phases and started

[12] Throughout the paper I will use the terms „category seed" and „ category label" interchangeably.

with the presentation of a blank screen for a period of 1 s. The presentation phase presented a tag cloud for a period of 30 s. Participants were told to study each tag cloud and try to make an inference as to what were the main interests of the person being represented in each tag cloud. In addition, they were told to remember the individual words presented in the tag cloud for a word recognition test. Participants were not informed that there were four main interests (categories) listed in each tag cloud. The category retrieval phase was self-paced but had a maximum allowable time of 120 s and participants were informed of this time limit in the instructions. This phase asked the participants to list the main interests of the tag cloud owner. Responses were collected in an essay form and participants had complete editorial freedom. They could use sentences or single words as descriptors and could list as many interests to better describe the tag cloud owner. The instructions also informed participants that once they had finished responding, they could click a continue link in order to go to the next phase. The category retrieval phase was also a distractor task and was meant to eliminate any recency effects for the memory test that followed. The recognition phase consisted of an old-new recognition test. It contained 16 targets, 12 semantically related distractors, and 4 unrelated distractors. Among the 12 semantically related distractors: 4 words were the category labels, 4 had high semantic similarities (similarity > .80) and 4 had medium semantic similarities (.60 > similarity > .80). The recognition phase presented one word at time, participants had to press the button labeled “True” if they thought the word had appeared in the tag cloud and the button labeled “False” if they thought the word had not appeared in the tag cloud. Pressing these buttons would allow participants to advance through the

recognition phase. This phase was also self-paced. However, if participants delayed their responses (no response after a period of 5 s), a warning message would appear advising participants to respond faster. Once the recognition phase was finished, participants were notified that the next trial was to begin. These three phases repeated until all trials were completed. See Appendix A for a diagram of an example trial.

Several factors were manipulated in this experiment. First, categorical structure varied among the categories presented. The measures of central tendency for the similarity vectors ranged between .572 and .957 and between .761 and .947 for the similarity matrices. Second, the tag cloud layout was manipulated. Each tag cloud was represented in four different fashions: Sequential with Alphabetical Sorting, Sequential with Frequency Sorting, Spatial Layout and Single Column List with Frequency Sorting (See Figure 6). Third, font size was manipulated into five different levels (F1 to F5, big to small) for the Sequential and Spatial Layouts. Font manipulation influenced the perceived prominence for each category. Although technically prominence is not a direct manipulation, it is a key variable of interest and I will refer to it as a manipulation. This manipulation was performed in systematic manner so that there would be three different levels of category prominence (high, medium and low) in every tag cloud. This prominence variable is observable in common tag clouds, where the most popular or most frequent terms are highlighted by means of either font size, weight or color (Kaser & Lemire, 2007; Rivadeneira et al., 2007). The assignment of the different levels of Font Size to each of the forty words in the tag cloud was translated into frequency of usage for each word. Thus level F1 not only represents the word with the largest Font Size but also the most

frequent occurring keyword in the tag cloud. For example, the word “patient” seen in Figure 6 is the most frequent term associated with documents tagged “patient” by the tag cloud owner. This frequency/font-size information was additionally used to construct the frequency layouts: Sequential with Frequency Sorting and Single Column List with Frequency Sorting. Table 2 presents how each level of prominence was manipulated. For example, the high prominent category contained one word with the largest font (F1), two words with the second largest font (F2), three words with the third largest font (F3) and four words with the second to smallest font (F4). Each tag cloud had four categories: one category with high prominence, two categories with medium prominence and one category with low prominence. Although the Single Column List with Frequency Sorting does not have a Font Size manipulation, it has a prominence manipulation. Prominence is operationalized in this case by sorting order. Words that appear higher in the list are those that otherwise would have larger fonts in the other three layouts. For example, the words “patient”, “physician”, “nurse”, “reservations” and “vineyards” are the five most largest words in the Sequential and Spatial Layouts. They correspond to the high prominent category “doctor” and to the two medium prominent categories “travel” and “winery”. These five words are the top five words in the Single Column List Layout.

Table 2

	Prominence			
	Low	Medium		High
Font Size	**(human rights)**	**(travel)**	**(winery)**	**(doctor)**
F1				patient
F2				physician
		reservations	vineyards	nurse
F3		Trips	wines	hospital
		sightseeing	chardonnay	clinic
	abuses	booking	cabernet	transplant
F4	amnesty	Fares	sauvignon	medical
	activist	destination	monterey	hmo
	repression	accommodations	tasting	surgery
	apartheid	lodging	gallo	outpatient
F5	advocates	Flier	oaks	
	freedoms	Tours	grapes	
	dissidents			
	torture			
	privileges			

Table 2. Table 2 presents an example on how Font Size, and thus Prominence, was manipulated. There were five different levels of Font Size (F1 thru F5, big to small), three levels of prominence (high, medium, low). Each tag cloud consisted of 4 categories with 10 words per category. There was one category with high prominence, two categories with medium prominence and one category with low prominence. This example shows the tag cloud represented in Figure 6, the high prominence category is "doctor", the medium prominence categories are "winery" and "travel", and the low prominence category is "human rights".

The practice trial was the same for all participants: the stimulus presented was a tag cloud in a Spatial Layout. There were twelve experimental trials in which tag clouds varied in layout. These trials were randomized for all participants. A counterbalancing scheme was used to diffuse any effects of category on layout (See Appendix B). For example, the "doctor" category (represented in Tag Cloud 1) was presented for Group 1 as a Single Column List with Frequency Sorting, for Group 2 as a Sequential Layout with Alphabetical Sorting, for Group 3 as a Spatial Layout and for Group 4 as a Sequential with Layout Frequency Sorting.

Results

Data Analysis

Two dependent variables were analyzed: correct category retrieval and recognition accuracy. A score was assigned to measure correct category retrieval. A point was given each time a subject correctly identified one of the categories presented in each tag cloud. Full credit was given when the category label or synonyms of the category label were used (i.e. "clinician" for "doctor"). Partial credit was given if the participants used words similar to the category (i.e. "surgery" instead of "doctor"). Four judges performed this scoring procedure. The inter-rater reliability was high (Average Measure Intraclass Correlation Coefficient = .955). There were no significant differences between the two groups of participants (IBM employees and UMD students), thus the data analysis will encompass all 30 participants.

Categorical Structure

I examined the correspondence between the percent of correct category retrieval and the two measures of categorical structure: means of similarity vectors and similarity matrices[13]. The unit of interest for the correlation analysis is at the category level. This implies averaging the scores from all participants and correlating that average score with each measure of categorical structure. Before averaging, I first tested whether the data was stationary. The equation to test homogeneity of correlations (Hedges & Olkin, 1985) is:

[13] These two measures of categorical structure were significantly correlated ($r(46)= .816$, $p< .05$). This makes sense because of the mathematical origin of both measures.

$$Q = \sum_{i=1}^{k} (n-3)\left(Z_i - \overline{Z}\right) \qquad [2]$$

In Equation 2, n represents the sample size used to estimate a particular correlation; Z_i represents the Fisher Z_i-transformed correlation[14]; and $\overline{Z}$ represents the average correlation. The Q statistic has $k-1$ degrees of freedom and is distributed as a chi-square distribution. Obtaining a non-significant Q implies that one cannot reject the possibility that the correlations come from the same population. Table 3 shows that the data is stationary and can be averaged.

Table 3

	Categorical Structure	
	Mean of Similarity Vector	**Mean of Similarity Matrix**
Correct Category Retrieval		
Pearson Correlation (r)	.280*	.308*
df	46	46
R^2	.078	.095
Gamma Correlation (G)	.220*	.223*
Q-statistic	17.34	17.05
df	29	29
	n.s.	n.s.

*p< .05

Table 3. Relationship between Category Retrieval and Categorical Structure as measured by the mean of each category's similarity vector and by the mean of each category's pairwise similarity matrix. Q values were derived from Equation 2 and represent tests of homogeneity.

Table 3 and Figure 7 show positive relationships between categorical structure and category judgments. Both the Pearson[15] and Gamma (Gonzalez & Nelson, 1995)

[14] Fisher Z_i-transformed correlation: $Z_i = \frac{1}{2}\ln\left(\frac{1+r}{1-r}\right)$

[15] A potential problem with the correlations found is the influence that outliers may exert on the results. Cook's distances (Cohen, Cohen, West & Aiken, 2003) were calculated to detect data points with unusual leverage. The findings are robust; new correlations obtained through analysis of influence statistics are in line with the results presented.

correlations were significant. The strength of these relationships was moderate[16]. Note that categorical structure appears to vary more when similarity vectors measure it. Structures appear to "tighten up" when pairwise similarities among all category members are introduced into categorical structure measures as those given by similarity matrices.

Figure 7

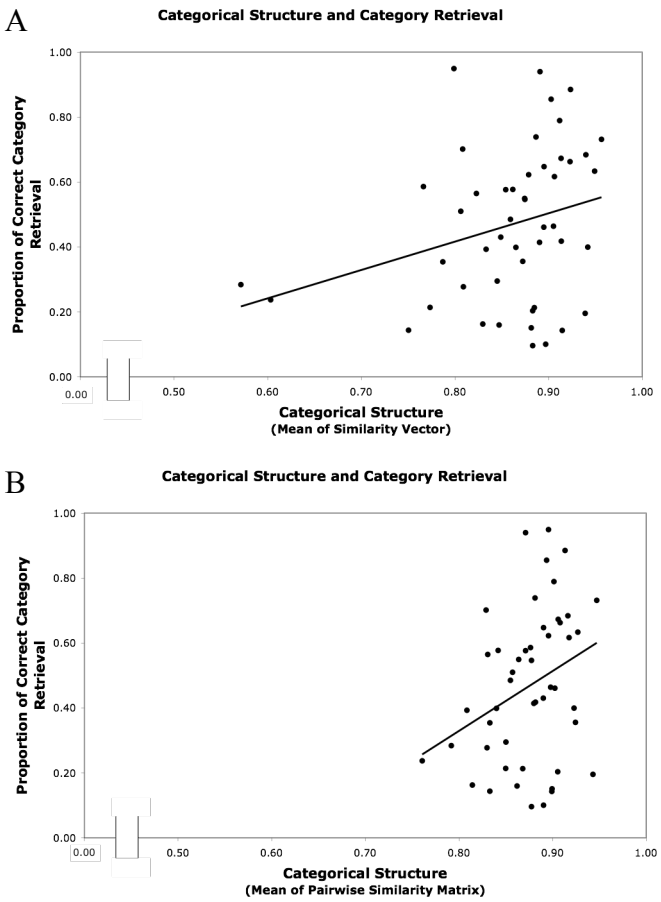

Figure 7. Effect of Categorical Structure on category retrieval: Panel A illustrates the relationship between Category Retrieval and Categorical Structure as measured by the mean of each category's similarity vector. Panel B illustrates the relationship between Category Retrieval and Categorical Structure as measured by the mean of each category's pairwise similarity matrix.

---

[16] Cohen (1988, pp. 78-83) provides the following guidelines for interpreting the size of correlational effects:

R^2 = .01 is a small effect
R^2 = .09 is a moderate effect
R^2 = .25 is a large effect

Format

Category Retrieval

I first examined the effect of prominence on category retrieval. Prominence was manipulated by two variables: Font Size and Order. There were no significant main effects of prominence on category retrieval (See Table 4). It appears that regardless of the prominence given to a category, they are all retrieved with an accuracy rate that varies around fifty percent.

Table 4

Prominence (manipulated by Font Size)	**Low** M (SE)	**Medium** M (SE)	**High** M (SE)	
Correct Category Retrieval	.497 (.044)	.441(.023)	.524 (.033)	F(2,58)=1.48, p> .05
Prominence (manipulated by Font Size and Order)	**Low** M (SE)	**Medium** M (SE)	**High** M (SE)	
Correct Category Retrieval	.486 (.043)	.445(.021)	.513 (.031)	F(2,58)=1.10, p> .05

Table 4. Effect of Prominence on category retrieval compares three different levels of prominence. The top panel summarizes the results for the Prominence factor when it is manipulated by Font Size. The bottom panel summarizes the results for the Prominence factor when it is manipulated by Font Size and Order.

Next, I examined the effect of layout on category retrieval. There was a main effect of layout ($F(3,87)= 5.81$, $p< .01$, $\omega^2 = .107$). There is some evidence that the Single Column List with Frequency Sorting transmitted more accurate information than the other layouts as shown in Table 5. A post-hoc analysis[17] showed that this layout had a slightly higher but significantly different accuracy rate when compared

[17] Bonferroni adjustments were performed on all post-hoc analyses, $\alpha_{adjusted} = \frac{\alpha}{4} = \frac{.05}{4} = .0125$.

to the Sequential Layout with Alphabetical Sorting and the Spatial Layout, but not when compared to the Sequential Layout with Frequency Sorting.

Table 5

	Correct Category Retrieval	
Layout	**M**	**SE**
Sequential Layout with Alphabetical Sorting	.456	.027
Spatial Layout	.435	.019
Sequential Layout with Frequency Sorting	.467	.021
Single Column List with Frequency Sorting	.532	.028

Table 5. Effect of Layout on category retrieval compares four different types of layout – Sequential Layout with Alphabetical Sorting, Spatial Layout, Sequential Layout with Frequency Sorting, Single Column List with Frequency Sorting.

Recognition

Table 6 summarizes results showing that recognition for words with a larger font size was significantly higher than for words with a smaller font size ($F(4,116)=96.17$, $p< .001$).

Table 6

	Correct Recognition	
	M	**SE**
Font Size 1 (High)	.822	.027
Font Size 2	.746	.024
Font Size 3	.599	.026
Font Size 4	.454	.027
Font Size 5 (Low)	.381	.026

Table 6. Effect of Font Size on accuracy of recognition compares five different levels of font size. Accuracy of recognition is given as a proportion of correct recognition.

There was no significant effect of layout on recognition for either targets or distractors ($F(3,87)<1$). As can be seen in Table 7, semantically related distractors had more false positives than unrelated distractors ($F(2,58)=292.27$, $p< .001$). There were three types of semantically related distractors: (a) the category label, with a .38

rate of false alarms; (b) a high semantic lure, with a .28 rate of false alarms; and (c) a medium semantic lure, with a .15 rate of false alarms[18]. Note the rate of false alarms for the category labels and note the rate of hits for words with the smallest font size – items that were not presented were falsely recognized at about the same rate as those that were presented, albeit those with the least favorable prominence.

Table 7

	Proportion of Hits		**Proportion of False Alarms**			
	Targets		**Sem. Related Distractors**		**Sem. Unrelated Distractors**	
Layout	**M**	**SE**	**M**	**SE**	**M**	**SE**
Sequential Layout with Alphabetical Sorting	.593	.027	.281	.036	.047	.017
Spatial Layout	.596	.021	.291	.033	.069	.024
Sequential Layout with Frequency Sorting	.608	.020	.284	.026	.042	.014
Single Column List with Frequency Sorting	.598	.025	.331	.028	.067	.020

Table 7. Effect of Layout on proportion of hits and false alarms compares four different types of layout – Sequential Layout with Alphabetical Sorting, Spatial Layout, Sequential Layout with Frequency Sorting, Single Column List with Frequency Sorting and compares targets and semantically related distractors (Sem. Related) and semantically unrelated distractors (Sem. Unrelated).

