A Hybrid Approach to Faceted Classification Based on Analysis of Descriptor Suffixes

Aaron Loehrlein, Corresponding Author
Classification-based Search and Knowledge Discovery, Indiana University, Bloomington, IN 47405.
aloehrle@indiana.edu

Elin K. Jacob
Classification-based Search and Knowledge Discovery, Indiana University, Bloomington, IN 47405.
ejacob@indiana.edu

Kiduk Yang
Classification-based Search and Knowledge Discovery, Indiana University, Bloomington, IN 47405.
kiyang@indiana.edu

Seungmin Lee
Classification-based Search and Knowledge Discovery, Indiana University, Bloomington, IN 47405.
seungmin@indiana.edu

Ning Yu
Classification-based Search and Knowledge Discovery, Indiana University, Bloomington, IN 47405. nyu@indiana.edu

This study explores the construction of a faceted vocabulary that can be used as a mechanism for organizing Web-based resources. After analyzing the manual process of faceted vocabulary construction using existing organizational structures to identify heuristics for automating the construction process, we modeled a hybrid, semi-automatic approach to facet generation that integrates the strengths of manual and automatic methods. This approach is based on the organization of terms according to the meanings of their suffixes. Although this heuristic could not usefully organize the majority of terms of the lexicon bases to which it was applied, it nevertheless shows promise as a component of hybrid classification.

Introduction

Enumerative classification schemes have long provided an effective tool for organizing a collection of resources by assigning each resource to a single class in a set of predefined and mutually exclusive classes. Although librarians have traditionally relied on classification schemes such as the Library of Congress Classification (LCC) and the Dewey Decimal Classification (DDC) to provide access to physical resources, the use of machine-based full-text searching undermined the perceived utility of classification for information discovery and retrieval. However, growing frustration with the huge retrieval sets and numerous false drops that accompanied do-it-yourself searching on the Web has generated renewed interest in classification, categorization and the power of controlled vocabularies -- interest reflected across the Web landscape, from Web directories such as Yahoo! to metadata initiatives associated with digital libraries and the Semantic Web.

There are many challenges to a classification-based approach to organizing the Web. For example, it is impossible to "organize" the whole Web due to its massive size and the diversity of Web resources. Even if such a feat were feasible, clustering approaches are not incremental and text categorization approaches are based on a static classification scheme, rendering them unable to deal with the dynamic nature of the Web corpus. A highly variable and dynamic environment such as the Web requires an organizational approach that not only provides flexibility of representation and accommodates the dynamic nature of human knowledge itself but is also able to respond to the information needs of a highly diverse and increasingly interdisciplinary population.

Because traditional classification schemes attempt to enumerate all knowledge in a given domain within a fixed set of

predetermined classes, they are ill-suited for organizing resources in the diverse and multidisciplinary environment of the Web. Recognizing the inherent rigidity of traditional enumerative structures, Ranganathan (1944; 1945) proposed a more flexible approach to organization that represented knowledge not as a set of static classes but as a set of concepts and relationships. This approach identifies the various aspects (characteristics or facets) of a given domain so as to derive a set of independent concept hierarchies that represent the range of characteristics relevant to that domain. Each such concept hierarchy is populated by the set of possible values (or isolates) that are used to describe that aspect for a given resource. Classes are created by combining isolates from this controlled vocabulary according to an established citation order, assuring collocation of related resources within a dynamically-generated hierarchy (Jacob and Priss, 2001). Thus, construction of a faceted organizational scheme neither prescribes a finite set of classes nor predetermines the relationships among classes. Rather, it establishes control over the formal semantics underlying the scheme and, in so doing, provides a conceptual basis for both the formation of classes and the establishment of relationships among the classes that comprise the resulting classification structure.

The dynamic and adaptive nature of a faceted vocabulary is more effective in organizing Web documents than traditional classification schemes that establish a fixed set of predefined and static classes. However, the manual construction of a faceted vocabulary is a resource-intensive process requiring considerable intellectual effort and its implementation on the Web is impractical.

One goal of the Classification-based Search and Knowledge Discovery (CSKD) research group is the exploitation of an existing body of manually classified documents to enhance information retrieval and knowledge discovery on the Web. CSKD research explores methods for leveraging both the ontological and link-structure knowledge embedded in classified corpora of Web documents for searching and organizing the Web. It is a multi-dimensional project that entails investigations in such area as machine learning, classification, clustering, link analysis, and fusion. Current projects include JiTTDL (The Just-in-Time Teaching Digital Library <http://134.68.135.1/jit/index.shtml>), where the CSKD approach is applied in building a digital library of teaching resources. The goal of this research is to discover a semi-automated method of faceted vocabulary construction that will make such an approach more viable for organizing collections of web-based resources.

This paper describes work in progress that investigates automated methods for streamlining and standardizing the process of constructing a faceted vocabulary. More specifically, it reports on an investigation of a semi-automated method that exploits suffixes and other ending strings of terms.

Construction of a Faceted Vocabulary

The fundamental organizing principles underlying the development of a faceted system are the grouping of that which is related and the separation of that which is unrelated. Unlike the fixed structure of classes produced by enumerative classification, faceting provides for the organization of concepts in modular hierarchies by splitting (separating) unrelated or dissimilar concepts and lumping (grouping) related or similar concepts. Relevant concepts are identified by partitioning domain terminology into mutually-exclusive baseline facets (Priss and Jacob, 1998) that are subsequently combined to form higher-order facets. Typically, development of the faceted vocabulary is an iterative process of analyzing a domain vocabulary and identifying clusters of relevant values (Batty, 1989). Initial clusters of values are aggregated into progressively more comprehensive groupings that identify general concepts and provide the initial set of baseline facets. These baseline facets are then combined to form modular hierarchies of superordinate facets. To create a classification scheme, values from this modular vocabulary are joined according to a standardized combinatorial order, generating a hierarchical structure of classes. In this way, a faceted structure of concepts and concept values ensures consistency of representation and coherence of structure within individual facets -- each concept appears only once in the vocabulary -- while assuring that the facets and the relationships between facets remain adaptable to context and usage (Priss and Jacob, 1998).

