Abstract

Semantic and syntactical judgments play a central role in the development of classification systems. Syndetic relationships are relationships between indexing terms, sometimes referred to as cross references. Due to the extensiveness of synonyms for different concepts in natural languages, subject indexing has historically focused on developing controlled vocabularies. A controlled vocabulary limits and prioritizes the terms that a searcher can choose from. This strategy has certain limitations in that it can redirect searchers away from the items that would best fulfill their information needs. Matching relationships tend to focus on nouns and noun type verbs. This article discusses how modern search engines have expanded the syntactical aspect of searching, opening new avenues beyond purely semantic concerns. The discussion then turns to an ontological inquiry into certain philosophical implications of these developments. An assessment of the direction for future research is provided.

Keywords: Classification, Controlled Vocabularies, Synonyms, Semantics, Syntactic Relationships

Introduction

There are many different kinds of items in the world, and there is information about them. Classification schemes are attempts to impose order on the sources of information that we encounter in our daily experiences. When we classify, we group, divide, and name, among other activities, and a structure develops. Structures have various relationships that exist between their component elements, or parts. When classification schemes are developed it is usually asserted that the system is concept based. (Marcella & Newton 1994, p.133-136) The terms that are employed to do the work of the system are synonyms for various concepts. There are, in many cases, varying types of notation for the terms. The descending order by ontological priority appears to be as follows: concept – term – notation. The entire inventory of concepts to which terms (or words) have been applied in a given language is sometimes known as the lexicon. According to this picture, concepts are the basic level at which structures are built. It is often said that concepts express relationships between subjects, and searching by subject has been the paradigm model for information retrieval studies. (Foskett 1996) Libraries and information centers typically arrange their collections by subject matter.

Subject search results are expected to refer to the appropriate items or entities; therefore semantics has historically been a focus of interest for classification development. Semantic relationships are usually placed in a hierarchic order. The two most common hierarchies are thing - kind (taxonomic) and part - whole (partonomic). (Tversky 1989) A common taxonomic example will take the following form: flowers – Larkspur, Hyacinth, Lily. By contrast, a common partonomic example will appear as follows: atoms – electrons, neutrinos, leptons.
Instantiation is a third type of hierarchic order. In this scenario an instance is presented from a broader class. For example, Lake Erie is an instance of the Great Lakes. In general, a semantic relationship can be inferred between terms by asking whether a term is a kind or part or instance of another.

A concept has relationships with other concepts, and the labels that are applied to concepts have relationships with other labels, or terms. The relationship between the terms is generally called a syntactic one. Syntax is commonly understood as any system of rules for combining terms in a natural language. The word order in sentences and the status of these words (noun, verb, etc.) is determined by syntactical considerations. One example of a purely syntactical relationship is “collection development”, as these terms do not have a semantic equivalence in the way that stars and planets (subsets of astronomical objects) or fruits and vegetables (subsets of plants) do. The latter have a semantic relationship – the terms “stand for” the concepts that they label.

Past research on the process of individuals learning classification schemes and using information retrieval systems often centered around what were called “information needs”, which were understood to be the desires or intentions of searchers to locate information. In recent years it has been suggested that information needs is a misnomer for information behaviors. (Case 2012, p.76-81) A need is something that at least partly exists in a person’s thoughts. A behavior by contrast is something that is acted out and can be directly observed. These behaviors can be understood as various searching strategies. Early information behavior theorists focused on the need element that is “hidden” in the mind, in the sense of not being directly observable. When information needs are expressed as questions or queries, they become observable, and can be assessed for comparative search strategy effectiveness.

Information Concepts

There are in this tradition skeptical claims that one can never be certain that what two or more people mean by a term is the same concept. (Kallestrup 2012, p.116-117) And it is true that an information need can only be met if there are relevant documents or information that exist to address it. In regards to labels, a concept can have a variety of synonyms corresponding to it. English in particular is a natural language that is rich in synonyms. As one would expect, information seekers tend to use the terms that occur to them, rather than a classification scheme or retrieval system of which they might be unaware or disinclined to use. Even if they are aware of these resources, a classifier and a natural language seeker may not choose the same terms.