Discussion

In Experiment 1, evidence suggested that the two measures of categorical structure are related to category judgments. Both the means of the similarity vectors and the means of the similarity matrices displayed a moderately positive relationship with correct category retrieval. This result is an indication that categories with tighter structures are easier to identify than categories with looser structures.

[18] These rates were calculated by collapsing all tag clouds and layouts.

The effect of tag cloud format studied in Experiment 1 produced mixed results. Experiment 1 was unable to find an effect of prominence on correct category retrieval. This lack of evidence could be due to the low number of participants in this study. The recognition test that followed all categorization judgments provided some evidence that prominence should not be discounted. Fonts with larger sizes resulted in higher recognition rates than those with smaller sizes. This variable should be further investigated before conclusions on its influence on categorization can be drawn. There was a moderate effect of layout on correct category retrieval, where the Single Column List with Frequency Sorting appeared to contribute to higher accuracy rates. This is surprising when one takes into consideration that participants needed to scroll in order to see the entire tag cloud. However, this layout was not significantly different from the Sequential Layout with Frequency Sorting. This may suggest that perhaps sorting is driving the effect found.

There was no evidence that layout influenced the memory recognition tests. An interesting result from Experiment 1 was the increase of false positives for semantically related distractors exhibited. This increase may be a result of participants encoding the categories presented in each tag cloud. This encoding process may inhibit their ability to correctly discriminate new items that belong to the same category. This result is similar to the creation of false memories in studies using the Deese-Roediger-McDermott (DRM) paradigm. Subjects are given lists of words that are all associated with a critical word, which is not presented. For example, if the critical word is *sleep*, the list would consist of the twelve words most highly associated with sleep: bed, rest, awake, tired, dream, wake, snooze, blanket, doze,

slumber, snore, nap, peace, yawn, drowsy. The DRM effect refers to the false recall or recognition of the critical words. This false recall or recognition often exceeds that of other high associate distractors and even the correct recall or recognition of low-associate targets[19] (Deese, 1959; Roediger & McDermott, 1995). This is similar to the results of Posner and Keele (1970) who showed dot patterns to participants that were distortions from a prototypic pattern. During the test, participants recognized the prototype at a higher rate than patterns that had been presented during study.

Another possible explanation for this increase can be obtained from Alba and Hasher's (1983) prototypical schema theory of memory. The term schema is used to describe the general knowledge a person has regarding a specific domain. Schema theorists are particularly interested in how information is encoded, stored and retrieved. The theory presented by Alba and Hasher (1983) assumes four encoding processes: selection, abstraction, interpretation and integration that occur sequentially after information is presented to a person. During selection, people discriminate which information to use for representation. Information that has been selected is further reduced by abstraction. This process stores the meaning of the information but not its original syntactic or lexical format. The interpretation process uses previous knowledge to assist comprehension. Integration uses the inputs generated by the previous three processes to form a single memory representation. The selection process is built upon traditional schema theories, such as Owens, Bower, & Black's

[19] Deese (1959) investigated the relationship between the percentage frequency of occurrence of the stimulus word as an intrusion in recall and the mean percentage frequency of the stimulus word as an association to items on the list and obtained a Pearson correlation of .873 ($p < .01$) (p. 19). Similarly, I calculated the correlation between the percentage of false alarms of category labels and the means of similarity vectors and matrices and obtained significant and positive Pearson correlations, albeit considerably lower than the one reported by Deese (Similarity vectors: $r = .372$, $p < .01$; Similarity matrices: $r = .315$, $p < .05$).

(as cited in Alba & Hasher, 1983), that contend that the ideas that are most important to the theme of the information are given special attention and will be remembered best. In this study, sixty-six percent of the times a recognition test item resulted in a hit, its category had been correctly retrieved. The abstraction process assumes that humans are resource efficient; and consequently, they will store lexical expressions with the same meaning into a single abstracted expression. This in turn will result in incorrect recall or recognition of words that are semantically related to the originally presented words (Alba and Hasher cite several studies that show these type of behavior p. 208). Table 7 shows that semantically related words had a higher proportion of false positives than unrelated words. Sixty-three percent of the times a semantically related word resulted in a false positive, its category had been correctly retrieved.

A serious objection to the findings from Experiment 1 is that categorization judgments needed to rely on memory. The paradigm used in Experiment 1 presented the phases serially: the category retrieval phase came after the presentation phase. The next two experiments will use concurrent phases. Participants will be asked to make categorization judgments while concurrently observing the tag cloud.

Chapter 3: Experiment 2

Experiment 1 provided an initial investigation of the effects of categorical structure and tag cloud format on categorization. Categorization judgments were in the form of category retrieval tasks. Judgments were elicited after the presentation of each stimulus. In Experiment 2 judgments were elicited in the form of category verification and these tasks were performed during stimuli presentation.

Methods

Participants

University of Maryland undergraduate students (n=123; 49 males and 74 females) participated in Experiment 2 for course extra credit. All subjects had normal or corrected-to-normal vision. Participants were run individually in single sessions lasting approximately 60 minutes.

Materials

Materials included sixty-eight tag clouds and were presented in PC-based equipment using MediaLab and DirectRT research software (Jarvis, 2006). The formatting scheme was the same as the one used in Experiment 1. All layouts were able to fit in a single screen; no scroll bars were used for Experiment 2.

Category, Words and Tag Clouds

Sixty-eight categories and 816 words were obtained using the same software and document collection as in Experiment 1.

Four categories appeared per tag cloud; one related to an occupation and the other three were either hobbies or locations. The category seed was not used as stimulus. Ten words per category were used for each tag cloud, for a total of 40 words. The category verification phase used the category label and a semantically unrelated distractor for each category.

Seventeen tag clouds were created in all four layouts for a total of 68 tag clouds.

Design and Procedure

Experiment 2 consisted of three phases: a presentation phase, a category verification phase, and a confidence judgment phase. Initial instructions welcomed the participants and provided them with the definition of a tag cloud and a general example of one. A tag cloud was said to represent the general interests of a person who was named "the tag cloud owner". Subjects performed three practice trials and 128 experimental trials. Participants were informed that no data would be collected during the practice trials. After the practice, participants were given the opportunity to ask questions before the experimental trials started. Participants were further informed of the details of the experimental procedure. Each trial encompassed all three phases. The presentation phase consisted of the presentation of a tag cloud for a period of 5 s. Participants were told to try to make an inference as to what were the main interests of the tag cloud owner. Participants were not notified that there were four main interests (categories) listed in each tag cloud. The category verification phase presented a statement regarding the interests of the tag cloud owner (Figure 8 shows a screen shot with a sample true statement regarding the tag cloud presented).

This statement was presented concurrently with the tag cloud. Participants had to press the letter "T" if they thought the statement was true and the letter "F" if they thought the statement was false. This phase was self-paced with the restraint participants had to view each statement for a minimum of 5 s before they were allowed to respond true or false. There was no time limit for their answers. Response time data was collected. Pressing these letters would allow participants to advance to the confidence judgment phase.

Figure 8

Figure 8. Figure 8 presents an example stimulus during the category verification phase. The statement shown is true.

The confidence judgment phase asked participants how confident they were that they provided the correct answer. They were given a six-point scale that varied between 50% and 100% as shown in Figure 9. In order to respond, participants were told to press the number assigned to each confidence level. Additionally, instructions attempted to explain how confidence judgments are given. Following is an excerpt of such instructions:

If you guessed the veracity of the statement, then you should say you are 50% confident. Since there are only two possible answers (true or false) you have a 50% chance of being correct. You should press "1".

If you are absolutely sure that you are correct then you should say that you are 100% confident. You should press "6".

For the other percentages you should proceed as this example:
If you assign an 80% confidence level to your answer, this means you believe your answer has an 80% chance of being correct. You should press "4".

This phase was also self-paced. Once the confidence judgment phase was finished, the next trial appeared. These three phases repeated until all trials were completed. To avoid fatigue and automatic responses, two one-minute breaks were interlaced within the trials. Participants were told to relax their eyes within each break. At the end of the break, participants heard a tone and saw a different color screen that notified them the break was over. Participants had to click the space bar in order to continue with the next set of trials.

Figure 9

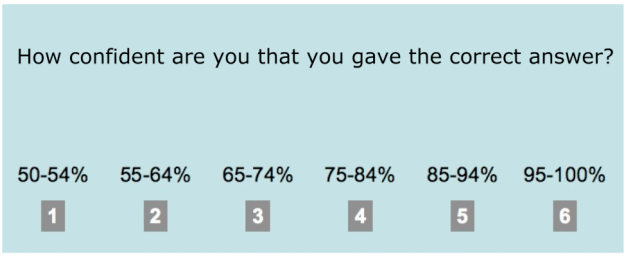

Figure 9. Figure 9 presents a clip from the screen participants viewed during confidence judgments.

The same factors as in Experiment 1 were manipulated in this experiment: categorical structure, prominence and layout. In Experiment 2, the measures of central tendency for the similarity vectors ranged between .531 and .957 and for the similarity matrices between .537 and .947. Categories were created automatically from the New York Time corpus by the Infomap software. This automated process

resulted in a significant number of categories with high central tendency measures of similarity. A potential problem with this negatively skewed distribution of categorical structure is the range restriction that it entails.

Practice was the same for all participants. It consisted of three trials that presented the same tag cloud in three different layouts: Sequential with Frequency Sorting, Spatial Layout, and Single Column List with Frequency Sorting. Among the three trials, two had true statements and one had a false statement. The experimental trials presented sixteen tag clouds that varied in layout. A counterbalancing scheme was used to diffuse any effects of category on layout (See Appendix B). For example, the "doctor" category (represented in Tag Cloud 1) was presented for Group 1 as a Sequential Layout with Alphabetical Sorting, for Group 2 as a Sequential with Layout Frequency Sorting, for Group 3 as a Spatial Layout and for Group 4 as a Single Column List with Frequency Sorting. There were eight trials per tag cloud: four trials contained true statements in which the category label was included and four trials contained false statements in which the distractor was included. Experimental trials were randomized for all participants.

Results

Data Analysis

Three dependent variables were analyzed: correct category verification, response time and confidence judgments. All analyses that involved response time

(RT) were performed on its logarithmic transformation[20]. For ease of interpretation, the results will be summarized using the non-transformed data.

Categorical Structure

I examined the correspondence between the dependent variables and the two measures of categorical structure: means of similarity vectors and similarity matrices[21].

Category Verification

The unit of interest for the correlation analysis is at the category level. Before averaging the scores from all participants and correlating that average score with each measure of categorical structure, I first tested whether the data was stationary. The non-significant Q-statistic in Table 8 shows that the data is stationary.

Table 8

	Categorical Structure	
	Mean of Similarity Vector	**Mean of Similarity Matrix**
Correct Category Verification		
Pearson Correlation (r)	.313*	.326**
df	62	62
R^2	.098	.106
Gamma Correlation (G)	.156	.182*
Q-statistic	125.39	138.12
df	122	122
	n.s.	n.s.

* p< .05; ** p< .01

Table 8. Relationship between Category Verification and Categorical Structure as measured by the mean of each category's similarity vector and by the mean of each category's pairwise similarity matrix. Q values were derived from Equation 2 and represent tests of homogeneity.

[20] Logarithmic transformations are recommend for positively skewed measures of response time (Kirk, 1995). This is true for the current data.

[21] These two measures of categorical structure were significantly correlated (r(62)= .85, p< .001). This makes sense because of the mathematical origin of both measures.

Figure 10

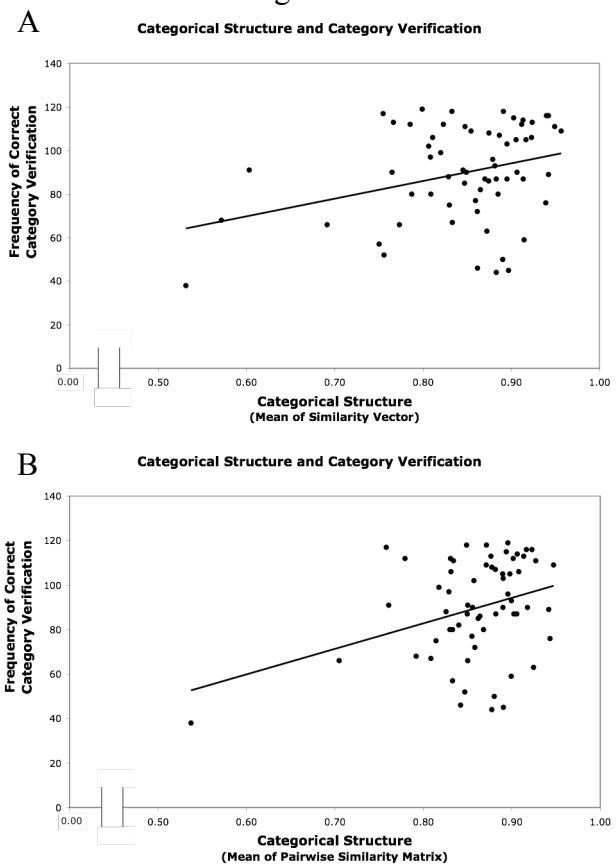

Figure 10. Effect of Categorical Structure on category verification: Panel A presents the relationship between Category Verification and Categorical Structure as measured by the mean of each category's similarity vector. Panel B presents the relationship between Category Verification and Categorical Structure as measured by the mean of each category's pairwise similarity matrix.

Table 8 and Figure 10 show positive relationships between categorical structure and category judgments. Both Pearson correlations were significant[22]. The strength of these relationships was moderate. The Gamma correlation for categorical structure as measured by the mean of each similarity matrix was significant, while the Gamma correlation for categorical structure as measured by the mean of each

[22] Cook's distances (Cohen, et al., 2003) were calculated to detect data points with unusual leverage. The findings are robust; new correlations obtained through analysis of influence statistics are in line with the results presented.

similarity vector was not. Note again that categorical structure appears to tighten slightly when pairwise similarity matrices are used to measure it.

Response Time

The unit of interest for the correlation analysis is at the category level. The non-significant Q-statistic in Table 9 shows that the data is stationary and can be averaged.

Table 9 and Figure 11 show a negative relationship between categorical structure and response time of category judgments. However, none of these correlations were significant.

Figure 11

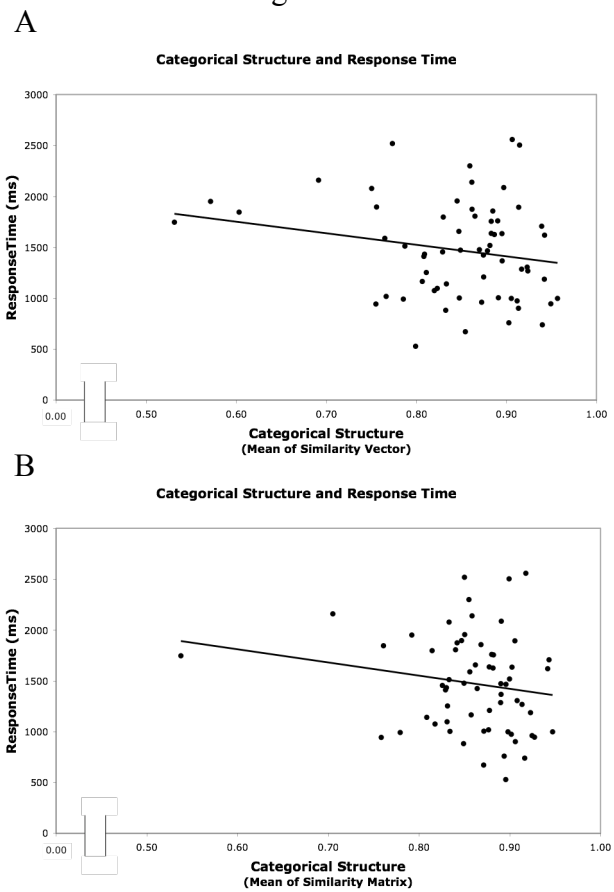

Figure 11. Effect of Categorical Structure on response time: Panel A presents the relationship between Response Time and Categorical Structure as measured by the mean of each category's similarity vector. Panel B presents the relationship between Response Time and Categorical Structure as measured by the mean of each category's pairwise similarity matrix.