A Hybrid Approach to Faceted Vocabulary Construction

The process of constructing a faceted classification scheme is generally described as "analytico-synthetic". Because construction of such a scheme begins with the collection and subsequent grouping of linguistic terms specific to a given domain, the process is generally described as "bottom-up", distinguishing it from the "top-down" process of division employed in the construction of enumerative classification schemes. The development of a faceted vocabulary necessarily begins with analysis of the linguistic terminology of the associated domain; but this analysis may not be effective if executed within a vacuum. For this reason, analysis of domain content should combine inductive (or "bottom-up") acquisition of the linguistic base and deductive (or "top-down") analysis of terms and term relationships based on the domain's conceptual framework. By employing a "middle-out" strategy that integrates

bottom-up and top-down approaches by analyzing the terminology of a domain within its existing conceptual framework (Priss and Jacob, 1998), the resulting vocabulary only identifies the most relevant concepts for the initial set of baseline facets but also maintains the relationships between concepts and concept hierarchies that are most meaningful within the domain context.

Bottom-up creation of a faceted vocabulary is prone to human error and inconsistency. And, because facet creation is intellectually labor-intensive, automation of the development process has not seemed feasible. However, we theorized that using a hybrid, middle-out approach could support automation of facet generation by integrating the processing capabilities of the machine with the analytical and evaluative capabilities of the human. This hybrid approach to facet generation would begin with identification of the heuristics or basic sorting strategies used by humans in the grouping process. Analysis of these heuristics would then indicate which strategies could be handled automatically by the machine to generate a set of candidate facets and values.

Analyzing the Faceted Vocabulary Construction Process

To assess the viability of an integrated, hybrid approach, we decided to begin the process of constructing the faceted vocabulary by identifying a lexicon of concepts from an existing representational system currently used to index a collection of Web documents. The representational system selected for this project was *EPA Topics*, available at <http://www.epa.gov/epahome/topics.html>, an indexing scheme used by the United States Environmental Protection Agency (EPA) to provide access to a collection of high-quality resources dealing with a range of environmental issues. EPA Topics is not a true classification scheme in that not all categories are mutually exclusive and any concept or category may be nested within more than one branch of the hierarchical tree structure. However, this representational system does provide a set of nested categories with each category represented by a chain of descriptors indicating its relationship within the overall hierarchical structure.

Table 1. A subset of the organizational system for suffix/meaning pairs

Suffix Class, First Hierarchical Level	Suffix Class, Second Hierarchical Level	Suffix Class, Third Hierarchical Level	Suffixes & Pseudo-suffixes	Example
abstractions	field, discipline, science	art, process, or science of measuring	*-metry*	*psychometry*
abstractions	field, discipline, science	fields, subjects	*-graphy*	*geography*
actions, processes			*-ence*	*an emergence*
actions, processes	actions		*-ing*	*a blessing*
actions, processes	processes	results of an action	*-ion*	*a rebellion*
actions, processes	processes	results of an action	*-ment*	*the recruitment*
characteristics			*-ive*	*coordinative*
characteristics	an action		*-ative*	*talkative*
characteristics	relating to degrees		*-imum*	*pessimum*
entities	chemicals, chemical compounds		*-dehyde*	*acetaldehyde*
entities	chemicals, chemical compounds	salt or ester of a carboxylic acid	*-oate*	*octanoate*
entities	performing an action		*-or*	*editor*
entities	places	towns, cities	*-boro*	*hillsboro*
entities	places	towns, cities	*-burgh*	*pittsburgh*
entities	proper nouns	companies or organizations	*-co*	*arco*
entities	things	books	*-book*	*handbook*
entities	things	drawings, writings, records	*-gram*	*spectrogram*
states, qualities, conditions			*-ment*	*amazement*
states, qualities, conditions			*-ence*	*dependence*

entities	things	drawings, writings, records	*-gram*	*spectrogram*
states, qualities, conditions			*-ment*	*amazement*
states, qualities, conditions			*-ence*	*dependence*

The first step in generating the faceted scheme involved the creation of two lexicon bases. The first lexicon base consisted of the approximately 700 unique, information-bearing terms in the set of descriptors used in the EPA Topics *category labels*. The frequency with which each term appeared was also recorded. This lexicon base was deemed to be highly relevant to the domain of EPA. The second lexicon base was taken from metadata fields such as *title*and *abstract*that describe the documents and other resources that are organized by EPA Topics. This lexicon base, known as *annotations*, consisted of approximately 13,000 terms and the frequencies with which the terms occurred in the EPA's collection of resources. These terms are presumed to vary widely with respect to their relevance to the specified domain.

The important aspect of this phase was investigation of the sorting heuristics. Based on the assumption that specification of the analytic strategies used by humans in analysis of a domain's lexicon would point to heuristics that could be automated to augment the manual process, we examined the analytic strategies used by two indexers to discover a set of heuristics that could both streamline and standardize the process of creating a faceted vocabulary (Ranganathan, 1944).

