Journal of Rhetoric, Professional Communication, and Globalization

Volume 1 | Number 1 Article 3

2010

Rethinking the Problem of Linguistic Categorization for Global Search Engines

Matthew McCool
Southern Polytechnic State University

Rhetoric, Professional Communication and Globalization

ISSN: 2153-9480. Volume 1, Number 1. December - 2010

Rethinking the Problem of Linguistic Categorization for Global Search Engines[1]

Matthew McCool
Southern Polytechnic State University, USA

Abstract

The fields of social psychology and neuroscience have known for several decades that culture affects the way people carve up the world. This perceptual difference is often, but not always, aligned with similar differences in linguistics categories. If correct, this problem of linguistic categorization may have considerable impact on search algorithms. This paper examines the relationship between culture and linguistic categorization for global search engines. A total of 43 American and Chinese participants completed two classification tests, one derived from social psychology and neuroscience and the other based on a common classification problem for full-text searching. These data suggest that Chinese participants are more field dependent, American participants are less field dependent, and that these results may offer important clues about adapting search algorithms for global computing systems.

Introduction

Conventional design protocols for adapting or "porting" computing systems to a global audience relies on little more than translation. This makes sense because language is an obvious barrier to communicating with people from other countries. A computer application developed in French must be translated into Japanese for users in Tokyo. This requirement of translating a word from one language into another is so obvious that it requires no discussion. For many computing systems, this essential change is all that is required to meet the needs of global users. Unfortunately, the same does not apply to search engines, a domain where natural languages and programming languages collide. And the best way to see this collision is through the lens of language and perception.

[1] A subset of these data was published with Kirk St. Amant in 2009 in the *Journal of the American Society for Information Science and Technology*, *60*(6), 1258-1266.

One function of perception is to pick things out in the environment. Psychologists refer to this perceptual process as the difference between *figure and ground*, and it is one of the most basic cognitive functions performed by the human mind (Masuda & Nisbett 2001; Nisbett 2003; Nisbett & Masuda 2003; Nisbett, Choi, Peng, & Norenzayan 2001; Paulesu et al 2000). A frog nestled among plants is an example of figure and ground in which the frog is the "figure" and the plants are the "ground." The degree to which a frog is embedded among some plants is known as *field dependence*. Remarkably, environmental factors can have a tremendous effect on whether people see or focus on frogs or plants, and it can also affect recall. But we do not just perceive things in the world, we also talk about them.

For instance, we have the ability to implant ideas into other minds through language. It is in this way that the grammar of language must interface with perception (Pinker 2003). The word "frog" represents the abstract category of cute amphibians and the word "plant" represents the abstract category of things that make a green pigment called chlorophyll. I can make you think about a frog when I use the word, and you can do so without having to think of plants. It is this link between perceiving things in the world and our ability to connect that perception with words and rules that makes this kind of communication distinctly human (Pinker 2003). It may also have a profound effect on search engines. Consider an example from astronomy. Suppose a user queries the term *galaxy* based on a general interest in galactic bodies. The term *galaxy* may return a wide range of results that seem non-specific and range from any of the 110 Messier objects, the general catalog of Deep Sky Objects, objects from the Herschel 400 catalog, the New General Catalogue, any of the 109 items of the Caldwell catalog, or any of the millions of other celestial objects accessible through modern optics. Like people, galaxies come in the form of all kinds of strange morphologies. There are barred, barred elliptical, dwarf, dwarf spheroidal, irregular (or peculiar), lenticular, ring, spiral, starburst, and unbarred spiral galaxies, to name a few. The point here is that a generic search term such as *galaxy* has limited power to return a specific result. The problem is that users are normally looking for something more specific. The dilemma is a logical problem of categorization. And the difficulty of figuring out which words or queries are ideal for a specific user population is the subject of this paper.

Background

Rethinking the problem of linguistic categorization for search engines must account for two issues. The first problem is with linguistic categorization, which is a branch of linguistics concerned with the way language is used to create (or reflect) categories. The second problem is with search engines. Although the engines that drive a search query are written in computer programming languages, they work by processing natural languages. This is why search algorithms, a deep problem in computer science, are further complicated by crossing countries, languages, and cultures. It is for these reasons that linguistic categorization and search engines must be addressed.

Linguistic Categorization

There is a clear link between perceiving things in the world and naming them. The ability to pick out things in the environment, attach them to concepts, and then couple them with words is a special trick of language. The mechanism that binds a thing in the world with a word is a concept, which has given rise to a special area of study known as conceptual semantics. This ability to assign a conceptual meaning to a word and then use it to implant an idea in another mind is so commonplace that it goes unnoticed. Yet, the ability to communicate ideas among other minds is, in every real sense, a remarkable ability. At the same time, there are always problems when trying to communicate with others. No matter how hard one tries, the veil of ambiguity is always present. Such is the case with categories, a fundamental problem in linguistics and, more recently, evolutionary psychology (Pinker 2003; Taylor 2003).