Table 9

	Categorical Structure	
	Mean of Similarity Vector	**Mean of Similarity Matrix**
Response Time		
Pearson Correlation (r)	-.202	-.170
df	62	62
R^2	.041	.029
Gamma Correlation (G)	-.101	-.122
Q-statistic	106.34	119.17
df	122	122
	n.s.	n.s.

Table 9. Relationship between Response Time and Categorical Structure as measured by the mean of each category's similarity vector and by the mean of each category's pairwise similarity matrix. Q values were derived from Equation 2 and represent tests of homogeneity.

Confidence Judgments

A calibration analysis was performed to investigate the correspondence between confidence judgments and categorical structure[23]. The unit of interest for the correlation analysis is at the category level.

Brier (1950) proposed an overall measure of judgment accuracy – the Brier Score (PS). Low values indicate good judgment. A Brier Scores is given by

$$\overline{PS}(f,d) = \left(\frac{1}{N}\right)\sum_{i=1}^{N}(f_i - d_i)^2 \qquad [3]$$

In Equation 3, N is the number of judgments; f denotes the probability assigned to the target event the judge is trying to predict and d is the outcome index for the target event. If the target event occurs then $d = 1$; if the target event does not occur then $d = 0$. The target event was defined as "My preferred answer is correct".

[23] There were no significant correlations between mean confidence judgments and categorical structure (Categorical Structure as measured by the Mean of Similarity Vectors: r(62)= .128, p> .05; Categorical Structure as measured by the Mean of Similarity Matrices: r(62)= .101, p> .05).

In order to provide separate measures of different aspects of judgment accuracy, Murphy (1973) proposed a decomposition of the Brier Score:

$$\overline{PS}(f,d) = \bar{d}(1-\bar{d}) + \left(\frac{1}{N}\right)\sum_{j=1}^{J} N_j\left(f_j - \bar{d}_j\right)^2 - \left(\frac{1}{N}\right)\sum_{j=1}^{J} N_j\left(\bar{d}_j - \bar{d}\right)^2 \qquad [4]$$

In Equation 4, the mean outcome index, $\bar{d}$, is the proportion correct; N is the number of judgments; j indexes the response category; J is the number of response categories ($J = 6$ in this study); N_j is the number of responses in category j; f_j is the probability assigned by the judge; and $\bar{d}_j$ is the proportion of correct responses in category j.

The first term in Murphy's decomposition, $\bar{d}(1-\bar{d})$, is the variance of the outcome index (VOI). The second term in Murphy's decomposition, $\left(\frac{1}{N}\right)\sum_{j=1}^{J} N_j\left(f_j - \bar{d}_j\right)^2$, is a measure of the extent that the probabilistic judgments are well calibrated (i.e. that the proportion correct at each level of confidence equals the stated level of confidence). This term is known as reliability-in-the-small (Yates, 1982) or calibration index (CI) (Ariely et al., 2000). The third term in Murphy's decomposition, $\left(\frac{1}{N}\right)\sum_{j=1}^{J} N_j\left(\bar{d}_j - \bar{d}\right)^2$, is the resolution of the collection of forecasts. This term is known as the Murphy resolution (Yates, 1982) or discrimination index (DI) (Ariely et al., 2000).

Table 10

	Categorical Structure	
	Mean of Similarity Vector	**Mean of Similarity Matrix**
Brier Scores (PS)		
Pearson Correlation (r)	-.323**	-.344**
df	62	62
R^2	.104	.118
Gamma Correlation (G)	-.161	-.180*
Outcome Index Variance (VOI)		
Pearson Correlation (r)	-.241	-.204
df	62	62
R^2	.058	.042
Gamma Correlation (G)	-.163	-.174
Calibration Index (CI)		
Pearson Correlation (r)	-.321*	-.404**
df	62	62
R^2	.103	.163
Gamma Correlation (G)	-.161	-.183*
Discrimination Index (DI)		
Pearson Correlation (r)	.067	.080
df	62	62
R^2	.004	.006
Gamma Correlation (G)	.036	.080

* p< .05; ** p< .01

Table 10. Relationship between Calibration and Categorical Structure as measured by the mean of each category's similarity vector and by the mean of each category's pairwise similarity matrix.

Table 10 and Figure 12 show a significant negative relationship between categorical structure and PS[24]. Table 10 and Figure 13 show that once the Brier Scores are partitioned; both VOI and the CI have negative relationships with categorical structure. However, only the correlations for CI are significant. The non-significant correlation analysis found for VOI is attributed to a restriction in range. Equation 4 indicates that VOI is a transformation of the proportion of correct scores.

[24] Cook's distances (Cohen, et al., 2003) were calculated to detect data points with unusual leverage. New correlations obtained through analysis of influence statistics are in line with the results presented.

The results presented in Table 8 showed that categorical structure had a significant relationship with proportion of correct scores. The Gamma correlations for both PS and the CI with the means of the similarity vectors decreased and were no longer significant. No relationship was found between DI and categorical structure. All the correlations presented in Table 8 may be affected by restrictions in range. Values for the X-axis – categorical structure – were more abundant in the higher end of the scale. Values for the Y-axis – PS, VOI, CI and DI – were close to zero. These correlations may increase if issues of range restrictions can be resolved.

Figure 12

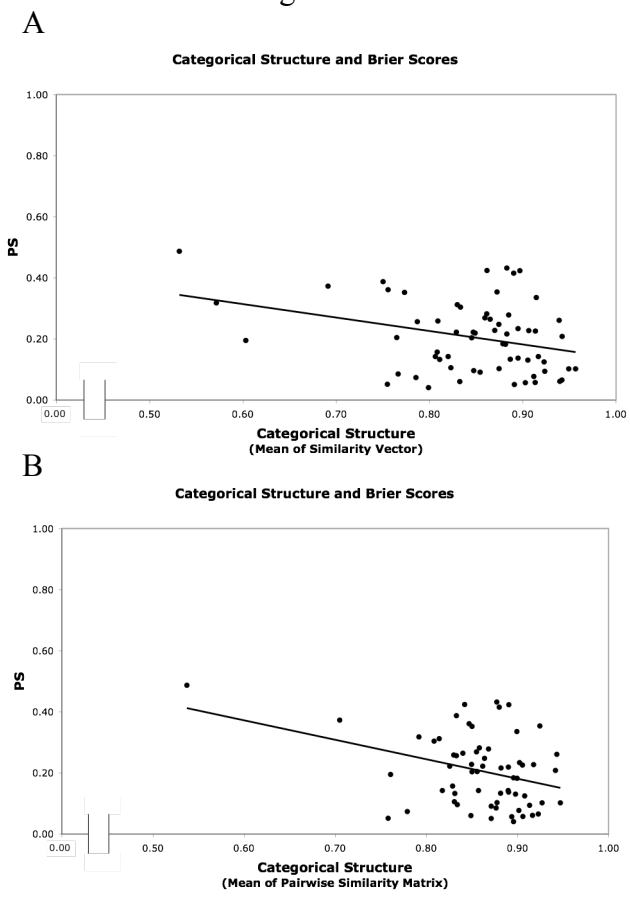

Figure 12. Effect of Categorical Structure on calibration: Panel A presents the relationship between Brier Scores (PS) and Categorical Structure as measured by the mean of each category's similarity vector. Panel B presents the relationship between Brier Scores (PS) and Categorical Structure as measured by the mean of each category's pairwise similarity matrix.

Figure 13

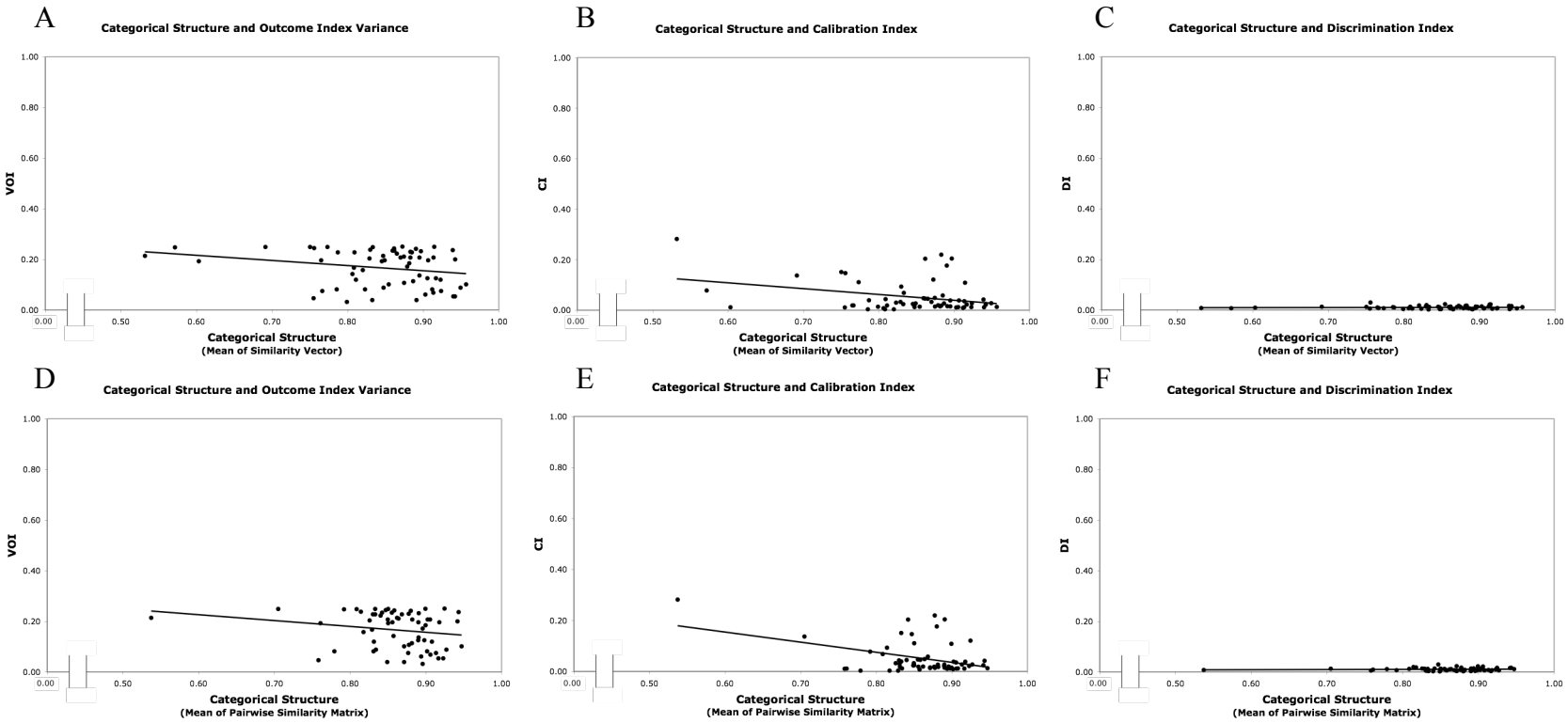

Figure 13. Effect of Categorical Structure on calibration: Panel A presents the relationship between Outcome Index Variance (OIV) and Categorical Structure as measured by the mean of each category's similarity vector. Panel B presents the relationship between Calibration Index (CI) and Categorical Structure as measured by the mean of each category's similarity vector. Panel C presents the relationship between Discrimination Index (DI) and Categorical Structure as measured by the mean of each category's similarity vector. Panel D presents the relationship between Outcome Index Variance (OIV) and Categorical Structure as measured by the mean of each category's pairwise similarity matrix. Panel E presents the relationship between Calibration Index (CI) and Categorical Structure as measured by the mean of each category's pairwise similarity matrix. Panel F presents the relationship between Discrimination Index (DI) and Categorical Structure as measured by the mean of each category's pairwise similarity matrix.

Format

Prominence

Table 11

Prominence (manipulated by Font Size)	**Low** M (SE)	**Medium** M (SE)	**High** M (SE)		
Correct Category Verification	.705 (.012)	.743 (.011)	.797 (.011)	F(2,244)=24.09 ***	ω^2 = .111
Response time (ms)	1514 (2.00)	1521 (1.08)	1216 (1.08)	F(2,244)=13.87 ***	ω^2 = .065
Confidence (%)	81 (0.90)	80 (0.80)	81 (0.80)	F(2,244)=4.04 *	ω^2 = .016
Prominence (manipulated by Font Size and Order)	**Low** M (SE)	**Medium** M (SE)	**High** M (SE)		
Correct Category Verification	.738 (.011)	.767 (.009)	.787 (.011)	F(2,244)=8.44 ***	ω^2 = .039
Response time (ms)	1500 (1.07)	1489 (1.07)	1236 (1.08)	F(2,244)=15.72 ***	ω^2 = .074
Confidence (%)	81 (0.80)	80 (0.80)	81 (0.80)	F(2,244)=1.67	ω^2 = .004

* p< .05; *** p< .001

Table 11. Effect of Prominence on category verification: compares three different levels of prominence. The top panel summarizes the results for the Prominence factor when it is manipulated by Font Size. The bottom panel summarizes the results for the Prominence factor when it is manipulated by Font Size and Order.

Prominence was manipulated by two variables: Font Size and Order. Table 11 shows a significant main effect of prominence on category verification. Higher prominence resulted in higher accuracies. There was a significant effect of prominence on response time. Higher prominence resulted in faster responses. The strength of association explained by Prominence was moderate[25]. There was a stronger effect on category verification when Prominence was manipulated by Font

[25] Cohen (1988, pp. 284-288) provides the following guidelines for interpreting strength of association:
ω^2 = .010 is a small association
ω^2 = .059 is a moderate association
ω^2 = .138 or larger is a large association

Size than when it was manipulated by both Font Size and Order. There was a small effect on confidence when Prominence was manipulated by Font Size and no effect when it was manipulated by both Font Size and Order. On average, participants were overconfident – their average confidence judgments were higher than their average accuracy rates[26].

Layout

Table 12

	Correct Category Verification		Confidence Judgments (%)	
Layout	**M**	**SE**	**M**	**SE**
Sequential Layout with Alphabetical Sorting	.725	.014	81	0.8
Spatial Layout	.722	.014	81	0.8
Sequential Layout with Frequency Sorting	.800	.014	78	0.9
Single Column List with Frequency Sorting	.812	.012	80	0.9

Table 12. Effect of Layout on category verification and confidence, compares four different types of layout – Sequential Layout with Alphabetical Sorting, Spatial Layout, Sequential Layout with Frequency Sorting, Single Column List with Frequency Sorting.