The proliferation of indexes and classification schemes and the sheer quantity of resources available on the Web make fully manual approaches to the creation of representational systems too time-consuming and labor-intensive to be practical. By the same token, fully automated classification methods usually cannot match the intellectual quality produced by a trained classificationist unless such approaches are applied to documents that exhibit explicit constraints on the organization of content, as is the case with medical records and many scientific reports (Hahn, Romacker, and Schulz, 2002).

A hybrid classificatory process would combine the strengths of both manual and automatic approaches to the construction of classification systems. It analyzes the steps undertaken by a human classificationist and determines which of those steps, if any, could be automated or could benefit from an automated process. Simple automatic processes that extract terms from a set of documents and other resources could be used to give those terms an initial organization, from which the classificationist can craft a

fully functional classification system. Because this approach seeks to provide a "first draft" of a classification system, it is not necessary for the automatic classificatory processes to correctly place every term.

The Suffix Heuristic

One such classificatory strategy that can be automated is the classification of terms according to their suffixes. This approach differs from previous work with suffixes, some of which has employed stemming heuristics to achieve the conflation of terms (Harman, 1991; Savoy, 1993), while others have identified a term's position within a phrase (Okada, Ando, Lee, Hayashi, and Aoe, 2001). In contrast, this heuristic is based on simple grouping of terms that share a common suffix. Extensions and refinements of this heuristic include such processes as taking into account suffixes with multiple meanings, sets of suffixes that have similar meanings, the extent to which a suffix can usefully organize a given set of terms, and ending strings that are not suffixes but that can be used in a like manner.

Morphemes and Phonaesthemes

Many of these ending strings can be considered morphemes, which are the smallest meaningful units of a word (Bybee, 1988). For example, *-day* in *Wednesday* is a morpheme. Non-suffix morphemes such as *-field* and *-flow* proved useful when applied in the suffix heuristic. Many suffixes are also morphemes. There has been substantial research into the role of morphemes in cognitive organization (*ibid.*). However, to our knowledge, there is no research that seeks to use morphemes in order to create a classification system from existing lexicon bases.

Also potentially of interest are sub-morphemic units known as phonaesthemes. Phonaesthemes are simply components of words that seem to correspond with an aspect of a word's meaning (Bolinger, 1965; Bergen, 2004). For example, the phonaestheme *gl-* appears in many words that "refer to some aspect of light or vision", such as *glance*, *glimmering*, and *glint* (Bergen, *ibid.*). Phonaesthemes that occur at the ends of words are distinct from suffixes in that a phonaestheme's meaning is generally implicit,

while a suffix's meaning is explicit. Also, suffixes are more generalizable. The suffix *-able*can be used to modify a wide variety of verbs, turning them into adjectives. However, the phonaestheme *-ee*, which implies "festivities" (Bolinger, 1965), applies only to a small number of words (*jubilee, jamboree*, etc.), not to social events in general (e.g., it could not modify *reception*). Conversely, words with a phonaestheme generally lose all meaning when that phonaestheme is removed (e.g., *jubil*is meaningless). Research into phonaesthemes has examined the extent to which subjects tend to categorize nonce words (i.e., novel words) by phonaestheme when asked to conjugate the past tense of those words (Bybee and Moder, 1983) and the extent to which subjects can recognize words more quickly if they are grouped by phonaestheme (Bergen, 2004).

There are several reasons why phonaesthemes are ill-suited for hybrid approaches to classification. First, words with common phonaesthemes tend to be similar in ways that are abstract and subtle. For example, the phonaestheme *-amble*groups together *preamble*and *scramble*, two terms that bear obvious similarities involving motion, but which clearly belong to different domains of activity. The same applies to *-awn*, which groups *lawn*and *spawn*. Both appear to represent different aspects of the spreading of life, yet it is difficult to imagine the utility of grouping these terms together in a classification system. Second, many phonaesthemes imply other characteristics of the terms in which they occur, such as a limit of no more than two syllables. It is currently beyond the scope of this research to automatically detect the number of syllables for a given term. Third, many phonaesthemes apply to words with informal connotations. For example, the phonaestheme *sn-*tends to indicate a word that relates to the nose or mouth (Bergen, 2004). Many of these terms, such as *sniggered*, *snippy*, and *sniveling*, are too informal or imprecise to be of much use in a classification system. One should use phonaesthemes with caution when classifying terms. However, phonaesthemes might indirectly support the process of classification by providing clues as to the general tone or bias of a lexicon base in which a specific phonaestheme features heavily.

In our research, we refer to morphemes and other non-suffix ending strings that were used in the suffix heuristic as "pseudo-suffixes". Before the suffixes and pseudo-suffixes in the lexicon bases were analyzed, the terms were altered in only one respect: plural forms were converted to singular forms according to the rules described in Table 2. Many terms, such as *agreements*, have both an ending string that represents a plural (*-s*) and an ending string with a more specific meaning (*-ment*). The removal of the plural string facilitated the analysis of the more meaningful string.

Table 2. Method for the conversion of the term to singular form. This method ignores certain ending strings that do not represent a plural, such as *-polis*and *-ness*.