The possibility that language may play some role in mental processing is an idea that has been around since at least the early twentieth century. Commonly known as the Sapir-Whorf or Whorfian hypothesis, this theory states that language *may* affect perception and thinking (Connor 1996). Though controversial, it has been proposed that some social groups may perceive things in the world differently based on their native language (Masuda & Nisbett 2001; Nisbett 2003; Nisbett & Masuda 2003; Nisbett, Choi, Peng, & Norenzayan 2001; Paulesu et al 2000). This means that a native speaker of French has a fundamentally different way of carving up the world than someone whose native language is English. Although recent research in psychology and neuroscience is peeling back the layers of this complex theory, some of which is supportive, the jury is still out on whether language affects cognition (Masuda & Nisbett 2001; Nisbett 2003; Nisbett & Masuda 2003; Nisbett, Choi, Peng, & Norenzayan 2001; Paulesu et al 2000).

Regardless of how the Whorfian theory pans out, it is clear that language is related to the way broad groups of people organize things in the world. The difference between novices and experts provides a useful template. An expert potter will find greater variation between cups and bowls than a novice (Taylor 2003). The same argument seems to hold true for any kind of expertise. The belief is that years of experience lead to increased ability. Such experts are further defined by a specialized language and vocabulary that also contribute to more differences. Most people would have trouble naming four or five major lobes of the brain. A neurosurgeon, on the other hand, regularly works with the concept that the brain has over fifty distinct areas. The difference is based on variations in expertise.

Aristotle more or less believed that things in the world held a set of discrete traits (Aristotle 2001). Either a thing fit in a category or it did not. Kant believed in a kind of metaphysical essence (Kant 2008). Wittgenstein felt that language, or words, held various meanings that interfered with understanding (Wittgenstein 1965). And Whorfian scholars have advanced the theory that language is not only a window into the mind but also culture (Connor 1996). There is probably some truth to all of these claims, but they do little to help us understand the way that language shapes (or is shaped by) categories and their things. And for that we should address two basic ways that categories are used in the mind.

Research in psychology has revealed that the mind processes information through one of two categories (Pinker 2003). One category is discrete and rigid, as with Aristotle's classical definition. The second category is fuzzy and fluid, which is similar to Wittgenstein's work. Some things occupy clear and crisp categories and other things seem to cross boundaries. The difference between these two perceptual distinctions is partly based on degrees of expertise and partly based on real and imaginary differences. Consider the prototype for the category *bird*.

Birds are things that grow feathers, use beaks for food, live in nests, and lay eggs for offspring. There are nearly 10,000 different types of birds. There are penguins, falcons, owls, hummingbirds, and ostriches, to name a few. By any measure, birds are a diverse lot that make them difficult to categorize. Penguins are fat and swim in water and often live in cold climates. Hummingbirds are small and frenetic creatures with hearts the size of the tip of a ballpoint pen. There seems to be little in common between penguins and hummingbirds, yet any child can tell you that they belong to the same group. This seems like a remarkable feat of the mind when you consider that upon looking up the word *bird* in a dictionary, chances are that you will find a warbler. How can this be?
It turns out that things occupy degrees of centrality in a category. Penguins and hummingbirds may be birds, but they are not typical. Some birds have more "bird-like" qualities, and they occupy a central location in a category. The average or prototypical bird is small, round, small-beaked, flies, and lives in trees. This does not mean that ostriches are not birds but it does mean that they occupy the edge of the category *bird*.

All of this is important because the placement of things in categories has not only cognitive implications but also affects search engines. Querying a site on hummingbirds that provides information about all species of birds will have to be matched against more than 9,000 species. The word *bird* is too broad, *Mellisuga helenae* too technical, and *flying jewel* too metaphorical. And that is precisely the problem. Search queries are imperfect because people do not always know what it is they want, and when they do know they are not always certain how best to find it. Clearly, the names and categories of things are of the utmost importance for global search engines, and it all starts with a cognitive problem known as field dependence.

Field Dependence

Field dependence is conceptually wrapped around what is known as figure and ground. The concept of figure and ground is based on the notion that a focal object is visually distinct from its background. Examples of figure and ground include the difference between a tree and the horizon or a frog sitting among plants. The ability to pick out things like trees and frogs is an example of visual processing that reveals something important about how the mind works (Berry, Poortinga, Segall, & Dasen 2002; Chua, Boland, & Nisbett 2005; Hannah, Boland, & Nisbett 2005; Nisbett, Choi, Peng, & Norenzayan 2001; Norenzayan, Choi, & Nisbett 2002; Peng & Nisbett 1999). A capacity to make distinctions between frogs and plants requires one to have different mental concepts for the abstract categories *frog* and *plant*.