There was a main effect of layout on category verification ($F(3,366)= 14.28$, $p< .001$, $\omega^2 = .075$) and on confidence ($F(3,366)= 3.98$, $p< .05$, $\omega^2 = .018$) . A post-hoc analysis[27] showed that the Single Column List with Frequency Sorting and the Sequential Layout with Frequency Sorting transmit more accurate information than the other layouts (Sequential Layout with Alphabetical Sorting and Spatial Layout; see Table 12). Slightly smaller confidence judgments are given when interacting with

[26] Brier Scores were calculated for the three different levels of prominence. Participants increased their calibration as prominence increased (Low Prominence: PS= .206, Medium Prominence: PS= .200, High Prominence: PS=. 178).

[27] Bonferroni adjustments were performed on all post-hoc analyses, $\alpha_{adjusted} = \frac{\alpha}{4} = \frac{.05}{4} = .0125$.

the Sequential Layout with Frequency Sorting[28]. There were no effects of layout on RT ($F(3,366)= 1.72$, $p> .05$). All layouts had an average response time of approximately 1,400 ms.

Figure 14

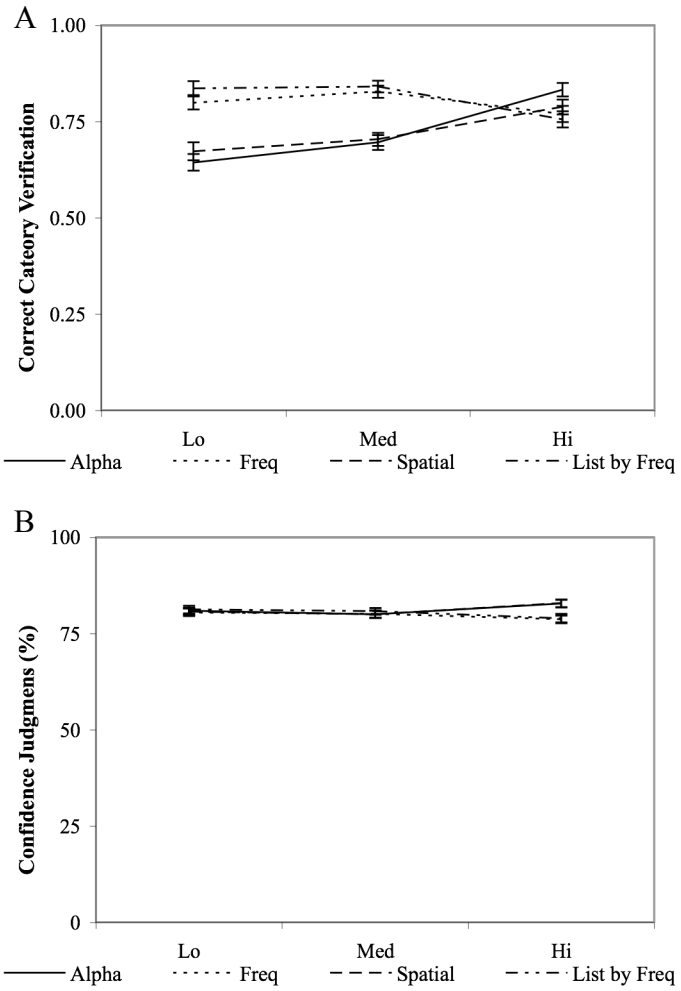

Figure 14. Panel A: Interaction between Layout and Prominence on the percent of correct category verification. Panel B: Interaction between Layout and Prominence on confidence judgments. Four different types of layout (Sequential Layout with Alphabetical Sorting (Alpha), Spatial Layout (Spatial), Sequential Layout with Frequency Sorting (Freq), Single Column List with Frequency Sorting (List by Freq)) and three levels of prominence (High (Hi), Medium (Med) and Lo (Low) are compared.

[28] Brier Scores were calculated for the four different layouts. Participants were slightly more calibrated when interacting with the frequency layouts (Sequential Layout with Alphabetical Sorting: PS= .135, Spatial Layout: PS= .131, Sequential Layout with Frequency Sorting: PS= .126, Single Column List with Frequency Sorting: PS=. 125).

There was a significant interaction between prominence and layout ($F(6,732)= 16.49$, $p< .001$) for category verification and for confidence judgments ($F(6,732)= 7.54$, $p< .001$). Although prominence had a significant influence on accuracy of category verification, it appears not to have influenced all layouts in the same degree. It can be observed in Figure 14 that prominence has an increasing effect on the accuracy rate for layouts with alphabetical and spatial sorting. Tests of simple effects[29] suggest this increasing influence for both alphabetical ($F(2,1464) = 26.27$, $p< .0125$) and spatial sorting ($F(2,1464) = 9.88$, $p< .0125$). However, for layouts with frequency sorting, prominence does not show this increasing influence. Tests of simple effects suggest a decreasing effect for the single column list with frequency sorting ($F(2,1464) = 6.39$, $p< .0125$) and no effect for the sequential layout with frequency sorting ($F(2,1464) = 2.31$, $p> .0125$). Tests of simple effects of prominence on confidence judgments were not significant.

Discussion

Experiment 2 provided additional evidence suggesting that the two measures of categorical structure are related to category judgments. Both the means of the similarity vectors and similarity matrices displayed a moderate and positive relationship with correct category verification. This result is an indication that categories with tighter structures are easier to authenticate than categories with looser structures. Another finding of Experiment 2 is that categorical structure and calibration have a moderate and negative relationship. The negative correlation with Brier Scores implies that participants are able to more accurately judge their

[29] Bonferroni adjustments were performed on all tests of simple effects, $\alpha_{adjusted} = \frac{\alpha}{4} = \frac{.05}{4} = .0125$.

performance with tighter structures. The correlation analyses using Murphy's (1973) partitions of the Brier Score suggest that the root of this relationship is provided by the calibration index and the variance outcome index. The first index is a local measure of calibration and the latter index is a measure of overall accuracy. Categorical structure has a correspondence with participants' reliability rather than their discriminability. It was surprising not to find a relationship between categorical structure and response time. Category membership verification of representative exemplars usually results in shorter response times than for non-representative exemplars (Mervis & Rosch, 1981). A tight category is akin to a collection of highly representative exemplars and would imply similar findings. The correlations for Experiment 2 were in the right direction (negative); but they were not significant.

Prominence had a significant effect on Experiment 2 as opposed to Experiment 1. Attention theory predicts such an effect, which made it puzzling not to see one in Experiment 1. Prominence resulted in higher categorization accuracy rates and faster response times. The stronger effect of the Font Size manipulation on category verification makes sense. Both the Sequential Layout with Alphabetical Sorting and the Spatial Layout rely solely on font size to transmit similarity measures. Larger fonts imply higher similarity scores. The effect of the Font Size and Order manipulation is somewhat smaller because it does not benefit all layouts equally. Both the Sequential Layout with Frequency Sorting and the Single Column List with Frequency Sorting benefit from both Font Size and Order to transmit similarity measures. Larger fonts and higher frequencies imply higher similarity scores.

There was a main effect of layout on correct category verification, where layouts with frequency sorting contributed to higher accuracy rates. This finding adds evidence to the suggestion presented in the previous section that sorting may be driving the effect. Confidence judgments for the Sequential Layout with Frequency Sorting were slightly smaller than for the other layouts.

Experiment 2 was a response to any objections that could be raised to Experiment 1, where categorization judgments relied on memory. The paradigm used in Experiment 2 set the categorization process to be performed concurrently with the presentation of the material. Categorization judgments were in the form of category verification. An objection that could be raised against Experiment 2 is that category verification may result in higher accuracies because participants are shown the category labels. This objection leads to Experiment 3 where categorization judgments will be in the form of category retrieval tasks. Participants are required to retrieve category labels from prior knowledge.

Chapter 4: Experiment 3

In Experiment 1, categorization judgments were in the form of category retrieval tasks. Judgments were elicited after the presentation of each stimulus. In Experiment 2, categorization judgments were in the form of category verification tasks. Judgments were elicited during stimuli presentation. Experiment 3 followed the pattern of Experiment 2; judgments were elicited during stimuli presentation. Categorization judgments were in the form of category retrieval tasks.

Methods

Participants

University of Maryland undergraduate students (n=119) participated in Experiment 3 for course extra credit. All subjects had normal or corrected-to-normal vision. Participants were run individually in single sessions lasting approximately 30 minutes.

Materials

Materials included fifty-two tag clouds and were presented in PC-based equipment using MediaLab research software (Jarvis, 2006). The formatting scheme was the same as the one used in Experiment 1 and 2. All layouts were able to fit in a single screen; no scroll bars were used for Experiment 3.

Category, Words and Tag Clouds

Sixty categories and 720 words were obtained using the same software and document collection as in Experiments 1 and 2.

Categories were of the following type: a profession, a sport, a hobby and a location. Each of these four categories appeared per tag cloud. The category seed was not used as stimulus. Ten words per category were used for each tag cloud, for a total of 40 words. There was one distractor per category to be used in the priming phase.

Fifteen tag clouds were created in all four layouts for a total of 60 tag clouds.

Design and Procedure

Eliciting categorization retrieval judgments implicated an open-ended response format where participants were allowed to input in a free-text field. Previous experience with open-ended responses, as in Experiment 1, suggested the need to reduce the variability of responses. In order to facilitate coding, a priming phase was designed for this experiment. Participants were presented with a list of six possible categories associated with the tag clouds they were about to see. Half of the list contained category labels and half contained distractors. The main purpose of this priming phase was intended to reduce response variability and not to investigate or manipulate responses based on priming.

Trials in Experiment 3 consisted of three phases: a priming phase, a presentation phase and a category retrieval phase. Initial instructions welcomed the participants and provided them with the definition of a tag cloud and a general example of one. A tag cloud was said to represent the general interests of a person who was named “the tag cloud owner”. Subjects performed three practice trials and

twelve experimental trials. Participants were informed that no data would be collected during the practice trials. After the practice, participants were given the opportunity to ask questions before the experimental trials started. Participants were further informed of the details of the experimental procedure. Each trial was composed of one initial priming phase and three presentation and category retrieval phases. Blocks of three tag clouds were used. Each block was associated with only one type of interest (profession, sport, hobby or location). The priming phase notified participants that the next group of people was interested in some of the following professions (sports, hobbies or locations). A list of six such interests was given. Three were category labels associated with the tag clouds and three were distractors. Initial instructions warned participants that the list contained correct and incorrect answers. The priming phase lasted 10 s. Three presentation and category retrieval phases followed – one for each tag cloud within the block. The presentation phase consisted of the presentation of a tag cloud for a period of 10 s. Participants were told to try to make an inference as to what were the main interests of the tag cloud owner. They were encouraged to look for the type of interest mentioned during priming (profession, sport, hobby or location). The category retrieval phase asked participants to enter the tag cloud owner's interest based on the tag cloud, type of interest within the block and the list of words used during priming. An example question would be: "Based on the list of words and the tag cloud, what profession is this person interested in?" After typing their answers, participants had to press the Enter-key in order to continue. Blank responses were not permitted, if participants did not know the answer they were instructed to type "I don't know". This phase was self-paced with no time

limit for responses. These two phases repeated until the block of three tag clouds was finished. The next trial would start with a priming phase and a block for three presentation and category retrieval phases. To avoid fatigue and automatic responses, two one-minute breaks were interlaced within the trials. Participants were told to relax their eyes within each break. At the end of the break, participants heard a tone and saw a different color screen that notified them the break was over. Participants had to click the space bar in order to continue with the next set of trials.

Practice was the same for all participants. It consisted of three trials; the first two trials had two tag clouds per block and the third trail had three tag clouds per block. The practice trials differed from the experimental trials with respect to feedback. After each trial, participants were informed what were the correct and incorrect answers. Correct answers were further stressed by highlighting the words in the tag cloud associated with such answers (See Figure 15). Feedback was provided to illustrate the task. Participants were notified that feedback would only be given during practice.

The same factors as in Experiment 1 and 2 were manipulated in this experiment: categorical structure, prominence and layout. In Experiment 3, the measures of central tendency for the similarity vectors ranged between .572 and .957 and for the similarity matrices between .732 and .947. Categories were created manually using the Infomap software and the New York Time corpus. A seed was used to create an initial list of words representing a category. Then these words were used iteratively as seeds to obtain additional words with greater semantic distances from the original seed. This process was done in order to create stimuli with a varying

degree of categorical structure and to avoid the skewed distribution that appeared in Experiment 2. However, the concern of range restriction is still valid for Experiment 3. It was quite difficult to obtain categories with low categorical structures.

Figure 15

Figure 15. Figure 15 presents an example of feedback given during practice in Experiment 3.

The experimental trials presented twelve tag clouds that varied in layout. There were four presentations per tag cloud – one presentation per category. A different layout was used for every presentation. Thus, participants would see a specific tag cloud in each of the four layouts. Experimental trials were randomized for all participants. A counterbalancing scheme was used to diffuse any effects of type of interest and category on layout (See Appendix B). For example, when profession was primed the "doctor" category (represented in Tag Cloud 1) was presented for Group 1 as a Sequential Layout with Alphabetical Sorting, for Group 2 as a Single Column List with Frequency Sorting, for Group 3 as a Sequential with Layout Frequency Sorting, and for Group 4 as a Spatial Layout.

Results

Data Analysis

Two dependent variables were analyzed: correct category retrieval and response time. A score was assigned to measure correct category retrieval. The scoring procedure was the same as described in Experiment 1. Four judges performed this scoring procedure. The inter-rater reliability was high (Average Measure Intraclass Correlation Coefficient = .964). All analyses that involved response time (RT) were performed on its logarithmic transformation. For ease of interpretation, the results will be summarized using the non-transformed data.

Categorical Structure

I examined the correspondence between the percent of correct category retrieval and the two measures of categorical structure: means of similarity vectors and similarity matrices[30].

Category Retrieval

The unit of interest for the correlation analysis is at the category level. Before averaging the scores from all participants and correlating that average score with each measure of categorical structure, I first tested whether the data was stationary. The non-significant Q statistic in Table 13 shows that the data is stationary and can be averaged.

[30] These two measures of categorical structure were significantly correlated ($r(46)= .826$, $p< .05$). This makes sense because of the mathematical origin of both measures.

Table 13 and Figure 16 show positive relationships between categorical structure and category judgments. Both the Pearson and Gamma correlations were significant[31]. The strength of these relationships was moderate. Note again that categorical structure appears to tighten slightly when pairwise similarity matrices are used to measure it.

Figure 16

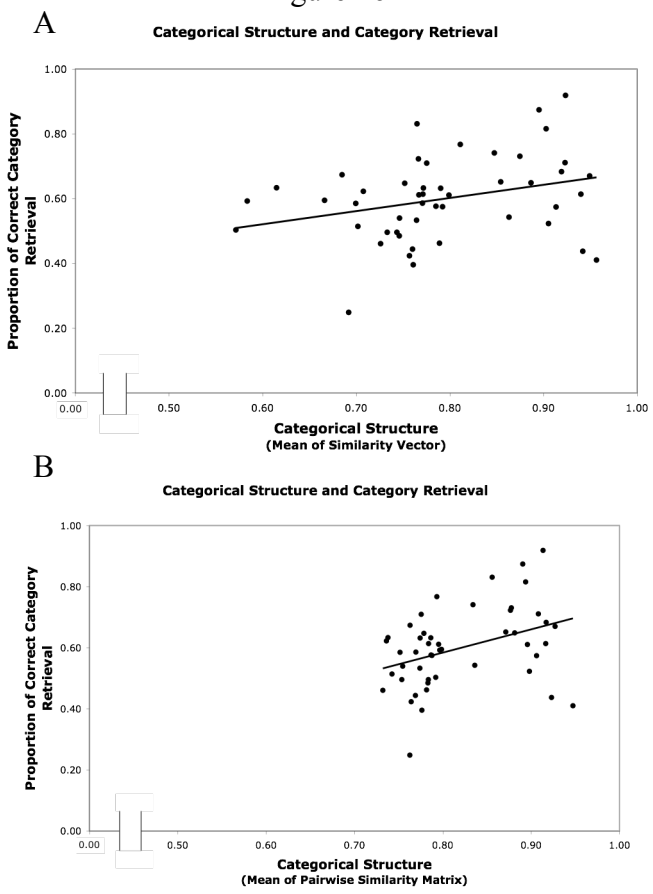

Figure 16. Effect of Categorical Structure on category retrieval: Panel A presents the relationship between Category Retrieval and Categorical Structure as measured by the mean of each category's similarity vector. Panel B presents the relationship between Category Retrieval and Categorical Structure as measured by the mean of each category's pairwise similarity matrix.