Due to these concerns, the focus has historically been on subject indexing, which typically results in lists of key words for indexed documents. Indexing maps the contents of an item onto the searchable structure used to find that item. If a searcher uses concepts about the item that are not reflected in the search structure, the item will not be found. Cross references and the relationships between indexing terms are referred to as syndetic relationships; these include broader than, narrower than, equivalence, and associative. In a controlled vocabulary (as discussed below), a preferred term is chosen from among equivalents or synonyms. The standard set of thesaural cross references are the following.

- USE: A reference from a non-preferred term for a preferred term
- UF: Use for: a reference from a preferred term for a non-preferred term
- BT: Broader term: terms which are more general in scope
• NT: Narrower term: terms which are more specific in scope
• RT: An associative term that is related in a way other than BT or NT (Broughton 2006, p.54)

USE and UF are the devices for vocabulary control, and denote synonyms. BT and NT relationships define a classification hierarchy where the broader term is associated with a class and the narrower terms are instances of this class. RT relationships are associative and are dependent upon the specific context and particular needs of information searchers. They are not items of a BT/NT inverse relationship and they are not synonyms.

Associative relationships are relationships between terms that are neither equivalent nor hierarchical. Terms that are mutually exclusive or could not be confused with one another are not considered associative. Associative relationships are employed between separate concepts that may be confused for each other. One question that can be raised in this context is “if someone is interested in this, might they also be interested in that”. These terms are sometimes labeled as “see also”. Terms in an associative relationship are determined by syntactical considerations and are often defined negatively, as whatever is non-hierarchical or non-equivalent. Defining something in a negative way usually introduces a high level of ambiguity. Often the related terms may have overlapping meanings.

Semantic And Syntactic Relationships

The assumption that informs this approach is that natural languages are deceptive. Synonyms can cause us to miss relevant information; homonyms can cause us to retrieve irrelevant information. A controlled indexing language restricts the number of terms that an indexer or an indexing mechanism, and a user of the index, can choose from. The motivation for doing so is to limit synonyms, thereby requiring searchers to use the same words for a concept. Judgements are made about which words are “useless” or words that no one would search for. Preferred terms are established for a concept and are often called “descriptors”.

The privileged term is the descriptor; the other terms are variants or “Used For”. One may sense how this could be misleading in a variety of cases. The index defines items in a sense more so than the item's content, because the index is how an item is found. If an algorithm process eliminates a processing token before it is indexed then a search query will not find the items in the database, even if they are there. For example, an indexer may choose what she considers to be representative of American Comedians. Consider a Parisian who wanted information on someone she had seen on late night television, Ellen Degeneres. If, when searching the database compiled by that indexer, her search for American Comedians is redirected to Bill Cosby, the information need will not have been met. In a similar example, consider “Cars: Use For Automobiles”. If a searcher for railroad cars is automatically referred to automobiles, she will not find what she is looking for.

In the controlled language, or limited expression vocabularies, it was often recommended for search retrieval that only the present tense be used. Auxiliary verbs (should, would, could) were also avoided. Only the present, immediate demand for the properly named object was considered an optimal search strategy. The vast computing power of modern search engines, with the ability to produce near spellings and other combinatorics instantly has made much of these guidelines obsolete. What continues to be of interest, however, is the way that these ideas reflect the societal preoccupations in which they were produced. Once a relationship or object is
labeled in a certain way, its nature is defined, and the roles that it may occupy are circumscribed. The central goal in topic matching and best matching theories is to retrieve all but only those items in a collection considered to be relevant to a specific query. (Bean & Green 2001, p.117-119) One of the primary reasons for the vast improvement for searching in Google and other automated search engines is that over-regulation of indexical reference is not necessary.

When developing a controlled vocabulary for classification purposes, words such as “an, another, out, the, up, and,” etc. are not considered words that searchers use, they are not sought terms. Articles, conjunctions, and prepositions were also not considered significant words. Adjectives and adverbs were avoided, if possible. What has been favored are nouns and noun forms of verbs because (it was claimed) these are what searchers look for. Nouns and noun-type verbs (gerunds) were said to be open classes, allowing the introduction of new members. Pronouns, prepositions, and conjunctions were said to be closed classes, in that nothing new can be added. And yet, considered from the point of view of relationships, the hierarchical and equivalent relationships are closed, whereas the relationships expressed by the syntagmatic terms are open, allowing an indefinite number of combinations (see table 1). As a search strategy this would have been unwieldly in the pre-computing era, but it is entirely possible in the current automated environment.