This perceptual difference is deeply rooted in the way people categorize. Social factors such as education, geographic region, and urban density have all been implicated in affecting figure and ground (Masuda & Nisbett 2001; Nisbett & Masuda 2003; Norenzayan, Choi, & Nisbett 2002). More recently, these same perceptual differences have been linked with broad social groups. Based on a variety of experimental tests, some eastern cultures appear to pay more attention to an object's background while some western cultures appear to focus on the object (Masuda & Nisbett 2001; Nisbett 2003; Nisbett & Masuda 2003; Nisbett, Choi, Peng, & Norenzayan 2001; Paulesu et al 2000). The degree to which an object is perceived to be embedded within its background is called field dependence.

Field dependence is directly related to linguistic categorization, and there are many examples from which to draw that help explain this remarkable connection. A poor example may be found in the myth that the Eskimo have over 200 words for snow (Pullum 1991). This unfortunate myth, which is now disappearing, is based on the faulty assumption and erroneous belief that there is something special about the Eskimo in respect to snow. Perhaps it is a heightened visual awareness, a keener aptitude for arctic topography, or some feature innate to arctic-dwelling peoples. The myth, as it has been retold by scholars and journalists alike, is typically rooted in a deep cognitive difference among the Eskimo. If the Eskimo have more words for snow than other people, so the argument goes, then they must be *seeing* something in their topography that others cannot. Their level of field dependence must be fundamentally different from other people, which explains their robust snow vocabulary.

A better example is the theory that linguistic categorization and field dependence are related, but not necessarily because of some causal function. One such account in linguistics is the lack of a word for aquamarine in Russian. This seemingly banal observation might seem odd from an English speaker's perspective, but not necessarily because it is based on some deep cognitive difference. If the Eskimo theory for snow vocabulary were true, then Russians do not have a word for *aquamarine* because they simply cannot see it. If this theory were correct, the Cold War could have been avoided with a skilled artist. In fact, there is no compelling theory that explains this peculiar instance in the Russian language, but it is clearly not from some cognitive deficit among Russians.

The value of connecting field dependence with linguistic categorization is important not simply on theoretical grounds. Perception and language are interconnected because we use language to describe and understand the world, and to share that world with other minds. The problem is that categories can sometimes be messy, languages seem to have built-in ambiguities that seem more like a feature than a bug, and people bring different assumptions about the world. All of these things help contribute to the increasingly important (and researched) areas of social psychology, neuroscience, and linguistic categorization. The real question is whether these aspects of natural language are affected by the programming languages of search engines.

Search Engines

One of the more interesting and important language problems for computing systems has to do with search engines. Finding a specific piece of information in an increasing sea of density is one of the most important tasks of today's users. The same problems that one finds within one's own culture are magnified on a global scale. It is for this reason that Google is the primary search engine for native speakers of English, Baidu is the primary engine for Chinese speakers, and Yandex is the primary search engine for Russian speakers.

The notion that Google knows how to deliver appropriate information to Chinese users is not only misguided but wrong. Despite numerous confounding variables, such as culture and online access, the main reason for the variety of search engines around the world is a practical matter. Google has never figured out how, according to their mission statement, "to organize the world's information and make it universally accessible and useful." According to Baidu's mission statement, they are not even interested in the world's information. Baidu is applying "avante garde technology to the world's most ancient and complex language," claims that are flat out wrong (Connor 1996). And Russia's premiere search engine Yandex wants to provide "homegrown world-class technology" for the "Russian internet." Of these three companies, only one is interested in the world's information. The problem is that Google has had little success on a global scale.

One of the reasons for this has to be about a basic misunderstanding in the way language is used. Any translator will tell you that converting a word from one language to the next is not always easy. Sometimes there are no direct equivalents for a word. Sometimes the word is highly contextualized within a complex phrasal structure. And sometimes the concept is foreign or avoided. The problem can be grasped through a brief explanation. Searching is based on the user's query matching keywords based on relevance. Many full-text search engines allow for either natural language or Boolean operators, or both. This means a user may type either "Orion nebula" or "Orion AND nebula." While these two search queries may retrieve different pages, they perform the same function. Many full-text search engines comb entire pages or databases to match keywords based on relevance. People typically assume that word frequency in a database yields a higher ranking. This is wrong. In fact, words that are used less frequently receive higher rankings. The exception to this rule is for disposable words such as definite articles, which are not factored into a search ranking.