[31] Cook's distances (Cohen, et al., 2003) were calculated to detect data points with unusual leverage. New correlations obtained through analysis of influence statistics are in line with the results presented.

Table 13

	Categorical Structure	
	Mean of Similarity Vector	**Mean of Similarity Matrix**
Correct Category Retrieval		
Pearson Correlation (r)	.299*	.376**
df	62	62
R^2	.089	.141
Gamma Correlation (G)	.222*	.261**
Q-statistic	90.15	113.92
df	118	118
	n.s.	n.s.

* p< .05; ** p< .01

Table 13. Relationship between Category Retrieval and Categorical Structure as measured by the mean of each category's similarity vector and by the mean of each category's pairwise similarity matrix. Q values were derived from Equation 2 and represent tests of homogeneity.

Response Time

The unit of interest for the correlation analysis is at the category level. The non-significant Q-statistic in Table 14 shows that the data is stationary and can be averaged.

Table 14

	Categorical Structure	
	Mean of Similarity Vector	**Mean of Similarity Matrix**
Response Time		
Pearson Correlation (r)	-.318*	-.471**
df	46	46
R^2	.101	.222
Gamma Correlation (G)	-.241**	-.309***
Q-statistic	102.34	142.47
df	118	118
	n.s.	n.s.

* p< .05; ** p< .01; *** p< .001

Table 14. Relationship between Response Time and Categorical Structure as measured by the mean of each category's similarity vector and by the mean of each category's pairwise similarity matrix. Q values were derived from Equation 2 and represent tests of homogeneity.

Table 14 and Figure 17 show a negative relationship between categorical structure and response time of category judgments. Both the Pearson and Gamma correlations were significant[32]. The strength of these relationships was moderate for vectors and strong for matrices.

Figure 17

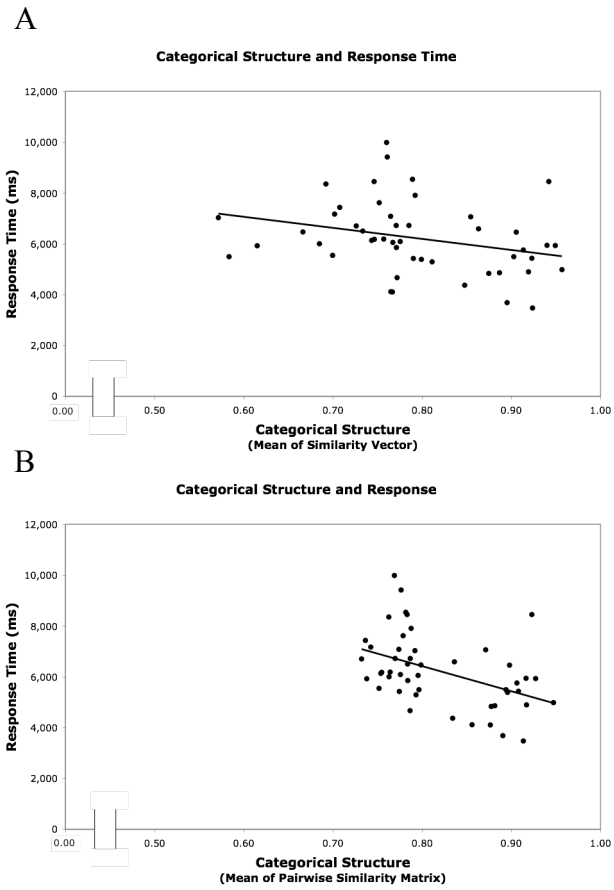

Figure 17. Effect of Categorical Structure on response time: Panel A presents the relationship between Response Time and Categorical Structure as measured by the mean of each category's similarity vector. Panel B presents the relationship between Response Time and Categorical Structure as measured by the mean of each category's pairwise similarity matrix.

[32] Cook's distances (Cohen, et al., 2003) were calculated to detect data points with unusual leverage. New correlations obtained through analysis of influence statistics are in line with the results presented.

Format

Prominence

Prominence was manipulated by two variables: Font Size and Order. Table 15 shows a significant main effect of prominence on category retrieval. Higher prominence resulted in higher accuracies. There was a significant effect of prominence on response time. Higher prominence resulted in faster responses. The strength of association explained by Prominence was large for accuracy and moderate for response time.

Table 15

Prominence (manipulated by Font Size)	**Low** M (SE)	**Medium** M (SE)	**High** M (SE)		
Correct Category Retrieval	.552 (.017)	.587 (.014)	.689 (.018)	F(2,236)=28.12 ***	ω^2 = .132
Response time (ms)	6339 (1.05)	6310 (1.05)	5521 (1.05)	F(2,236)=11.54 ***	ω^2 = .056
Prominence (manipulated by Font Size and Order)	**Low** M (SE)	**Medium** M (SE)	**High** M (SE)		
Correct Category Retrieval	.558 (.015)	.587 (.013)	.673 (.015)	F(2,236)=23.93 ***	ω^2 = .114
Response time (ms)	6295 (1.05)	6237 (1.04)	5535 (1.04)	F(2,236)=14.03 ***	ω^2 = .068

* p< .05; *** p< .001

Table 15. Effect of Prominence on category retrieval: compares three different levels of prominence. The top panel summarizes the results for the Prominence factor when it is manipulated by Font Size. The bottom panel summarizes the results for the Prominence factor when it is manipulated by Font Size and Order.

Layout

There was no effect of layout on category retrieval nor on response time ($F(3,354) < 1$). All layouts had an average accuracy rate of 60%. All layouts had an average response time of approximately 6,000 ms. There were no significant interactions between prominence and layout for either category retrieval or response time.

Discussion

Experiment 3 provided additional evidence suggesting that the two measures of categorical structure are related to category judgments. Both the means of the similarity vectors and similarity matrices displayed a moderate and positive relationship with correct category retrieval. This result is an indication that categories with tighter structures are easier to identify than categories with looser structures. Both the means of the similarity vectors and similarity matrices displayed moderate and strong negative relationships with response time. This result is an indication that categories with tighter structures are identified faster than categories with looser structures. The significant relationship between categorical structure and response time found in Experiment 3, as opposed to Experiment 2, is predicted by theory. Studies investigating categorical structure and typicality have reported a decrease in categorization time with the increase of structure and typicality (Rosch, 1975; Rosch, et al., 1976; Rips et al. 1975). Experiment 3 had higher average response times than Experiment 2. This is a product of the nature of the different category judgment tasks used for each experiment. In Experiment 2, participants performed a category verification task: for each trial, the category label was presented and they had to respond true or false. In Experiment 3, participants performed a category retrieval task: participants had to retrieve from long term memory the category label associated with each trial. The difference between Experiment 2 and 3 in average response time and correlation results may be due to the nature of the different tasks employed.

Experiment 3 provided additional evidence on the influence of prominence. Prominence resulted in higher categorization accuracy rates and faster response times.

Again, there was a stronger effect of the Font Size manipulation. The effect of the Font Size and Order manipulation is somewhat smaller because it does not benefit all layouts equally.

There was no effect of layout on correct category retrieval or response time. This lack of evidence is at odds with the previous two experiments that found effects of layout on category judgments. A possible explanation for this null effect could be a reduction in the familiarity participants may have perceived with a particular tag cloud's layout. To illustrate this point, please refer to Appendix B. In Experiment 1, a participant in Group 1 was required to look at Tag Cloud 4 presented in a Spatial Layout for 30 s before being allowed to answer. In Experiment 2, a participant in Group 1 was required to look at Tag Cloud 3 presented in a Spatial Layout for a minimum of 10 s before being allowed to answer. This was performed eight times – four for each category and four distractors. In Experiment 3, a participant in Group 1 was required to look at Tag Cloud 1 for a minimum of 10 s before being allowed to answer. This was performed four times – four for each category. However for each of those four times, Tag Cloud 1 was presented in one of the different four layouts. The familiarity the variable layout may have provided for a particular tag cloud in Experiments 1 and 2 may not have been transmitted in Experiment 3. Every time a particular tag cloud appeared, it did so in a different layout.

This experiment was designed as a result of any objections that could have been raised against Experiment 1 due to its reliance on memory. Experiments 2 and 3 complement each other; categorization judgments have been elicited in terms of category verification (Experiment 2) and category retrieval (Experiment 3).

Chapter 5: General Discussion

The purpose of this research was to examine how categorical structure and tag cloud format affect categorization. The influence of the degree of within-category association on judgments of category membership was examined. Semantic distances were calculated to measure similarity between category members. These distances were obtained from a latent semantic analysis performed on a general text corpus. In order to represent categories, exemplars and their inter-item similarities were used as coordinates of similarity vectors and matrices. Categorical structure was operationalized as a central tendency measure of said similarity vectors and matrices. Regarding tag cloud format, different known layouts used for tag clouds were compared to investigate their effect on categorization judgments. Other formats that were investigated were font size and sorting order. These formats were suggested to determine the perceived prominence of the presented stimuli.

Several general findings from the experiments are of importance. First, a relationship between categorical structure and categorization was found. Loose structures result in lower rates of categorization accuracy and tighter structures result in higher rates of categorization accuracy. Second, prominence positively influenced categorization. Prominence was operationalized by manipulating font size and sorting order. Larger fonts contribute to higher rates of categorization accuracy. The sorting of category exemplars played a role on categorization. Layouts with frequency sorting produce more accurate judgments. Frequency was operationalized by semantic distance to the category label. Terms that were more similar to the category label

were assigned higher frequency. Font size was operationalized in this same manner. Terms more similar to the category label were assigned larger fonts. Similarity is the driving element of these results.

Table 16

		Experiment 1		Experiment 2		Experiment 3	
		Vector	Matrix	Vector	Matrix	Vector	Matrix
Categorization Judgments	***r***	+ ●	+ ●	+ ●	+ ●	+ ●	+ ●
	G	+ ●	+ ●	+ ○	+ ●	+ ●	+ ●
Response Time	***r***			- ○	- ○	- ●	- ●
	G			- ○	- ○	- ●	- ●
Calibration: PS, CI	***r***			- ●	- ●		
	G			- ○	- ●		

● significant results; ○ non significant results

Table 16. Summary of correlational analyses between Categorical Structure and Categorization Judgments, Response Time and Calibration. PS= Brier Scores, CI= Calibration Index, + = positive correlations, - = negative correlations.

The influence of categorical structure on categorization was evaluated by its relationship with category retrieval, category verification, calibration and response time. A summary of the direction and statistical significance of these relationships is presented in Table 16. Categorical structure showed a positive relationship with category retrieval (Experiments 1 and 3) and category verification (Experiment 2, except for the Gamma correlation for the Similarity Vectors); tighter structures result in higher accuracies. Categorical structure showed a negative relationship with response time (Experiment 2: negative correlation but not statistically significant, Experiment 3: significant negative correlation); tighter structures result in faster

categorization judgments. Categorical structure showed a negative relationship with calibration (Experiment 2, except for the Gamma correlation for the Similarity Vectors); tighter structures result in better judgments of participants' performance.

Table 17

	Experiment 1		Experiment 2		Experiment 3		Effect Size ω^2		
	Prom	Layout	Prom	Layout	Prom	Layout	Prom		Layout
							FS	FSO	
Categorization Judgments	○	●	●	●	●	○	.121	.076	.081
Response Time			●	○	●	○	.061	.071	n.s.
Confidence			●○	●			.016	n.s.	.018

● significant results; ○ non significant results;
●○ one variable with significant results and one without

Table 17. Summary of the effects of format on Categorization Judgments, Response Time and Confidence. Effect Sizes are weighted averages of a significant variable's individual effect size in each experiment. Prom = Prominence, FS = Font Size Manipulation, FSO = Font Size and Order Manipulation.

The influence of tag cloud format on categorization was evaluated by examining the effects of prominence (font size and sorting order) and layout on category retrieval, category verification, confidence judgments and response time. A summary of the effects of format is presented in Table 17. The effect sizes presented in the summary table are weighted averages of a significant variable's individual effect size in each experiment; the weights are the corresponding number of participants. The variables were analyzed as fixed effects; thus, their effect size is descriptive of a participant's performance in this study. Prominence influenced

category retrieval (Experiment 3[33]) and category verification (Experiment 2); higher prominence results in higher accuracy. Prominence influenced response time (Experiment 2 and 3); higher prominence results in faster responses. The effect of prominence on confidence judgments is unresolved. An effect was found for font size but not for order. In addition, the direction of this effect was not interpretable; both low and high prominent categories resulted in slightly higher confidence judgments than categories with medium prominence. Layout influenced category retrieval (Experiment 1) and category verification (Experiment 2); layouts with frequency sorting result in higher accuracy. However, there was no effect of layout on category retrieval in Experiment 3. It was suggested that this null effect could be explained by a reduction in familiarity with the layouts assigned to specific tag clouds. Layout did not influence response time (Experiment 2 and 3). Layout influenced confidence judgments (Experiment 2); the sequential layout with frequency sorting resulted in smaller confidence judgments.

Theoretical Implications

The measure of categorical structure presented in this study was obtained through an automated analysis of the English language. Similarity between two words is defined by the usage and distribution of those words in a linguistic context. The co-occurrence statistics of a set of words in a general text corpus translate into a measure of semantic distance. For each category, semantic distances between category members and the category label can be computed. These distances can be arranged as a similarity vector with words as coordinates or as a matrix of pairwise similarities. In

[33] The non-significant influence of prominence on category retrieval on Experiment 1 is ascribed to its low statistical power.

this study, categorical structure is defined as a measure of central tendency for both vectors and matrices.

A vector represents categorical structure by computing the relationship between members and the category label. A matrix represents categorical structure by computing the relationship within all members of the category – including the label. A theoretical question that researchers usually face relates to the selection of measures. A matrix's mean of pairwise similarities is a comprehensive measure; it takes into account all elements of a class and all relationships between these elements. A vector's mean of similarities is a practical measure; it requires simpler computations, not all inter-item similarities are required. It is interesting to note that one measure of categorical structure displays more sensitivity with the same stimuli set as opposed to the other measure (Figures 7, 10-13, 16, 17). Similarity vectors show a wider range for the same stimuli than matrices. Is a more sensitive measure superior to a less sensitive measure? The first step in answering this question is to investigate if the higher sensitivity found in the similarity vectors accurately represents the measured construct. For all categorical structure analyses, the conclusions resulting from the Gamma and Pearson correlations were in accordance except for three occasions (Experiment 2: analyses of Category Verification, Brier Scores and Calibration Index). Gamma correlations for the means of similarity vectors led to non-significant conclusions, while Pearson correlations led to significant correlations. In addition, the less sensitive measure – means of similarity matrices – is explaining *slightly* more variance than the more sensitive measure (the increase of R^2 was between .01 and .12). This small increase would suggest that

either measure – means of vectors or means of matrices – can be used to represent categorical structure. However, the discrepancy found in some of the correlation analyses would suggest using similarity matrices would produce more consistent results.