Suffix	
-s	Truncate the final character (*-s*)
-ses	Truncate the final two characters (*-es*)
-ies	Truncate the final three characters (*-ies*) and add *-y*

Organization by Suffix

The list of suffixes consists of every English-language suffix according to Merriam-Webster Online (2002 <http://unabridged.merriam-webster.com>) and totals 1,571 suffixes. Once the set of suffixes that occurred in at least one of the lexicon bases was identified, each of those suffixes was manually assigned meanings based on their definitions in Merriam-Webster Online. This qualitative approach provided relatively standardized meanings for the suffixes. For example, one of the definitions in Merriam-Webster of the suffix *-al*is "of, relating to, or characterized by", while one of the definitions of *-ative*is "of, relating to, or connected with". Both of these meanings were represented simply as "characterized by". If a suffix had multiple meanings, a separate record was provided for each meaning. For example, the multiple meanings of *-ment*include "the result of an action" and a "state, quality, or condition". In addition, meanings can have multiple suffixes. For example, the concept of "states, qualities, or conditions" can be exemplified by words with the suffixes *-ment, -tude*, and *-ence*. Each record of a suffix meaning was represented as a suffix/meaning pair.

These suffix/meaning pairs were then grouped into general classes of meaning, such as "actions and processes", "characteristics", and "entities". This manual grouping allowed terms to be associated on the basis of shared properties. For example, *-tron* /"device for the manipulation of subatomic particles" and *-gram* /"drawings, writings, records" were both considered types of *things*, which along with such classes as *people*and *places*were considered types of *entities*.

Once the meanings were identified for each suffix that appeared in at least one lexicon base, the terms in the lexicon bases were grouped by suffix in order to determine which of a suffix's multiple meanings, if any, applied to the majority of the terms (see Table 3). The meaning that applied to the majority of terms was flagged so that terms were grouped only according to the specified

suffix/meaning pair. Although most suffix/meaning pairs that were useful in one lexicon base were also useful in the other lexicon base, a separate record of flags was kept for each lexicon base. There were a few cases when both lexicon bases used a suffix, but in different senses. For example, both lexicon bases used the suffix *-ery*, but most terms ending in *-ery* in the *category labels* lexicon base represented "states, qualities, conditions" (e.g., *slavery*), while most terms ending in *-ery* in the *annotations* lexicon base were "places that are characterized by" (e.g., *nursery*).

Table 3. A subset of the terms ending in the suffix *-en*. Because the majority of terms do not correspond to either meaning, this suffix was not used to organize terms.

	Meanings for *-en*		
terms ending in *-en*	cause to be, come to be	made of, consisting of	neither
between			x
broaden	x		
children			x
golden		x	
green			x
heighten	x		
oven			x
raven			x
screen			x
token			x
woolen		x	

Occasionally a term was grouped according to its suffix/meaning pair, but at a more general level of meaning than is typically associated with the suffix. For example, *-oate* /"salt or ester of a carboxylic acid" could be considered to usefully classify terms, but only at the more general level of "chemicals, chemical compounds". Because the former meaning is nested under the latter (which is

in turn nested under *entities*), it was relatively simple to assign a more general level of meaning to a suffix/meaning pair that occurs in a given lexicon base.

Terms were not grouped by a suffix if that suffix was a sub-string of another suffix (e.g., *-ar*is a sub-string of *-lar*). In some cases, the two suffixes have different meanings, such as *-ess*and *-ness*. In cases where both suffixes have the same meaning, the longer suffix usually returns words at a higher level of precision. This gives the classificationist the option of increasing precision at the expense of recall by unflagging the shorter and less successful suffix.

Identification of Pseudo-Suffixes

In order to take fullest advantage of the organizational potential of ending strings of terms, ending strings were identified that occurred in many terms, but did not correspond to genuine suffixes. For every term in the *annotations*lexicon base, several ending strings were identified that ranged in size from two characters in length up to, when applicable, seven characters in length. For example, from the term *memoranda*were extracted the ending strings *-da, -nda, -anda, -randa, -oranda*, and *-moranda*, while from the term *plea*were extracted the ending strings *-ea*and *-lea*. These strings were filtered out if the string itself ended in a suffix. For example, *-dable, -rtable,*and *-ortable*each end in the suffix *-able*and were therefore not considered. In addition, ending strings of terms were filtered out if they were also the ending string of a suffix. For example, the string *-ng*is the final two characters of the suffix *-ing*and was therefore not considered. After filtering out suffixes, ending strings that end in a suffix, and ending strings that are the ends of suffixes, the remaining ending strings were used to group the terms in each lexicon base.

A manual process was used to determine which of these strings produced useful groups. Some ending strings were useful even though they were not meaningful. For example, the ending string *-ento*has no meaning and produced a heterogeneous set of terms, except that most of the terms in the lexicon bases that end in *-ento*are Spanish. Other ending strings of words are themselves words. For example, while the ending string *-book*is a word, not a suffix, it groups together terms such as *notebook*and *yearbook*. Some ending strings appear to be informal contractions, such as *-tech*, which groups terms such as *airtech, biotech*, and *envirotech*. Finally, some ending strings appear to be meaningful even though they are not words or suffixes. For example, Merriam-Webster does not consider the ending string *-illu*to be a suffix and does not contain any words that end in *-illu*. However, this ending string groups the

terms *bacillu*and *aspergillu*, which appear in articles concerning proteins and cellular biology. This sort of ending string is also an example of a pseudo-suffix. Table 4 provides further examples of pseudo-suffixes.