The natural language full-text search engine is a popular strategy for delivering specific information to user queries. The problem is when wrong or inadequate keywords and description words are used. A user that searches for "nebulas" may find many web pages on an astronomy site. But if the same user types in "Orion nebula," then she is likely to retrieve pages that are more specific to her needs. Again, full-text search engines rank less common words higher and more common words lower. Typing in Messier 41 (also known as the Orion nebula) will return more accurate results than typing in "Orion" or "nebula" because Orion is also a constellation and nebulas are everywhere. All of this works fine until a user starts looking for information in places the developer never considered. In other words, the user is categorizing information in a way that is different

from the developer. Thus, the user faces the problem of linguistic categorization for global search engines.

Methods

The methods of this study were designed for the purpose of assessing the relationship between linguistic categorization and full-text searching. A total of 43 participants responded to two categorization tasks. The first task was adapted from existing research in social psychology and neuroscience (Masuda & Nisbett 2001; Nisbett 2003; Nisbett & Masuda 2003; Nisbett, Choi, Peng, & Norenzayan 2001; Paulesu et al 2000). This task provided a benchmark for assessing the degree of field dependence among participants. The second task was adapted from a common problem in full-text search engine algorithms for computing systems. The word catalog used for the second categorization task was derived from a set of cancer and cancer-related terms provided by the National Cancer Institute. The classification task was selected for its presence among general purpose users around the world. Specifically, the National Cancer Institute provides information on over 200 different types of cancer for the general user. A broad general audience and its global vision were the two primary reasons for selecting the site. The selection of these two classification tasks were based on the working assumption that culture may affect classification, as recent research in social psychology and neuroscience have unveiled (Feldman & Turvey 1980; Hannah, Boland, & Nisbett 2005; Heden et al 2008; Masuda & Nisbett 2001; Nisbett 2003; Nisbett & Masuda 2003; Nisbett, Choi, Peng, & Norenzayan 2001; Paulesu et al 2000).

Participants

A total of 43 subjects participated in the study. Twenty-six Chinese subjects (5 male, 17 female) and 12 American subjects (5 male, 7 female) participated in the study. The mean age of the Chinese cohort was 21.8 years of age, while the mean age of the American cohort was 22.3 years of age. All of the Chinese subjects were born in China, earned degrees in their native country, and moved to the U.S. to pursue additional undergraduate studies. Further, all of the Chinese students were bilingual, speaking not only Chinese but also a sufficient level of English to gain entry into a U.S. university. All of the U.S. subjects spoke English as their native language, and none had sufficient knowledge of a second language to be considered bilingual.

Materials

The first classification test was derived from social psychology and neuroscience, and is commonly known as the cow, chicken, and grass test (Masuda & Nisbett 2001; Nisbett 2003; Nisbett, Choi, Peng, & Norenzayan 2001). Rather simple, the test consists of three caricatures—one cow, one chicken, and a small tuft of grass. The images were taken from a now-standard social psychology test and presented in a linear fashion. Each image was set to monochrome (black and white) and presented in a liner fashion (chicken, grass, cow). Each caricature shared a similar set of traits. In particular, each image was drawn from hand and subject to the same classification studies in social psychology (Masuda & Nisbett 2001; Nisbett 2003; Nisbett & Masuda 2003; Nisbett, Choi, Peng, & Norenzayan 2001; Paulesu et al 2000).

The second classification test was derived from a common problem in full-text search algorithms. Derived from the National Cancer Institute's site, the second classification test consisted of organizing a small corpus of over 200 different types of cancers. Each cancer was highly specific, such as "Brain Tumor, Pineal Parenchymal Tumors of Intermediate Differentiation, Childhood" and "Squamous Neck Cancer with Occult Primary, Metastatic." The goal of the second classification task was to assess the degree of linguistic categorization for different types of cancer. The importance of this task is based on the assumption that the National Cancer Institute is a global entity with users from around the world who disseminate critical health information to the general population. It is their desire to reach people from around the world that makes this task not only global but also salient and critical.

Procedures

A total of two classification tasks were administered to the subjects. The first test was the cow, chicken, and grass test. This test was drawn from social psychology and neuroscience because of its increasingly common use in cross-cultural studies. Studies that rely on the cow, chicken, and grass test do so because they provide not only a baseline for cross-cultural research, but also because it can ferret out differences in figure and ground across cultures. Using the cow, chicken, and grass test as a baseline helps establish the degree of "normalization" across the study.

The second classification task asked subjects to organize or classify different types of cancers into specific categories. A series of over 200 different types of cancer were available for classification, each of which could be subdivided into six types. These six different categories included the most common type (frequency), all cancer types, an alphabetical list of cancers (A to Z), cancers by location in the body (body location/system), childhood cancers, cancers common to adolescents and young adults, and cancers specific to women.