This study has presented two measures of categorical structure that have been able to explain a moderate amount of variance in categorization judgments. The advantage of using the proposed measures lies on how they are obtained. Categorical structure for artificial stimuli is usually defined a priori by the experimenter with the application of a predetermined rule (Homa & Cultice, 1984; Rosch, et al. 1976). Categorical structure for natural stimuli is usually obtained by a post hoc analysis of ratings of semantic distance (Rips et al., 1975). All experiments in this study utilized natural stimuli, in the sense that words representing general interests were used. Findings relating to the categorical structure variable were in line with theoretical predictions. Categorical structure values were obtained by a latent semantic analysis of The New York Times corpus.

The use of Latent Semantic Analysis (LSA) to obtain measures of semantic distance and categorical structure is superior for several reasons. First, the measures obtained are not subjective – they do not require explicit human judgments; they are calculated based on how the stimuli are distributed in a text corpus – similarity and meaning of words are implicitly expressed by the authors of the text being analyzed. Second, narrow measures are possible; specific domains can be investigated if the researcher has access to a corpus relevant to that particular domain. Third, the calculations are effortless; LSA software packages are widely available. However,

there are weaknesses with this application of LSA. Using a deterministic measure of similarity – as the one obtained from a static corpus – reduces the inherent noise related with people. There will be some noise present as a consequence of the number of authors associated with the corpus or as a result of a change in writing style over time. Multidimensional analyses of participant's similarity ratings are capable of capturing more noise than LSA. Another possible weakness is that the measure of similarity obtained through LSA is dependent on the corpus used to construct the latent space. It could be argued that the measures obtained are only applicable to the population for which the corpus is intended. For example, the New York Times corpus provides different similarity measures than the Wall Street Journal corpus. I believe that this last weakness of LSA could be construed as a strength, as it illustrates the flexibility of the similarity construct. Judgments of similarity are dependent on the context in which they are elicited.

Applications of Research

Depending on the contexts in which they appear, tag clouds can support user tasks ranging from locating specific items to providing an overview of the underlying content. Such tasks can include: (a) *Search*. Locating a specific term or a desired concept; (b) *Browsing*. Using tag clouds as a means to browse; (c) *Impression Formation*. Looking at the tag cloud as a means to form a general impression of the underlying data set; (d) *Recognition/Matching*. Recognizing which of several sets of information or entities a tag cloud is likely to represent. An example is determining which of two John Smith's is the one you met at a conference based on their personal

tag clouds (Rivadeneira et al., 2007). The results found in this study apply to situations that do not require precise navigation.

The prominence and layout of tag clouds were used to investigate how format affects categorization judgments. The study found that highly prominent categories resulted in higher accuracies and faster response times. Prominence was operationalized by font size and order. Layouts sorted by frequency produced higher accuracy rates when compared to the other layouts.

One of the goals of this study was to provide a set of guidelines on how to visually present tags so that the information they represent could be accurately transmitted. The results suggest the following recommendations:

1. *Take advantage of Font Size.* More important tags should be represented with larger fonts.

Font Size is the formatting variable that generated higher degrees of influence. When prominence was uniquely manipulated by Font Size, it had a moderate-to-large effect on categorization accuracy and a moderate effect on response time[34] (See Table 17). Note that the designer should consider the ratio of change used to increment the size of a word from one level to the next. If the tag cloud is populated by a majority of words with large fonts, the effect of font size may dissipate. This study used 2:1 increment between levels.

2. *Use layouts with frequency sorting.* More important tags should be listed first in a sequence.

[34] See Footnote 25 for guidelines for interpreting strength of association.

When Sorting Order was included in the prominence manipulation, it had a moderate effect on categorization accuracy and a moderate effect on response time (See Table 17). In addition, the post-hoc tests mentioned in both Experiments 1 and 2 suggested that layouts ordered by frequency resulted in higher categorization accuracies.

3. *Maintain consistency.* If the tag cloud does not change over time, keep it consistent[35].

The effect of Layout on categorization accuracy was moderate when consistency was maintained (Experiments 1 and 2) and it was eliminated when tag clouds appeared in different formats (Experiment 3). There was no effect of Layout on response time.

The recommendations presented are for tag clouds used in overview tasks. A specific search task, for example, would benefit more from an alphabetical sorting than a frequency sorting (Halvey & Keane, 2007).

The guidelines provided are aimed for tag cloud designers and do not include information regarding categorical structure. It is assumed that the designer is not responsible for the content associated with the tag cloud. However, if the designer is also responsible for a semantic analysis of the underlying content, there is one last recommendation:

4. *Increase categorical structure.* Reduce the number of unique tags by combining synonyms or highly semantically related terms. It may be better to have less tags with high degrees of within-category association than more tags with low degrees of within-category association.

[35] I have met tag cloud designers that have purposefully added an element of randomness to their tag cloud algorithms in order to produce a sense of freshness to their tag clouds, not realizing the adverse effect this decision entails.

The strength of the relationship between categorical structure and categorization accuracy was moderate, as evidenced in the R^2's for all experiments. It is hypothesized that the size of these correlational effects are smaller in this study than in actuality. The measures of categorical structure for the stimuli experienced range restriction. It was quite difficult to feed the Infomap algorithm (Computational Semantics Lab from Stanford University, n.d. b) with seeds that would result in categories with low degrees of categorical structure.

Categorization, Tag Clouds and Social Perception

A parallel exists between research in impression formation and categorization. Researchers in impression formation present participants with stimuli, which are persons described by a list of personality traits. Participants are usually requested to give a rating in some measure of likeability or a judgment of fit for a particular context (Fiske, Neuberg, Beattle, & Milberg, 1987). Researchers in categorization provide participants with stimuli, which are described by a set of dimensions. Participants are requested to give a judgment of fit for a particular category. In fact, some categorization experiments use stimuli with personality traits and ask participants to form impressions before classification (e.g. Experiments 2 and 6 in Medin et al., 1987).

Generally, when first impressions are formed, a number of perceptual cues are available for a person to process. Taylor, Fiske, Etcoff, and Ruderman (1978) proposed a two-step process for social perception. First, people process information about social groups by categorizing the group members as a way of organizing information about them. Categorization reduces within-group differences and reveals

between-group differences. Following categorization, the behavior of members of the new subgroups is interpreted in stereotyped terms. “Stereotypes can be thought of as attributes that are tagged to category labels (e.g., race, sex) and imputed to individuals as a function of their being placed in that category, much as attributes of other categories are imputed to objects placed in those categories” (Taylor et al., p. 792). Processes that underlie person perception and impression formation have much in common with those that underlie object perception, particularly categorization processes.

The use of tag clouds in social computing is quite beneficial. In addition to providing a navigation mechanism, tag clouds provide an overview of the information they represent. A tag cloud can provide a meaningful reflection of the topics of interest and/or expertise of a particular user. Hearst and Rosner (2008) conducted a series of interviews with visualization designers and found evidence that the primary reason people use tag clouds is because of their perceived social component. The interviewees believed that tag clouds were able to communicate what a person or a group of people is interested in.

Tag clouds can be regarded as the stimuli presented in an impression formation task. The findings of this study are applicable for person perception. High categorical structure, high prominence and layouts with frequency sorting will provide more accurate perceptions of the individual represented in the tag cloud. In the same manner, the findings are applicable to issues regarding impression management. Impression management is the process through which people try to shape the impressions of others. If the impression manager wants to transmit a

specific idea in a tag cloud, the idea should have a high categorical structure and be represented by tags listed first in a sequential layout and with high prominence.

Future Research

There are several lines of research still pending. Other possible measures of categorical structure based on semantic distance can be investigated. Several come to mind: the determinant of a similarity matrix, the cohesion factor of a similarity matrix, the surface area of the category represented in multidimensional space and the number of factors a category has (based on a factor analysis of the similarity matrices). The data obtained for this study could be used for follow up analyses with these suggested measures.

An interesting question refers to the effect of judging category membership in the context of multiple categories. The study presented tag clouds with four categories. Participants in categorization retrieval and categorization verification studies are usually asked to judge one category at a time. Participants in sorting and classification studies are presented with exemplars from multiple categories. The accuracy results reported in all three experiments are considerably lower than those reported in categorization retrieval and categorization verification studies. Perhaps the presence of other categories reduced participants' performance. An experiment can be designed to compare stimuli with multiple categories versus one category. It is predicted that categorical structure and tag cloud format will influence categorization accuracy in the same manner as this study; however, the accuracy rate will increase for tag clouds with only one category.

Several studies performed by Homa and colleagues provided evidence that the degree of positive transfer to new instances is influenced by category size, resulting in superior transfer for the categories defined by a larger number of stimuli (Homa, Cross, Cornell, Goldman & Shwartz, 1973; Homa & Vosburgh, 1976). Similar results were found in a study performed by Hintzman (1988, Experiment 1), where accurate recognition of studied category members and false recognition of lures from that category increased as a function of category size. It would be interesting to find a categorical structure and category size trade off: how big does a category with low structure have to be so that it reaches the same accuracy level as one with a high structure?

The present research is not an attempt to construct a model of semantic memory. However, the fitting of models to the data of the present research is considered a further stage of investigation.

Appendix A

Figure A1

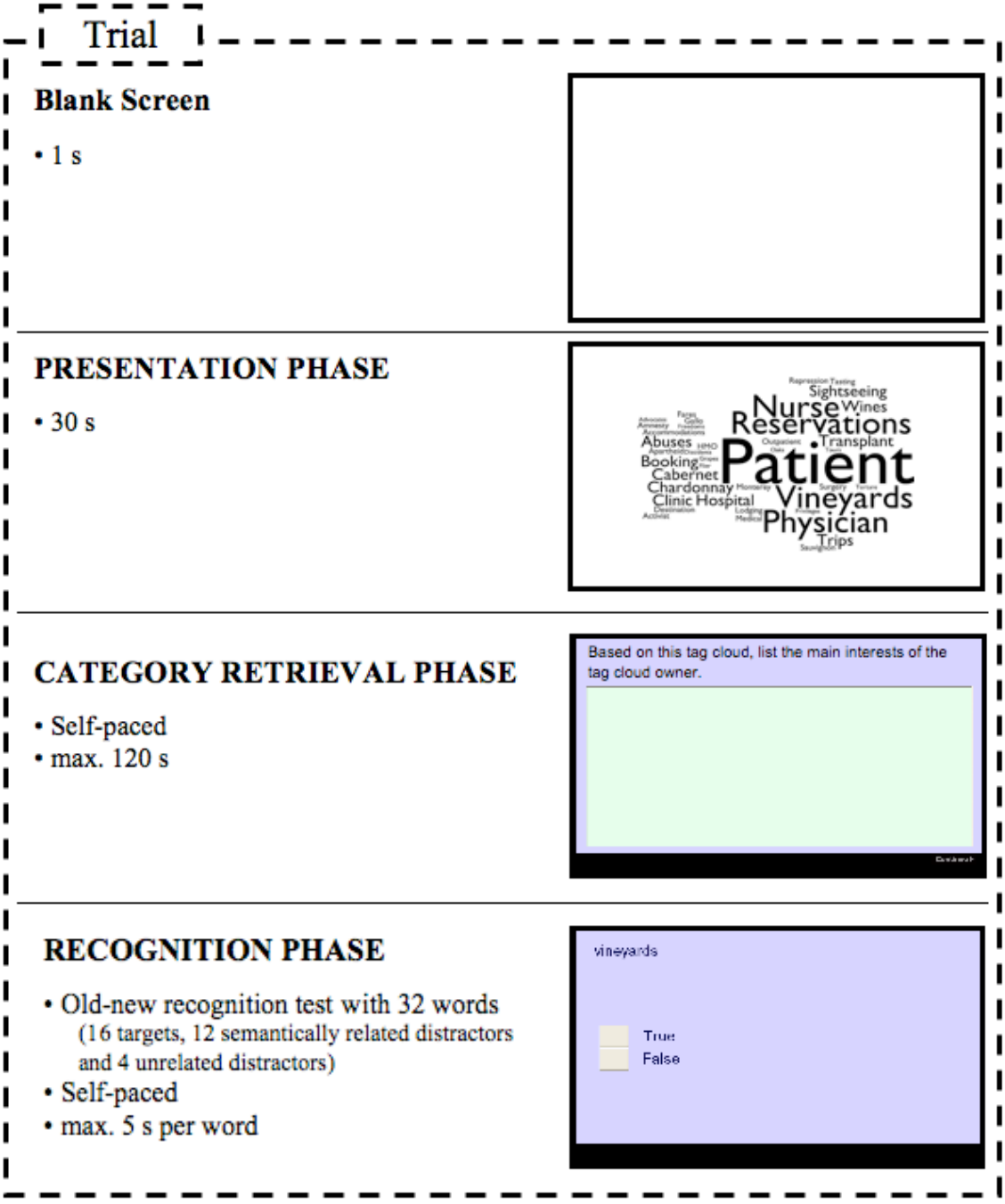

Figure A1. Diagram of an example trial in Experiment 1.