Matching relationships between noun terms as they have been commonly understood are semantic relationships, in the sense that they could be substituted for each other in the same part of a sentence, such as the subject of predication. This differs from syntagmatic relationships in that syntax, as has been mentioned, is concerned with the way that words work with the other words in a sentence. This distinction can be made clearer by the use of a table that places the semantic relationships of terms in a proposition on a vertical axis and the syntagmatic relationships on a horizontal axis (Table 1). In the following example, sandals, shoes, and boots stand in the relation of hyponomy because they all belong to the same semantic group: footwear. Likewise, turquoise, red, and brown are hyponyms of the color group. Margaret and Linda are examples of humanity. The substitution words for “bought” and “pair” are synonyms (acquired, couple) and antonyms (sold, single) which are semantic relationships.

**Table 1**

In this example, substituting the terms that have semantic relationships does not affect the syntactical pattern. The grammatical picture that results from this use of syntactical expressions allows for an open combination of terms.

<table>
<thead>
<tr>
<th></th>
<th>acquired</th>
<th>couple</th>
<th>Margaret</th>
<th>bought</th>
<th>a</th>
<th>pair</th>
<th>of</th>
<th>turquoise</th>
<th>sandals</th>
<th>for</th>
<th>Linda</th>
</tr>
</thead>
<tbody>
<tr>
<td>Linda</td>
<td>sold</td>
<td>set</td>
<td>blue</td>
<td>boots</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Margaret</td>
</tr>
<tr>
<td></td>
<td>pawned</td>
<td></td>
<td>red</td>
<td>shoes</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

The syntagmatic words in this example (a, of, for) are not about meaning, they are about grammatical patterns. Rather than allocating words to a semantic category, the syntagmatic relationships collate the terms in a certain direction. They are not collected into sets as single key words, but must be focused on as phrases and clauses. Searchers for information are not walking dictionaries (filled with singular terms); they are searching for items that usually require more
A philosophical question that can be raised at this point is whether any given relationship under consideration is language independent. One of the reasons that the information retrieval paradigm has been strongly in favor of nouns and noun forms is the appeal of set relationships such as transitivity, Boolean, and Vector models. These metrics provide an aura or semblance of objectivity in contrast to the subjectivism that language dependence seems to imply. However, the collection of items and the assumed choices that accompanied the selection of these items into sets can obscure the actual variation that is involved within natural languages. The application of mathematical relationships to common language use may produce a misleading context for the organic patterns that searchers sometimes follow.

One of the assumptions that may work toward not giving syntagmatic patterns adequate attention is the association of choice with the noun concepts. In sense perception, one sees the noun concepts, but there is no particular sensible content for “one, and, all, it, then, or, no,” etc. Meaning appears to be created when an item represented by the noun is chosen. Much of the research that has focused on this process has centered around the concept of relevance, which is subject to user interpretation. (Buck, 2017). However, as some philosopher have noted, the particulars themselves are “unintelligible” – in the context of information studies one might say unrecognizable - without categorical form. The structure within which choices are made and in fact makes the choice possible may not be noticed or is neglected. A classification scheme in which noun-objects are not syntactically structured presents a nominalized view of content. When this is coupled with a controlled vocabulary the effect is a compression and redirect of meaning that may limit the possibilities that could satisfy a search query.

Considering language relationships in this way has certain affinities with discussions about apperception Although there are a number of senses to this term in different disciplines, it is understood here as the awareness of one’s own ideas or sensations. This occurs when searchers are aware that they are searching for a particular result and that the item (for example) both has a medium and that the medium is print, digital, audio, etc. Notice the ontological character of this context, in which the transitivity relationship does have some bearing on search activities. Consider three search items labeled A, B, and C that are obtained by a searcher. If the same searcher that locates A also locates B, then this is also the searcher that locates C. If the searcher is conscious of the items as what she in fact intended, then she is also conscious of her own identity as the subject that has successfully executed the search requests. The ways in which linguistic elements are combined to achieve results through queries connects the searcher with the retrieval item in one common link of associations. As the influential philosopher Edmund Husserl points out, syncategorematic terms provide ordering within which categorematic meanings can be combined. (Lampert 1995, p.88-107) The dependent character of syncategorematic terms both requires and stimulates the activities of classification.