The function of these two classification tests was twofold. First, the cow, chicken, and grass test was used to determine the degree to which subjects conformed to current statistical data regarding figure and ground, a measure of field dependence. Second, the classification test derived from the National Cancer Institute site was designed to assess the classification strategies of U.S. and Chinese subjects regarding different cancer types. If the intercultural theory of figure and ground is correct, which states that culture affects the way people classify things in the world, then it is critical to assess the degree to which culture affects classification. Such differences in linguistic categorization have at least some affect on search engines.

Results

The data presented in the following section were obtained from two classification tasks, the cow, chicken, and grass test and the linguistic categorization test derived from the National Cancer Institute site. Data included both frequency and percentage of the specific data set. The first test was analyzed in total. The second test was analyzed for the first four rankings. The purpose of the ranking system for the second categorization test was to determine the degree of field dependence for each cohort. As will become clear in

the following section, the degree of field dependence varied, depending on the sequence in the ranking.

The first test (see Table 1) presents data that conflict, at least to some degree, with the current literature on figure and ground in cross-cultural studies. U.S. subjects present with a 66% rating for the cow and grass category, which suggests a symbiotic relationship. Conversely, Chinese subjects present with a 38% rating for the cow and grass test. On the other end of the spectrum, U.S. subjects present with a 33% rating for the cow and chicken test. Chinese subjects, on the other hand, group the cow and chicken 61% of the time. As will be examined in the Analysis section, these results contradict current literature on linguistic categorization in cross-cultural psychology and neuroscience.

Table 1
Results of the cow, grass, and chicken test between American and Chinese participants.

	American (*n=12*)	Chinese (*n=26*)
Cow and grass	8 (.66)	10 (.38)
Cow and chicken	4 (.33)	26 (.61)

The second ranking of cancer classification (see Table 2) provides the first hint of data that supports the current literature on linguistic categorization across cultures. U.S. subjects present with a 50% response for the Most Common category while Chinese subjects present with a 38% response. The second category, All Types (of cancer), offers a similar data set. Approximately 26% of U.S. subjects present All Types (of cancer) as their preferred categorization strategy while 15% of Chinese subjects present the same category as their preference. Winding out the U.S. cohort was the category of Alphabetic, which rounded out at 33%. The Chinese cohort presents with a 15% response for Alphabetic. The U.S. cohort reported no responses for Body Type, Children, and Women. The Chinese cohort reported a 30% response for Body Type and, like the Americans, no response for Children and Women.

Table 2
The first ranking of cancer classification broken down by American and Chinese participants.

	American (*n=12*)	Chinese (*n=26*)
Most common	6 (.5)	10 (.38)
All types	2 (.26)	4 (.15)
Alphabetic	4 (.33)	4 (.15)
Body type	--	8 (.3)
Children	--	--
Women	--	--

The second tier ranking (see Table 3) offers a slightly different perspective on the differences between U.S. and Chinese subjects regarding linguistic categorization for cancer types. U.S. participants, for instance, presented with a 33% response for the category Most Common. Chinese subjects, on the other hand, presented with a much lower 7%. U.S. subjects presented with a 26% response for All Types while Chinese subjects presented with a response of 15%. Twenty-six percent of U.S. subjects presented with a preference for Alphabetic categorization of cancers for their second tier rankings. Chinese participants presented with a similar ranking of 30% percent for their second tier ranking of All Types. U.S. subjects presented with 26% for categorizing cancer based on Body Type, while Chinese subjects presented with 46% for Body Type. Neither U.S. nor Chinese subjects selected Children for their second tier rankings. The final ranking, Women, was met with a 26% response from U.S. subjects.

Table 3
The second ranking of cancer classification broken down by American and Chinese participants.

	U.S. (*n=12*)	Chinese (*n=26*)
Most common	4 (.33)	2 (.07)
All types	2 (.26)	4 (.15)
Alphabetic	2 (.26)	8 (.3)
Body type	2 (.26)	12 (.46)
Children	--	--
Women	2 (.26)	--

Third tier rankings (see Table 4) present a slightly different portrait of the results. U.S. subjects reported no results for the Most Common categorization while Chinese subjects presented with 38%. U.S. subjects reported a 26% response for All Types of cancer while Chinese subjects reported a 15% response. U.S. and Chinese subjects were nearly identical for the Alphabetic category. U.S. subjects presented with 33% for the Alphabetic category while Chinese subjects presented with 38%. The category for Body Type saw a slightly different result, with 50% for American subjects and 0% for Chinese subjects. The category for Children saw 0% for both U.S. and Chinese subjects. U.S. subjects reported 0% response for the third tier ranking for the category of Women, while Chinese subjects reported 7%.