Table 4. A subset of the pseudo-suffixes and their assigned meanings

Pseudo-suffix string	Assigned Meaning	Example	Frequency in *annotations*
-day	days	*yesterday*	11
-doxy	opinion or doctrine	*orthodoxy*	0
-duce	to move or bring into being	*induce*	6
-dum	results of actions related to a document	*addendum*	3
-ello	towns, cities	*Monticello*	3
-ena	proper nouns	*Wadena*	9
-endo	non-English	*aprendiendo*	3
-ente	proper nouns	*Fuente*	9
-ern	characteristics	*western*	21
-ett	proper nouns	*Everett*	12
-ez	proper nouns	*Suarez*	9
-field	places	*Mansfield*	27
-flow	characterized by flowing	*inflow*	7
-group	groups	*workgroup*	3
-heim	proper nouns	*Mannheim*	3
-house	houses	*warehouse*	11
-illu	related to genetics	*bacillu*	3

Findings

The effectiveness of the suffix heuristic was evaluated against both the *category labels* lexicon base, which was small but highly-relevant, and to the *annotations* lexicon base, which consisted of a larger set of terms that ranged from highly-relevant to irrelevant. In *annotations*, 168 suffixes and pseudo-suffixes organized 42.66% of the 13,189 terms into 105 groups (see Table 5). In the *category labels* lexicon base, sixty-eight suffixes and pseudo-suffixes organized 49.79% of the 723 terms into forty-five groups. Of those, genuine (i.e., non-pseudo) suffixes accounted for 85.07% of terms and 57.14% of the groups in the *annotations* lexicon base and 90.23% of the terms and 66.67% of the groups in the *category labels* lexicon base. These results indicate that the suffix heuristic provided an initial classification for approximately half of the terms in each lexicon base. The identification of pseudo-suffixes did not result in a notable increase in the number of terms organized, though it did appear to dramatically increase the number of groups into which the terms were placed. In addition, of the 174 "final cluster" (i.e., ending string) phonaesthemes that appear in the Dictionary of English Phonaesthemes (Shisler, 1997), none produced groups in which at least half of the terms were appropriated placed in either lexicon base. The only pseudo-suffixes that were effective were those that we had previously identified from one of the lexicon bases. Overall, it appears that suffixes alone are not sufficient when seeking to organize a large number of terms by their meanings.

Table 5. Results of the suffix heuristic, in terms of number of suffixes used, number of terms organized into groups, and total number of groups.

Lexicon base	Suffixes and pseudo-suffixes used	Number of terms organized	Number of terms in lexicon base	Percent of terms organized	Number of groups of terms
Category labels					
suffixes	48	325	723	44.93%	30
pseudo-suffixes	20	35	723	4.86%	15
Total	**68**	**360**	**723**	**49.79%**	**45**
Annotations					
suffixes	87	4786	13,189	36.29%	45
pseudo-suffixes	81	840	13,189	6.37%	60
Total	**168**	**5626**	**13,189**	**42.66%**	**105**

In addition, groups that consist of large numbers of terms are likely to be of much less utility than groups with a small number of terms. If the suffix heuristic creates a group that consists of, e.g., four hundred terms, the classificationist must manually break that group of terms down into more manageable sub-set that have more specific meanings. For this reason, a very long list of terms is unlikely to form the basis for a class or set of classes in a classification system.

Generally speaking, sets are most manageable when they contain no more than thirty items (Batty, 1989; Foskett, 1996; Taylor, 1995). In addition, groups that consist of only one term do not help organize the lexicon base into manageable sections. Therefore, suffix/meaning pairs are most useful when they create groups that range in size from two terms to thirty terms. Taking all terms into account, 75.69% of the suffixes and pseudo-suffixes used in the *annotations* lexicon base created groups of two to thirty terms (see Figure 1). However, in the same lexicon base, only 21.52% of the terms were placed into groups that consisted of two to thirty items (see Figure 2).

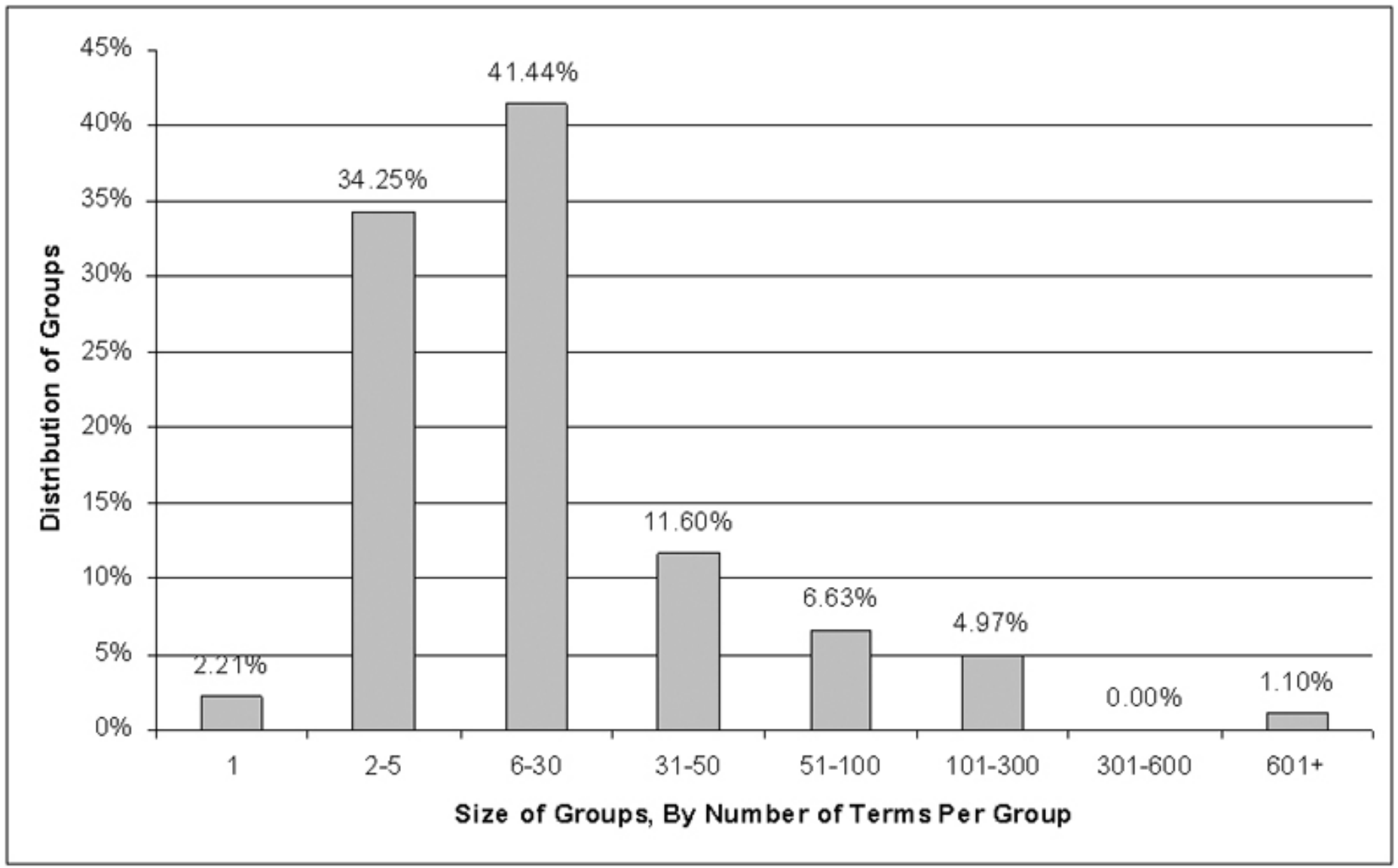
45%
40%
35%
30%
25%
20%
15%
10%
5%
0%
Distribution of Groups
41.44%
34.25%
11.60%
6.63%
4.97%
2.21%
0.00%
1.10%
1
2-5
6-30
31-50
51-100
101-300
301-600
601+
Size of Groups, By Number of Terms Per Group