Figure A2

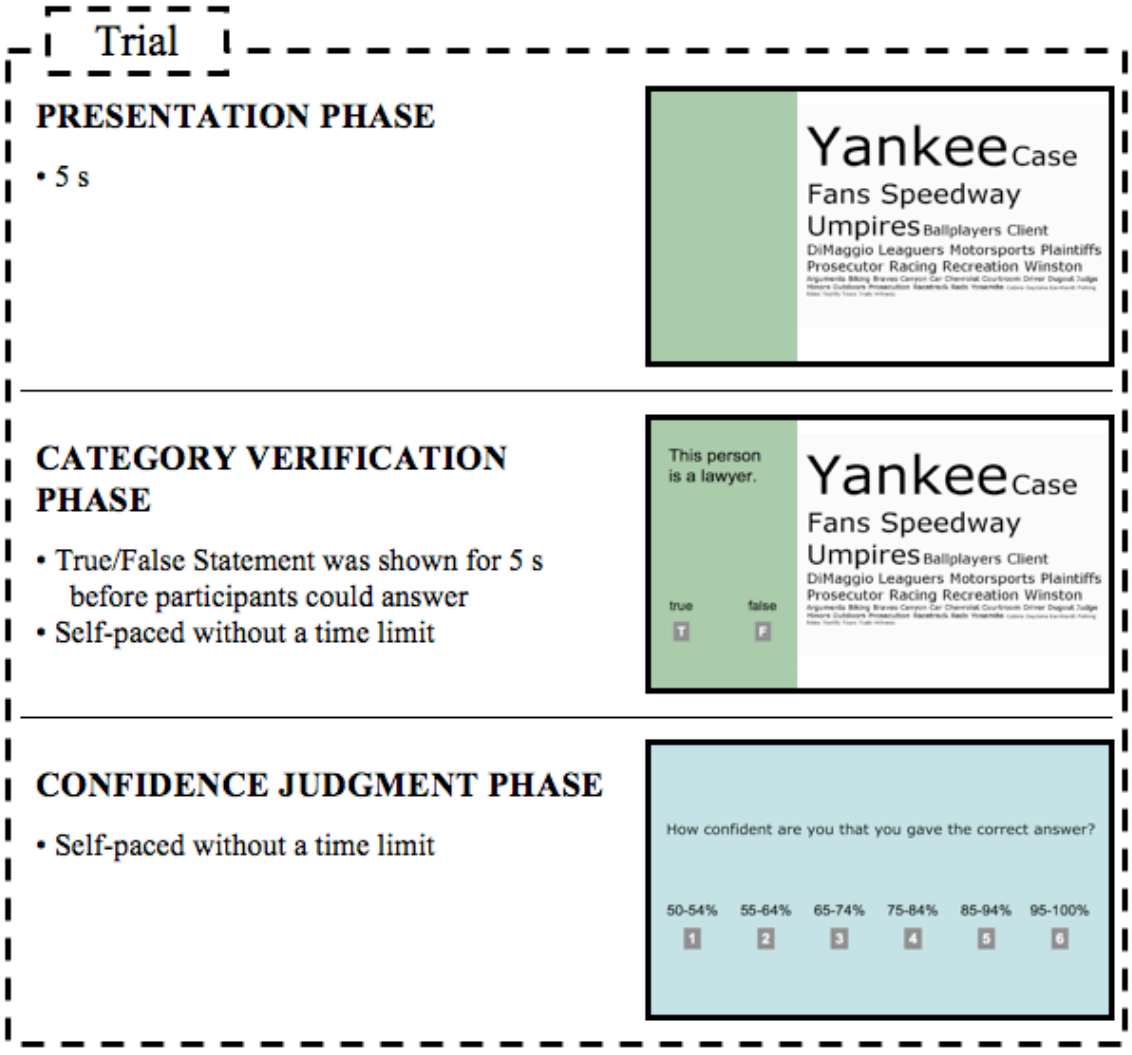

Figure A2. Diagram of an example trial in Experiment 2.

Figure A3

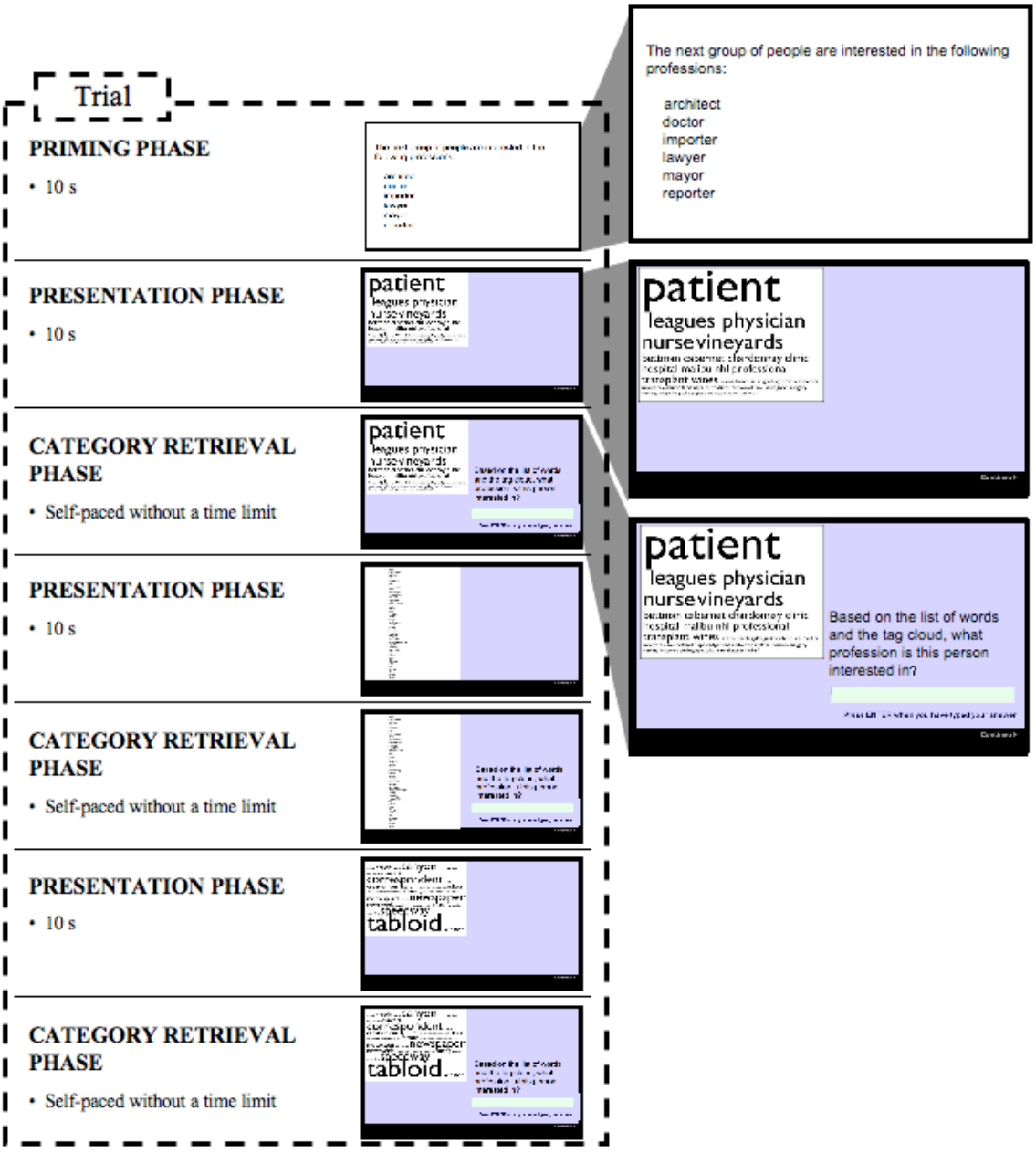

Figure A3. Diagram of an example trial in Experiment 3.

Appendix B

Table B1

		Group 1	Group 2	Group 3	Group 4
Practice Trial	Spatial	Tag Cloud 0	Tag Cloud 0	Tag Cloud 0	Tag Cloud 0
Experimental Trials	Alpha	Tag Cloud 7	Tag Cloud 1	Tag Cloud 4	Tag Cloud 10
		Tag Cloud 8	Tag Cloud 2	Tag Cloud 5	Tag Cloud 11
		Tag Cloud 9	Tag Cloud 3	Tag Cloud 6	Tag Cloud 12
	Freq	Tag Cloud 10	Tag Cloud 4	Tag Cloud 7	Tag Cloud 1
		Tag Cloud 11	Tag Cloud 5	Tag Cloud 8	Tag Cloud 2
		Tag Cloud 12	Tag Cloud 6	Tag Cloud 9	Tag Cloud 3
	Spatial	Tag Cloud 4	Tag Cloud 10	Tag Cloud 1	Tag Cloud 7
		Tag Cloud 5	Tag Cloud 11	Tag Cloud 2	Tag Cloud 8
		Tag Cloud 6	Tag Cloud 12	Tag Cloud 3	Tag Cloud 9
	List by Freq	Tag Cloud 1	Tag Cloud 7	Tag Cloud 10	Tag Cloud 4
		Tag Cloud 2	Tag Cloud 8	Tag Cloud 11	Tag Cloud 5
		Tag Cloud 3	Tag Cloud 9	Tag Cloud 12	Tag Cloud 6

Table B1. Table B1 presents the counterbalancing scheme for Experiment 1. Sequential Layout with Alphabetical Sorting (Alpha), Spatial Layout (Spatial), Sequential Layout with Frequency Sorting (Freq), Single Column List with Frequency Sorting (List by Freq).

Table B2

		Group 1	Group 2	Group 3	Group 4
Practice Trials	Freq	Tag Cloud 0	Tag Cloud 0	Tag Cloud 0	Tag Cloud 0
	Spatial	Tag Cloud 0	Tag Cloud 0	Tag Cloud 0	Tag Cloud 0
	List by Freq	Tag Cloud 0	Tag Cloud 0	Tag Cloud 0	Tag Cloud 0
Experimental Trials	Alpha	Tag Cloud 1	Tag Cloud 4	Tag Cloud 3	Tag Cloud 2
		Tag Cloud 5	Tag Cloud 8	Tag Cloud 7	Tag Cloud 6
		Tag Cloud 9	Tag Cloud 12	Tag Cloud 11	Tag Cloud 10
		Tag Cloud 13	Tag Cloud 16	Tag Cloud 15	Tag Cloud 14
	Freq	Tag Cloud 2	Tag Cloud 1	Tag Cloud 4	Tag Cloud 3
		Tag Cloud 6	Tag Cloud 5	Tag Cloud 8	Tag Cloud 7
		Tag Cloud 10	Tag Cloud 9	Tag Cloud 12	Tag Cloud 11
		Tag Cloud 14	Tag Cloud 13	Tag Cloud 16	Tag Cloud 15
	Spatial	Tag Cloud 3	Tag Cloud 2	Tag Cloud 1	Tag Cloud 4
		Tag Cloud 7	Tag Cloud 6	Tag Cloud 5	Tag Cloud 8
		Tag Cloud 11	Tag Cloud 10	Tag Cloud 9	Tag Cloud 12
		Tag Cloud 15	Tag Cloud 14	Tag Cloud 13	Tag Cloud 16
	List by Freq	Tag Cloud 4	Tag Cloud 3	Tag Cloud 2	Tag Cloud 1
		Tag Cloud 8	Tag Cloud 7	Tag Cloud 6	Tag Cloud 5
		Tag Cloud 12	Tag Cloud 11	Tag Cloud 10	Tag Cloud 9
		Tag Cloud 16	Tag Cloud 15	Tag Cloud 14	Tag Cloud 13

Table B2. Table B2 presents the counterbalancing scheme for Experiment 2. Sequential Layout with Alphabetical Sorting (Alpha), Spatial Layout (Spatial), Sequential Layout with Frequency Sorting (Freq), Single Column List with Frequency Sorting (List by Freq).

Table B3

	Group 1			Group 2			Group 3			Group 4		
	Type of Interest	**Layout**	**Tag Cloud**	**Type of Interest**	**Layout**	**Tag Cloud**	**Type of Interest**	**Layout**	**Tag Cloud**	**Type of Interest**	**Layout**	**Tag Cloud**
Practice Trials	profession	Alpha	0-3	profession	Alpha	0-3	profession	Alpha	0-3	profession	Alpha	0-3
		Freq	0-2		Freq	0-2		Freq	0-2		Freq	0-2
	hobby	List by Freq	0-1	hobby	List by Freq	0-1	hobby	List by Freq	0-1	hobby	List by Freq	0-1
		Spatial	0-2		Spatial	0-2		Spatial	0-2		Spatial	0-2
	hobby	Spatial	0-3	hobby	Spatial	0-3	hobby	Spatial	0-3	hobby	Spatial	0-3
		Alpha	0-1		Alpha	0-1		Alpha	0-1		Alpha	0-1
		List by Freq	0-2		List by Freq	0-2		List by Freq	0-2		List by Freq	0-2
Experimental Trials	profession	Alpha	1	location	Spatial	1	hobby	List by Freq	1	profession	Alpha	2
		List by Freq	2		List by Freq	11		Alpha	2		List by Freq	7
		Freq	3		Freq	12		Freq	12		Freq	8
	sport	Spatial	7	profession	Freq	2	location	Alpha	1	hobby	Spatial	2
		Freq	8		Spatial	3		List by Freq	2		Freq	3
		List by Freq	9		Alpha	4		Spatial	12		List by Freq	8
	location	Spatial	10	sport	Spatial	2	profession	Alpha	3	location	Alpha	4
		Alpha	11		Freq	3		List by Freq	4		Freq	6
		List by Freq	12		List by Freq	4		Freq	5		Spatial	11
	hobby	Alpha	4	location	Alpha	2	sport	List by Freq	3	profession	Spatial	9
		Spatial	5		List by Freq	3		Alpha	4		Alpha	10
		Freq	6		Freq	4		Spatial	5		List by Freq	11
	location	Alpha	7	hobby	List by Freq	2	location	Freq	3	sport	Freq	1
		List by Freq	8		Alpha	3		Spatial	4		List by Freq	6
		Freq	9		Spatial	4		Alpha	5		Alpha	11
	hobby	List by Freq	7	location	Spatial	5	hobby	Spatial	3	location	Freq	2
		Alpha	8		Alpha	6		Freq	4		Spatial	7
		Spatial	9		List by Freq	7		List by Freq	5		List by Freq	9
	profession	Spatial	4	hobby	Freq	5	location	List by Freq	6	hobby	Alpha	1
		Alpha	5		List by Freq	6		Freq	7		Spatial	6
		List by Freq	6		Alpha	7		Spatial	8		Freq	11
	sport	Alpha	10	profession	List by Freq	5	hobby	Alpha	6	location	List by Freq	5
		Spatial	11		Freq	6		Spatial	7		Freq	10
		Freq	12		Spatial	7		Freq	8		Alpha	12
	hobby	Freq	10	sport	Alpha	5	profession	Spatial	6	hobby	List by Freq	4
		List by Freq	11		Spatial	6		Alpha	7		Alpha	9
		Alpha	12		Freq	7		List by Freq	8		Spatial	10
	location	List by Freq	4	hobby	Spatial	8	sport	Freq	6	profession	Spatial	1
		Freq	5		Freq	9		List by Freq	7		Freq	4
		Spatial	6		List by Freq	10		Alpha	8		Alpha	6
	sport	List by Freq	1	location	Freq	8	hobby	List by Freq	9	sport	Spatial	4
		Alpha	2		Spatial	9		Alpha	10		Freq	9
		Spatial	3		Alpha	10		Spatial	11		List by Freq	10
	profession	Freq	7	sport	List by Freq	8	location	Alpha	9	hobby	Alpha	5
		Spatial	8		Alpha	9		List by Freq	10		Freq	7
		Alpha	9		Spatial	10		Freq	11		List by Freq	12
	sport	Freq	4	profession	Alpha	8	sport	Spatial	9	location	List by Freq	1
		List by Freq	5		List by Freq	9		Freq	10		Spatial	3
		Alpha	6		Freq	10		List by Freq	11		Alpha	8
	profession	List by Freq	10	sport	Alpha	1	profession	Freq	9	sport	List by Freq	2
		Freq	11		Freq	11		Spatial	10		Alpha	3
		Spatial	12		List by Freq	12		Alpha	11		Spatial	12
	hobby	Spatial	1	profession	List by Freq	1	sport	Spatial	1	profession	List by Freq	3
		Freq	2		Spatial	11		Freq	2		Spatial	5
		List by Freq	3		Alpha	12		Alpha	12		Freq	12
	location	Freq	1	hobby	Freq	1	profession	Freq	1	sport	Freq	5
		Spatial	2		Alpha	11		Spatial	2		Alpha	7
		Alpha	3		Spatial	12		List by Freq	12		Spatial	8

Table B3. Table B3 presents the counterbalancing scheme for Experiment 3. Sequential Layout with Alphabetical Sorting (Alpha), Spatial Layout (Spatial), Sequential Layout with Frequency Sorting (Freq), Single Column List with Frequency Sorting (List by Freq).