Table 4

Third ranking of cancer classification broken down by U.S. and Chinese participants.

	U.S. (*n=12*)	Chinese (*n=26*)
Most common	--	10 (.38)
All types	2 (.26)	4 (.15)
Alphabetic	4 (.33)	10 (.38)
Body type	6 (.5)	--
Children	--	--
Women	--	2 (.07)

The fourth and final category (see Table 5) is an aggregate of the first and second tier rankings. The reason for this is because the results for the first two tiers were difficult to analyze unless they were combined. But when combined, the results are stunning because they appear to confirm at least some cross-cultural research in social psychology and neuroscience. U.S. subjects present with a 41% response for the Most Common category of cancer categorization. Conversely, Chinese subjects present with a 23% for the same Most Common category. U.S. subjects report a 26% response rate for the All Types category while Chinese subjects report a 15% response for the same group. U.S. and Chinese responses were nearly identical for the Alphabetic category. U.S.s present with a 25% rate for the Alphabetic category while Chinese subjects present with a 23% rate. The category of Body Type presents a slightly different perspective. U.S. subjects present with an 8% rate while Chinese subjects present with a 38% rate. Neither U.S. nor Chinese subjects reported any response for the Children category for the first two tiers of rankings. The final category, Women, saw a slight difference. U.S. participants present a 26% response while Chinese subjects present a 0% response.

Table 5
Aggregate of the first and second tier rankings of cancer classification broken down by U.S. and Chinese participants.

	U.S. (*n=12*)	Chinese (*n=26*)
Most common	10 (.41)	12 (.23)
All types	4 (.26)	8 (.15)
Alphabetic	6 (.25)	12 (.23)
Body type	2 (.08)	20 (.38)
Children	--	--
Women	2 (.26)	--

In summary, the results were partitioned into two distinct sections. The first section presents the results from the first categorization test—the chicken, cow, and grass test. The second section presents the results from the second categorization test. The first subdivision of this part examines the first three tier rankings and a fourth aggregate of the first two tiers with the aim of surfacing meaningful differences between U.S. and Chinese subjects.

Analysis

The initial results of this study did not support the theory that culture affects linguistic categorization. Based on work in social psychology and neuroscience (Masuda & Nisbett 2001; Nisbett 2003; Nisbett & Masuda 2003; Nisbett, Choi, Peng, & Norenzayan 2001; Paulesu et al 2000), culture may be implicated in explaining the way people from different cultures organize things in the world. In particular, this extant research advances the claim that one's environment shapes or "conditions" the way people view the world. Subjects from Asian countries such as Japan, Korea, and China have been implicated as responding to greater field dependence. This means that Asian subjects observe, and sometimes recall, a large focal object only in respect to its original backdrop. An eastern subject is less likely to recall a frog (figure) if it is presented against a novel background (ground). Conversely, western subjects remains somewhat indifferent to a figure's background. If this research is believed to be true, a frog presented against a novel background presents fewer obstacles for western subjects.

The data from this study did not agree with these results. Instead, these data support a counterintuitive claim that culture is inversely related to categorization. U.S. subjects, for example, present with a 66% response for organizing the cow and grass. Chinese subjects, on the other hand, present with a 61% response for organizing the cow and chicken. The theoretical inference from these data are difficult to rationalize. Based on these results, U.S. subjects are more likely to organize things in the world based on relationships. This is an astounding claim, as it conflicts with current research. Similarly, Chinese subjects report a greater interest in organizing things in the world based on categories. Unlike grass, a cow and a chicken share the same properties of being in a class of animals. If Chinese subjects were more field dependent, according to the theory, then they should be categorizing things like cow and grass together. This is not what happened, which poses serious concern for the theory that culture affects the mind.
The second stage of the study provides a far different result on this theory. This second stage assessed subjects on their categorization approach toward cancer types. These data not only confirm the intercultural theory of mind, but also support it. The U.S. preference for the Most Common category (50%) appears to align with the assumption that native English-speaking cultures are pragmatic, empirical, and inclined toward statistical analysis. Chinese subjects consider the Most Common category 38% of the time, which is a similar rate of analysis. This surely has something to do with the practical and arithmetic value of frequencies for understanding cancer types. Cancer is serious business, and it makes sense to organize them according to frequency.