Categorical forms are not objects of sense perception, they must be thought; concepts are not simply noun referents, they involve and reflect the activity of thinking. It has been said that syntactical expressions change animal sounds to human speech. They do so by allowing wholes to be built from parts, the categorical parts of language. Our interpretations transform marks into meaningful signs. This is not to diminish the importance of concrete nouns in searching.
Particularly in regards to place, it is important to have the right name for a hospital, school, residence, or other tactical structures. One can also see how deception can play a strategic role, for example in a military operation, providing the enemy with the wrong location and so forth. An enemy can be understood in a number of ways, such as a competitor, intruder, rival, etc., which is always identified in the singular as a noun. Likewise explosive materials and other ordnance are often identified by the so called “mass nouns”.

In natural languages, the meanings of words “drift” over time, usually due to an associative connection with other words. (Sinclair 2004, p.155) Academic disciplines and other professions select specific words and develop them into terms, assigning them a fixed meaning. Terms or words that are not adopted by an institution have various patterns of use within the natural languages. In closely related methods to how classification schemes used controlled vocabularies, information retrieval systems often establish a preferred term among synonyms, a process sometimes called “disambiguation”. Among other effects, this results in an approximation between the vocabulary of an information retrieval system and the naturally occurring data, which varies over time and place. From this point of view the determination of preferred terms for a concept is a delimitation of meaning from the environment in which it developed.

Finally, it is important to note that there are cases where quantifiers like “everything” and “something” relate to semantic content rather than syntactic categories. “That - clauses” can occur in referential positions. Consider the following examples:

*Linda imagined something that Margaret might never suggest.*

*Margaret suggested everything that Linda had long thought about.*

What the clauses refer to in these examples is the propositional content that both Linda and Margaret share. Moreover, they refer to something that would not be discovered based upon the proper names that share the content. The view that such expressions are non-referring because they are not nouns that pick out a particular object is misguided.

Although technical advances have made it easier to understand how words that were variously considered “useless”, as “incomplete” in their ability to refer to objects, are words that do have a role to play in making “sought terms” possible, there were philosophers in earlier ages who had a strong interest in syntactical relations. G.W. Leibniz, as one example, discussed the possibility of an alphabet of human thought in which all concepts could be understood as combinations of simple elements. (Lampert 1995, p.103) In recent times there has been an emphasis on the cultural assumptions that inform classification systems. The semantic foundations are often determined by decisions that reflect societal and cultural values. Careful attention can reveal the biases, motivations, and opinions of the cultures in which they are created. Terms and concepts that are christened as “useful”, “important”, “valuable”, and so forth are made primary in the system, but should remain open to inquiry and revision. These trends reflect the development of a more outward ontological orientation than a controlled vocabulary presupposes.

**Concluding Remarks**

Certain types of problems are endemic to systems that utilize intermediaries that limit what is accessible. A selection model of whatever kind, whether it be a classification scheme or
information retrieval system, will have presuppositions built into the structure, and there is the possibility that what is not included is precisely what the searcher had been seeking. The standard picture for information retrieval before the advent of the internet was topic matching among nouns, establishing preferred terms among synonyms. Doubt was often cast upon a searcher’s ability to adequately express previously acquired knowledge. Although most individuals have an internalized view of the world, they may not have the ability to coherently express themselves, making a controlled vocabulary necessary. Apart from the somewhat parochial attitude implied here, the ontological commitments of this view - one might say the inward ontological orientation - favored particularized content over relations. However, technological advances have made it ever more possible for searchers to utilize a personalized view of the world without recourse to standardized ontologies. It is clear from trends that future research will focus on relevant information that is related inferentially, or by analogy, but may not contain the same particular terms. The goal is to open the field of human cognitive associations, rather than to enforce the vocabulary of a selection model. But what is also important is the recognition of limits about what can or should be made available. The implicit unfairness of a demand that every request lead to a present, satisfactory result, even when no such results are possible, should be recognized and accepted.

BIBLIOGRAPHY