References

Adams, A.S., & Edworthy, J. (1995). Quantifying and predicting the effects of basic text display variables on the perceived urgency of warning labels: tradeoffs involving font size, border weight and colour. *Ergonomics, 38*, 11, 2221-2237.

Alba, J.W., & Hasher, L. (1983). Is memory schematic? *Psychological Bulletin, 93,* 2, 203-231.

Ashby, F. G., & Maddox, W. T. (1993). Relations between prototype, exemplar, and decision bound models of categorization. *Journal of Mathematical Psychology, 37*(3), 372-400.

Battista, J., & Kalloniatis, M. (2002). Left-right word recognition asymmetries in central and peripheral vision. *Vision Research, 42*, 1583-1592.

Berlin, B., & Kay, P. (1969). *Basic color terms: their universality and evolution*. Berkeley: University of California Press.

Brier, G.W. (1950). Verification of forecasts expressed in terms of probability. *Monthly Weather Review, 78,* 1-3.

Cohen, J. (1998). *Statistical power analysis for the behavioral sciences* (Second Edition). Hillsdale, NJ: Erlbaum.

Cohen, J., Cohen, P., West, S. G., & Aiken, L. S. (2003). *Applied multiple regression/correlation analysis for the behavioral sciences* (Third Edition). Mahwah, NJ: Erlbaum.

Computational Semantics Lab from Stanford University. (n.d. a). Information mapping project. Retrieved February 4, 2008, from http://infomap.stanford.edu

Computational Semantics Lab from Stanford University. (n.d. b). Infomap algorithm description. Retrieved February 5, 2008, from http://infomap-nlp.sourceforge.net/doc/algorithm.html

Daehler, M., Lonardo, R., & Bukatko, D. (1979). Matching and equivalence judgments in very young children. [Abstract] *Child Development, 50*(1), 170-179.

Deese, J. (1959). On the prediction of occurrence of particular verbal intrusions in immediate recall. *Journal of Experimental Psychology, 58,* 17-22.

Dorow, B., & Widdows, D. (2003). Discovering corpus-specific word senses. *Proceedings of the 10th. Conference of the European Chapter of the Association for Computational Linguistics, Conference Companion (research notes and demos), Hungary*, 79-82.

Estes, W.K. (1986). Memory storage and retrieval processes in category learning. *Journal of Experimental Psychology: General, 115*, 155-175.

Fiske, S.T., Neuberg, S.L., Beattle, A.E., & Milberg, S.J. (1987). Category-based and attribute-based reactions to others: Some informational conditions of stereotyping and individuating processes. *Journal of Experimental Social Psychology, 23,* 399-427.

Gati, L., & Tversky, A. (1984). Weighting common and distinctive features in perceptual and conceptual judgments. *Cognitive Psychology, 16,* 341-370.

Garner, W.R. (1978). Selective attention to attributes and to stimuli. *Journal of Experimental Psychology: General, 107,* 287-308.

Golder, S., & Huberman, B.A. (2006). Usage patterns of collaborative tagging systems. *Journal of Information Science, 32*(2), 198-208.

Goldstone, R.L. (1994a). Similarity, Interactive Activation, and Mapping. *Journal of Experimental Psychology, Learning, Memory and Cognition, 20*(1), 2-28.

Goldstone, R.L. (1994b). The role of similarity in categorization: providing a groundwork. *Cognition, 52*, 125-157.

Gonzalez, R., & Nelson, T.O. (1995). Measuring ordinal association in situations that contain tied scores. *Psychological Bulletin, 119*(1), 159-165.

Goodman, N. (1972). Seven strictures on similarity. In N. Goodman (Ed.), *Problems and projects* (pp. 437-447). New York: Bobbs-Merrill.

Halvey, M., & Keane, M.T. (2007). An assessment of tag presentation techniques. *Proceedings of the 16th International World Wide Web Conference (WWW, 2007),* Canada, 1313-1314.

Hearst, M.A., & Rosner, D. (2008). Tag clouds: Data analysis tool or social signaler? *Proceedings of the 41st Hawaii International Conference on System Sciences, (HICSS, 2008),* IEEE Computer Society, Washington, DC, 160-169.

Hedges, L.V., & Olkin, I. (1985). *Statistical methods for meta-analysis.* New York: Academic Press.

Hersh, W., Buckley, C., Leone, T. J., & Hickam, D. (1994). OHSUMED: an interactive retrieval evaluation and new large test collection for research. *Proceedings of the 17th Annual international ACM SIGIR Conference on Research and Development in information Retrieval (Ireland).* W. B. Croft and C. J. van Rijsbergen, Eds. Annual ACM Conference on Research and Development in Information Retrieval. Springer-Verlag New York, New York, NY, 192-201.

Hintzman, D.L. (1988). Judgments of frequency and recognition memory in a multiple-trace memory model. *Psychology Review, 95,* 528-551.

Homa, D., Cross, J., Cornell, D., Goldman, D., & Shwartz, S. (1973). Prototype abstraction and classification of new instances as a function of number of instances defining the prototype. *Journal of Experimental Psychology, 101*(1), 116-122.

Homa, D., & Cultice, J. (1984). Role of feedback, category size, and stimulus distortion on the acquisition and utilization of ill-defined categories. *Journal of Experimental Psychology, Learning, Memory and Cognition, 10*(1), 83-94.

Homa, D. & Vosburgh, R. (1976) Category breadth and the abstraction of prototypical information. *Journal of Experimental Psychology: Human Learning and Memory, 2*(3), 322-330.

Jarvis, B. G. (2006). DirectRT (Version 2006.2) [Computer Software]. New York, NY: Empirisoft Corporation.

Jarvis, B. G. (2002). MediaLab (Version 2006.2) [Computer Software]. New York, NY: Empirisoft Corporation.

Joelson, J.M., & Herrmann, D.J. (1978). Properties of categories in semantic memory. *American Journal of Psychology, 91,* 101-114.

Kahneman, D. & Tversky, A. (1972). Subjective probability: A judgment of representativeness. *Cognitive Psychology, 3*, 430-451.

Kahneman, D., & Tversky, A. (1973). On the Psychology of Prediction. *Psychological Review, 80*, 237-251.

Kahneman, D. & Tversky, A. (1979). Prospect theory: An analysis of decision under risk. *Econometrica*, *47,* 263-292.

Kahneman, D. & Tversky, A. (1996). On the reality of cognitive illusions. *Psychological Review, 103*, 582-591.

Kaser, O. & Lemire, D. (2007). Tag-Cloud drawing: Algorithms for cloud visualization. *Proceedings of Tagging and Metadata for Social Information Organization (WWW 2007),* Canada. Retrieved February 16, 2008, from http://www2007.org/workshops/paper_12.pdf

Kirk, R. E. (1995). *Experimental design* (Third Edition). Belmont, CA: Thomson/Wadsworth.

Komatsu, L.K. (1992). Recent views of conceptual structure. *Psychological Bulletin, 112*(3), 500-526.

Krantz, D.H., & Tversky, A. (1975). Similarity of rectangles: An analysis of subjective dimensions. *Journal of Mathematical Psychology, 12,* 4-34.

Legge, G.E., Pelli, D.G, Rubin, G.S., & Schleske, M.M. (1985). Psychophysics of reading: I. Normal vision. *Vision Research, 2*, 239-252.

Luce, R., & Raiffa, H. (1957). Individual decision making under uncertainty. In R. Luce and H. Raiffa. *Games and Decisions: Introduction and Critical Survey*. New York, John Wiley & Sons, Inc., 275-306.

Lund, K., & Burgess, C. (1996). Producing high-dimensional semantic spaces from lexical co-occurrence. *Behavior Research Methods, Instruments & Computers, 28,* 203-208.

Mansfield, J.S., Legge, G.E., & Bane, M.C. (1996). Psychophysics of reading: XV. Font effects in normal and low vision. *Investigative Ophthalmology and Visual Science, 37*(8), 1492-1501.

Marlow, C., Naaman, M., Boyd, D.M., & Davis, M. (2006). HT06, Tagging Paper, Taxonomy, Flickr, Academic Article, To Read. In U.K. Wiil, P.J. Nürnberg, & J.Rubart (Eds.), *Proceedings of the Seventeenth ACM Conference on Hypertext and Hypermedia* (pp. 31-40). Odense, Denmark: ACM Press.

McCloskey, M.E., & Glucksberg, S. (1978). Natural categories: Well defined or fuzzy sets? [Abstract] *Memory & Cognition, 6*(5), 462-472.

Medin, D.L. (1989). Concepts and Conceptual Structure. *American Psychologist, 44,* 12, 1469-1481.

Medin, D.L., Goldstone, R.L., & Gentner, D. (1991). Respects for Similarity. *Psychological Review, 100*, 254-278.

Medin, D.L., Goldstone, R.L., & Markman, A.B. (1995). Comparison and choice: Relations between similarity processes and decision processes. *Psychonomic Bulletin & Review, 2*(1), 1-19.

Medin, D.L., & Schaffer, M. (1978). Context theory of classification learning. *Psychological Review, 85*, 207-238.

Medin, D.L., Wattenmaker, W.D., & Hampson, S.E. (1987). Family resemblance, conceptual cohesiveness, and category construction. *Cognitive Psychology, 19,* 242-279.

Mervis, C.B., & Crisafi, M.A. (1982). Order of acquisition of subordinate-, basic- and superordinate-level categories. *Child Development, 53*(1), 258-266.

Mervis, C.B., & Rosch, E. (1981). Categorization of Natural Objects. *Annual Review of Psychology, 32,* 89-115.

Millen, D.R., Feinberg, J., & Kerr, B. (2005). Social bookmarking in the enterprise. *ACM Queue, 3*(9), 29-35.

Miller, G.A. (1995). WordNet: A lexical database for English. *Communications of the ACM, 38*(11), 39-41.

Morikawa, K., & McBeath, M.K. (1992). Lateral motion bias associated with reading direction. *Vision Research, 32*(6), 1137-1141.

Murphy, A.H. (1973). A new vector partition of the probability score. *Journal of Applied Meteorology, 12,* 595-600.

Murphy, G.L., & Wright, J.C. (1984). Changes in conceptual structure with expertise: Differences between real-world experts and novices. *Journal of Experimental Psychology: Learning, Memory and Cognition, 10*(1), 144-155.

Nosofsky, R. (1986). Attention, similarity, and the identification of separable-dimension stimuli: A choice model analysis. *Perception & Psychophysics, 38*, 415-432.

Osherson, D.N., & Smith, E.E. (1981). On the adequacy of prototype theory as a theory of concepts. *Cognition, 9*, 35-58.

Posner, M.L., & Keele, S.W. (1968). On the genesis of abstract ideas. *Journal of Experimental Psychology, 77*, 353-363.

Posner, M.L., & Keele, S.W. (1970). Retention of abstract ideas. *Journal of Experimental Psychology, 83*, 304-308.

Pirolli, P., Schank, P., Hearst, M., & Diehl, C. (1996). Scatter/Gather browsing communicates the topic structure of a very large text collection. *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. CHI '96. ACM, Vancouver, BC, 213-220.

Reed, S.K. (1972). Pattern recognition and categorization. *Cognitive Psychology, 3,* 382-407.

Rips, L.J. (1989). Similarity, typicality, and categorization. In S. Vosniadu & A. Ortony (Eds.), *Similarity, analogy, and thought* (pp. 21-59). Cambridge: Cambridge University Press.

Rips, L.J., Shoben, E.J., & Smith, E.E. (1975). Semantic distance and the verification of semantic relations. *Journal of Verbal Learning & Verbal Behavior, 12*(1), 1-20.

Rivadeneira, A.W., Gruen, D.M., Muller, M.J. & Miller, D.R. (2007). Getting our head in the clouds: toward evaluation studies of tagclouds. *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. CHI '07. ACM, New York, NY, 995-998.

Roediger, H.L., & McDermott, K.B. (1995). Creating false memories: Remembering words not presented in lists. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 21*(4), 803-814.

Rosch, E. (1975). Cognitive representations of semantic categories. *Journal of Experimental Psychology: General, 104,* 192-133.

Rosch, E., & Mervis, C.B. (1975). Family resemblances: Studies in the internal structure of categories. *Cognitive Psychology, 7*, 573-605.

Rosch, E., Mervis, C. B., Gray, W. D., Johnson, D. M., & Boyes-Braem, P. (1976). Basic objects in natural categorization. *Cognitive Psychology, 8,* 382-439.

Rosch, E., Simpson, C., & Miller, S. (1976). Structural bases of typicality effects. *Journal of Experimental Psychology: Human Perception and Performance, 2*(4), 491-502.

Shepard, R.N. (1987). Toward a universal law of generalization for psychological science. *Science, 237*, 1317-23.

Sloutsky, V.M. (2003).The role of similarity in the development of categorization. *Trends in Cognitive Science, 7*(6), 246-251.

Smith, E.E. & Medin, D.L. (1981). *Categories and Concepts*. Harvard University Press.

Smith, E.E. & Sloman, S. (1994). Similarity- versus rule-based categorization. *Memory & Cognition, 22,* 377-386.

Smith, L.B. (1981). Importance of the overall similarity of objects for adults' and children's classifications. *Journal of Experimental Psychology: Human Perception and Performance, 7*(4), 811-824.

Spalek, T.M., & Hammad, S. (2005). The left-to-right bias in inhibition of return is due to the direction of reading. *Psychological Science, 16*(1), 15-18.

Steyvers, M., Griffiths, T.L., & Dennis, Simon. (2006). Probabilistic inference in human semantic memory. *Trends in Cognitive Sciences, 10,* 327-334.

Tanaka, J. W., & Taylor, M. (1991). Object categories and expertise: Is the basic level in the eye of the beholder? *Cognitive Psychology, 23*, 457-482.

Taylor, S.E., Fiske, S.T., Etcoff, N.L., & Ruderman, A.J. (1978). Categorical and contextual bases of person memory and stereotyping. *Journal of Personality and Social Psychology, 36,* 778-793.

Torgerson, W.S. (1952). Multidimensional scaling: I. Theory and method. *Psychometrika, 17*, 401.

Tversky, A. (1969). Intransitivity of preferences. *Psychological Review, 76*, 31-48.

Tversky, A. (1977). Features of similarity. *Psychological Review, 84*, 327-351.

Tversky, A. & Kahneman, D. (1974). Judgment under uncertainty: heuristics and biases. *Science, 185*, 1124-1131.

Tversky, A., & Kahneman, D. (1983). Extensional vs. intuitive reasoning: The conjunction fallacy in probability judgment. *Psychological Review, 91*, 293-315.

Vitu, F., Kapolua, Z., Lancelin, D., & Lavigne, F. (2004). Eye movements in reading isolated words: evidence for strong biases towards the center of the screen. *Vision Research, 44*, 321-338.

Widdows, D. (2003). Orthogonal negation in vector spaces for modeling word-meanings and document retrieval. *Proceedings of the 41st Annual Meeting of the Association for Computational Linguistics,* Japan, 136-143.

Widdows, D., & Dorow, B. (2002). A graph model for unsupervised lexical acquisition. *Proceedings of the 19th International Conference on Computational Linguistics (COLING 19),* Taipei, 1093-1099.

Whitten, W.B., Suter, W.N., & Frank, M.L. (1979). Bidirectional synonym ratings of 464 noun pairs. *Journal of Verbal Learning & Verbal Behavior, 18*(1), 109-127.

Zadeh, L. (1965). Fuzzy sets. *Information and control, 8*(3), 338-353.