But the second dimension to the first tier ranking provides a slightly different perspective. This is where the difference between U.S. and Chinese subjects begins to emerge. U.S. subjects present with a 33% response for the Alphabetic category while Chinese subjects present with a 30% response for the Body Type category. On the surface, this may seem like a benign difference. Statistically similar, there seems to be little to quibble about between Alphabetic and Body Type categories. But with a little digging, it is clear that this difference may be tapping into something far deeper than is seen on the surface. It has long been known that Chinese culture and their medicine has relied on what is called a holistic level of analysis. Instead of carving up the body into discrete organ types, as is

common in western medicine, eastern medicine tends to look at a person's entire physiology. This means that a patient is probably subject to a wide range of questions about their lifestyle and health, a method relatively unseen in the west. This difference, perhaps, accounts for the initial dissimilarity between U.S. and Chinese subjects.

Second tier rankings further advance the claim that Chinese participants appear to be more field dependent, based on their interest in organizing things in the world based on background (or body, in this particular instance). Of the six categories available for selection, Chinese subjects picked the Body Type category 46% of the time. This means that Chinese subjects believed that a specific region of the body was the best method for locating information about a specific type of cancer. This not only conflicts with data from the first stage of the study, but it also supports extant claims that culture affects linguistic categorization. Conversely, U.S. subjects selected the Most Common category 33% of the time. Again, this would seem to support the prevailing theory that western culture relies on numerical and statistical data rather than a more holistic way of viewing the world.

The differences found in the second tier rankings deserve special consideration, as they appear to hit on a critical issue. The well-known Chinese preference for acupuncture provides a nice example supporting this point. According to the doctrines of holistic medicine in general and acupuncture in particular, one region of the body is directly connected with other mutually exclusive areas of the body. The earlobe, for instance, may be connected with the tonsils, eyes, cheeks, and is even believed to affect blood pressure. Western medicine is quite different. Instead of seeing distinct organs as connected or interrelated, western physicians and medical schools tend to view the body as a series of mutually exclusive parts. The hand has little relationship with the foot. The earlobe has little relationship with the heart. It is this discrete difference in categorization, or carving up the body if you will, that separates western and eastern forms of medicine. And it is this difference that may very well account for the data differences in the second tier rankings. If correct, examples such as this may offer important clues about search engines for a global audience.

The last section to be analyzed is based on an aggregate of the first two tier rankings. The reason for this grouping is based on the fact that results across both groups may become particularly vivid when compared. In fact, when the data from the aggregate of the first two tier rankings are analyzed, it is clear that something significant is emerging. U.S. subjects present with a 41% response rate for the Most Common category, which is considerable. This means that U.S. subjects believe that the best way to organize a large corpus of cancer categories is through frequency. This makes sense because a general user of the National Cancer Institute site is likely a cancer patient or relation of someone recently diagnosed with cancer. On the Chinese side of the spectrum, subjects reported a preference for the category of Body Type (38%). This response is equally telling in that it supports the claim that eastern approaches toward physiology are based on a holistic and more field dependent style of categorization.

The implications of these data may have a profound effect on search full-text search algorithms. One of the most significant problems facing search engines is the task of presenting user-defined information across cultures. It is one thing to retrieve results based on frequency, popularity, or other statistical functions, but it is another matter to retrieve results based on user-specific content. It should be no surprise that cultural and linguistic factors must be considered to accomplish this goal.

Discussion

The goal of this study was to determine what, if any, relationship existed between culture and linguistic categorization and whether these differences surfaced in theory and practice. The guiding assumption of this study was that culture must have at least some effect on the way people carve up the world, which in turn affects the way people label things. If true, such results would appear to affect search engines.

For roughly sixty years, culture has been implicated in all kinds of effects on the way individual people think and behave. The most promising of this research has occurred in the past decade where very clear distinctions have been made. Numerous studies have concluded that culture unmistakably affects how people perceive and interact with the world. Such distinctions would seem to have enormous ramifications on social life, and that certainly would seem to be true for online interaction and search engines. Studying this phenomenon is not easy given the enormous difficulties of trying to normalize international and intercultural variables. The problem is compounded when trying to assess linguistic categorization for search engines.

One limitation of the study is based on a simple demographic problem. People who visit cancer sites, such as the one on which this study was based, are usually patients or friends and family of the recently diagnosed. Except for one subject who was recovering from a highly treatable form of cancer, this study relied on subjects who were in perfect health. This creates obvious difficulties when trying to assess the true nature of linguistic categorization for a site that aims to disseminate information on cancer.

The most interesting data of this study came not from the first stage but from the second. Here, subjects present with some rather interesting results that support current research in social psychology and neuroscience. Specifically, these data support the claim that culture affects linguistic categorization, which may influence the practicality of some search engines. U.S. subjects, for instance, report a 41% response rate for the Most Common category. According to recent work in social psychology, these results seem to conform to current research on the topic. Chinese subjects, on the other hand, present with a 38% response rate for Body Type. This result, while unexpected, clearly has a connection with current research in neuroscience.