Figure 1.

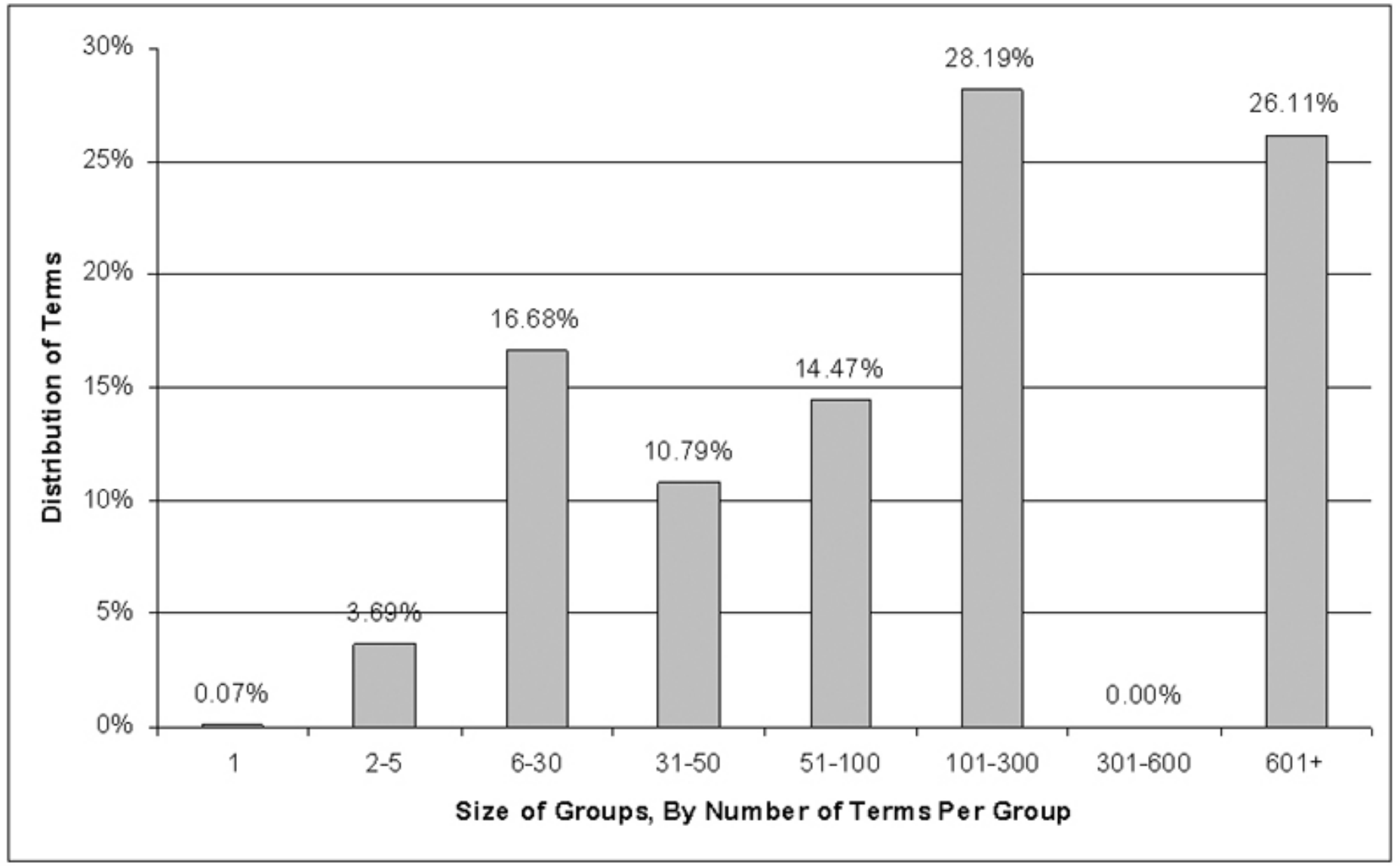

Figure 2.