While the theoretical implications of these data are unclear, there may be some connection with current medicinal practices among U.S. and Chinese cultures. U.S. and western culture have a history of dividing the body into discrete organ systems. The heart functions without respect to the liver, or so the western model seems to suggest. Chinese and eastern cultures have a history of taking in the body holistically. That is, such

approaches toward physiology seem to be of a slightly higher altitude whereby the heart has a clear link with the liver. The notion that organs are discrete or unrelated to other organ systems is a foreign concept in many of the cultures found in the east.

Finally, the issue of linguistic categorization for global search engines is actually made more complicated. While there appears to be some relationship between language, culture, and search engines, there are no clearly defined benchmarks for advancing the problem. These data do, however, make a good case for furthering research in this area. If information really has gone global, as many scholars suggest, then it is absolutely critical to address the growing market of content in the online environment.

References

Aristotle. (2001). Categories. In *Basic works of Aristotle* (Ed. McKeon). New York, NY: The Modern Library.

Chua, H.F., Boland, J.E., & Nisbett, R.E. (2005). Cultural variation in eye movements during scene perception. *Proceedings of the National Academy of Sciences, 102*(35), 12629-12633.

Connor, U. (1996). *Contrastive rhetoric: cross-cultural aspects of second language writing*. Cambridge, MA. Cambridge University Press.

Edelman, G. (2007). *Second nature: brain science and human knowledge*. New Haven, CT: Yale University Press.

Hannah, F.C., Boland, J.E., & Nisbett, R.E. (2005). Cultural variation in eye movements during scene perception. *Proceedings of the National Academy of Sciences, 102*(35), 12629-12633.

Harter, S. (1992). Psychological relevance and computing science. *Journal of the American Society for Information Science, 43*(9), 602 - 615.

Hedden, T., Ketay, S., Aron, A., Markus, H. R., and Gabrieli, J. D. E. (2008). Cultural influences on neural substrates of attentional control. *Psychological Science, 19*(1), 12-17.

Kant, I. (2008). *Critique of pure reason*. New York, NY: Penguin.

Masuda, T., & Nisbett, R.E. (2001). Attending holistically versus analytically: comparing the context sensitivity of Japanese and Americans. *Journal of Personality and Social Psychology, 81*, 992-934.

McCool, M. (2008). Negotiating the design of globally networked learning environments: the case of a collaborate online learning module about the Sonoran biosphere. In *Designing global learning environments: visionary partnerships, policies, and pedagogies* (Ed. Starke-Meyerring and Wilson, (pp.200-217). Amsterdam, The Netherlands: Sense Publishers.

Nisbett, R.E. (2003). *The geography of thought*. New York, NY: The Free Press.

Nisbett, R.E., & Masuda, T. (2003). Culture and point of view. *Proceedings of the National Academies of Sciences, 100*(19), 11263-11170.

Nisbett, R., Choi, I., Peng, K., & Norenzayan, A. (2001). Culture and systems of thought: holistic versus analytic cognition. *Psychological Review, 108*, 291-310.

Norenzayan, A., Choi, I., & Nisbett, R.E. (2002). Cultural similarities and differences in social inference: evidence from behavioral predictions and lay theories of behavior. *Personality and Social Psychology Bulletin, 28*, 109-120.

Paulesu, E., McCrory, E., Fazio, F., Menoncello, L., Brunswick, N. Cappa, S. F., Cotelli, M., Cossu, G., Corte, F., Lorusso, M., Pesenti, S., Gallagher, A., Perani, D., Price, C., Frith, C. D., and Frith, U. (2000). A cultural effect on brain function. *Nature Neuroscience, 3*(1), 91-96.

Peng, K., & Nisbett, R. (1999). Culture, dialectics, and reasoning about contradiction. *American Psycholinguist, 54*, 741-754.

Pinker, Steven. (2003) Language as an adaptation to the cognitive niche. In M. Christiansen & S. Kirby (Eds.), *Language evolution: states of the art* (pp.16-37). New York: Oxford University Press.

Pinker, S. (2005). So how does the mind work? *Mind and Language, 20*(1), 1-24.

Pullum, G. K. (1991). *The great Eskimo vocabulary hoax and other irreverent essays on the study of language*. Chicago: University of Chicago Press.

Taylor, J.R. (2003). *Linguistic categorization*. New York, NY: Oxford University Press.

Wittgenstein, L. (1965). *The blue and brown books*. San Francisco, CA: HarperPerennial.

Matthew McCool works in the Department of ETCMA at Southern Polytechnic State University in Atlanta, GA. He can be reached at mmccool@spsu.edu.