Refinements to the Heuristic

Of the terms in the *annotations* lexicon base, 44.83% appeared only once in the corpus of EPA resources. It could be argued that these terms are less likely to be essential to a classification scheme that is based on the domain of the EPA than terms that appear multiple times. While filtering out terms that appear only once could potentially produce a lexicon base that is more relevant to the subject matter, it only slightly increases the effectiveness of the suffix heuristic. If only terms that appear multiple times in *annotations* are considered, the percentage of suffixes that create manageable groups (i.e., groups of two terms to thirty terms) decreases from 75.69% to 75.00%, while the percentage of terms that are organized into manageable groups rises from 21.52% to 27.34%. The overall percentage of terms that were assigned to classes rises from 42.66% to 44.59%.

Other approaches to maximizing the number of terms that appear in manageable groups might be more effective. One method is simply to filter out of every group all but the thirty terms that appear most frequently in the lexicon base. This does indeed result in groups with manageable cognitive loads without having a discernable effect of the precision of each group of terms, i.e., whether the terms have been placed into appropriate classes (see Table 6). However, it exaggerates the significance of terms with suffixes or pseudo-suffixes that are not shared by many other terms. For example, if this approach were used in the *annotations* lexicon base, the list of "actions" would not include the term *mining*, which is the thirty-sixth most commonly-occurring term in the set of "actions", appearing in the lexicon base forty-six times. In contrast, the list of "doctrines, theories, and sciences" includes the term *archaeology*, which ties for the twentieth most commonly-occurring term in that group, even though it appears in the lexicon base only once.

Table 6. The precision scores for four suffix meanings. The first column takes into account every term in the lexicon base, while the second column takes into account only the top thirty most frequently-occurring terms for each meaning

Suffix meaning	Validated classification (precision ratings)			
	Every term		Thirty most frequent terms	
Characteristics of actions (e.g., *cumulative*, *regulatory*)	98 of 119	82.35%	23 of 30	76.67%
Results of an action (e.g., *assessment*, *assurance*)	106 of 139	76.26%	25 of 30	83.33%
States, qualities, and conditions (e.g., *priority*, *toxicity*)	129 of 173	74.57%	19 of 30	63.33%
Companies and organizations (e.g., *amoco*, *witco*)	36 of 52	69.23%	21 of 30	70.00%

Discussion

It is difficult to predict which terms from a lexicon base will be most "essential" to a classificationist. A massive list of stopwords might be effective in filtering out non-essential terms, but would require the classificationist to manually identify in advance the terms that would be most likely to be inessential to classification systems in general. In addition, many terms could be non-essential in most domains, but highly essential in a few domains. Common terms also express additional meaning when term frequencies are taken into account. For example, although the terms *harmful*and *unlawful*are both so common as to be unlikely to suggest any particular strategy to a classificationist, the fact that *harmful*appears in the *annotations*lexicon base fifteen times while

unlawful appears only twice may provide important information as to the principles on which to most usefully base a classification system.

The Utility of Proper Nouns

In addition, methods for the identification of proper nouns might be helpful in determining the central concepts of a classification system, since proper nouns in general may be more relevant to a given domain than other terms. Many pseudo-suffixes appear to have consistently identified proper nouns. Pseudo-suffixes such as *-boro, -burgh,* and *-town* were useful in identifying metropolitan areas, while pseudo-suffixes such as *-ah, -ett,* and *-ald* identified proper nouns in general such as *Beulah, Bennett,* and *Gerald* to a precision of 94.92%. The pseudo-suffix -co identified proper nouns to a precision of 98.08% and identified the more specific concept of "corporations or organizations" to a precision of 69.23%.

When validating the various groups of proper nouns, each term was submitted as a query to both the Google search engine <http://www.google.com> and to Merriam-Webster Online. For example, when validating the set of "companies and organizations", each term in the group was assumed to be a valid member of the set if at least 50% of the first ten results in Google were for a company or organization (not necessarily the same company or organization) and if Merriam-Webster Online either did not have a definition of the term or its definition referred to a concept that was highly unlikely to have anything to do with the domain of the EPA. For example, the *annotations* lexicon base included the term *casco*. A search for "casco" in Google resulted in links to information about "Casco Products, Int'l", "Canada Starch Operating Company Inc.", "Casco Communications", and the "Casco programme", as well as the "Casco Bay estuary" in Maine, while Merriam-Webster defines *casco* as "a long almost rectangular barge or lighter sometimes with sails used in the Philippines". Because the majority of the resources returned by Google referred to *casco* as a company or as an organization, and because the definition for *casco* provided by Merriam-Webster Online appeared to have very little to do with the subject matter of the lexicon base, *casco* was deemed to be correctly classed.

However, this example actually highlights a limitation of this methodology, since according to the Muskie School of Public Service at the University of Southern Maine (2004), Casco Bay was "designated an 'estuary of national significance' and included in the U.S. Environmental Protection Agency's National Estuary Program". This suggests that the string *casco* in EPA literature is likely to be a

reference to the Casco Bay estuary. Even so, this approach to validation seems to be useful since the majority of the terms in the class "companies and organizations" that were deemed to be correctly classed most likely were. Also, "companies and organizations" is a sub-class of "proper nouns", so although *casco*probably does not belong in "companies and organizations", it was still placed in the correct general class, making it much easier for a classificationist to find the term than if it had not been placed in a class at all.

The Manual Process

Once a method has been established for the organization of relevant terms into appropriate groups, the classificationist must still interpret the results in order to create an effective classification scheme. It is likely that the classificationist will not simply create class labels that are taken directly from the terms, but will also create new classes based on shared characteristics of terms. For example, the classificatory heuristic may provide the classificationist with a list of action terms, a subset of which is related to chemical or molecular changes, such as:

aeration
chlorinated
combustion
composting
desorption
incineration
irradiated
oxidation
polychlorinated
radiation
tanning

In this case, the classificationist might conclude that "chemical processes" would be an appropriate class for the corpus of resources, even though the terms *chemical*and *processes*do not necessarily appear in the lexicon base. However, other groups of terms, though commonly used, might be less likely to form useful classes, such as the following "results of actions":

measurement
nonattainment
performance
pretreatment
reimbursement
requirement
sediment
settlement
statement
substance
treatment

Apart from the fact that *substance*and possibly *sediment*appear to have been classified incorrectly, this set of terms is unlikely to form the basis for a class or set of classes. Terms such as *measurement*and *requirement*appear to have more dissimilar features than similar features. Therefore, classes that are based on these terms would group resources that are unlikely to have much in common. Furthermore, terms such as *statement*and *performance*are likely to occur in a very large number of resources. They are relatively unlikely to represent the essence of a resource. This is not to say that large sets of terms are always useless to a classificationist. Lists of chemicals, for example, may to the trained eye suggest certain industries or processes. However, it is beyond the scope of the suffix heuristic to parse the specific language used by documents and other resources in an effort to determine which chemicals are associated with each other.

Conclusions

In this paper, we have described a study that explored the feasibility of constructing a faceted vocabulary using an existing hierarchical classification structure. We have also generalized the findings from that study to outline a hybrid, semi-automatic approach to faceted scheme creation that combines the strengths of the human with the strengths of the machine: the intelligence, context awareness and evaluative judgment that the human brings to the construction of high-quality faceted schemes with the speed of processing, unlimited memory and consistency in repetition of the machine.

We have shown how suffixes may be used in a hybrid approach to classification, in which automatic classification assists the human classificationist. Although suffixes alone cannot provide a useful initial classification for every term in a lexicon base, it provides a simple means by which to classify many terms, particularly highly technical terms and proper nouns. Further research will examine how the suffix heuristic can interact with other simple heuristics to provide the most effective approaches to classification.

References

Batty, D. (1989). Thesaurus construction and maintenance: a survival kit. *Database 12*(1), 13-20.

Bergen, B.K. (2004). The psychological reality of phonaesthemes. *Language 80*(2), 290-311.

Bolinger, D.L. (1965). *Shivaree and the phonestheme. In Forms of English: Accent, morpheme, order*(pp. 227-229). Cambridge: Harvard University Press.

Bybee, J.L. (1988). Morphology as lexical organization. In *Theoretical morphology: Approaches in modern linguistics*(pp. 119-142). San Diego: Academic Press, Inc.

Bybee, J.L. and Moder, C.L. (1983). Morphological classes as natural categories. *Language 59*(2), 251-270.

Casco Bay Estuary Project, Muskie School of Public Service, University of Southern Maine, *Casco Bay Estuary Project (CBEP) -*

A Cooperative Effort to Protect the Health and Integrity of Casco Bay. Retrieved January 29, 2005, from http://www.cascobay.usm.maine.edu/

Foskett, A.C. (1996). *The subject approach to information*, 5th ed. London: Library Association Publishing.

Hahn, U., Romacker, M., and Schulz, S. (2002). Creating knowledge repositories from biomedical reports: the medSynDiKATe text mining system. In: R. B. Altman, A. K. Dunker, L. Hunter, K. Lauderdale &T. E. Klein (Eds.), *Pacific Symposium on Biocomputing*.

Harman, D. (1991). How effective is suffixing? *Journal of the American Society for Information Science, 42*(1), 7-15.

Jacob, E. K., and Priss, U. (2001). Non-traditional indexing structures for the management of electronic resources. In *Advances in classification research, vol.10*. Information Today for the American Society for Information Science, Medford, NJ, 73-90.

Okada, M., Ando, K., Lee, S.S., Hayashi, Y., and Aoe, J. (2001). An efficient substring search method by using delayed keyword extraction. *Information Processing &Management, 37*, 741-761.

Priss, U., and Jacob, E.K. (1998). A graphical interface for faceted thesaurus design. In *Proceedings of the 9th ASIS SIG/CR Classification Research Workshop*(Pittsburgh, PA, October 25, 1998). American Society for Information Science, Silver Spring, MD, 107-118.

Ranganathan, S. R. *Library Classification: fundamentals and procedures with 1008 graded examples and exercises*. Madras Library Association, Madras, 1944.

Ranganathan, S. R. (1945). *Elements of library classification: based on lectures delivered at the University of Bombay in December 1944*. N.K. Publishing House, Poona, 1945.

Savoy, J. (1993). Stemming of French words based on grammatical categories. *Journal of the American Society for Information Science, 44*(1), 1-9.

Shisler, B.K. (1997). *The Dictionary of English Phonaesthemes*. Retrieved June 28, 2005 from http://www.geocities.com/SoHo/Studios/9783/phonpap2.html#glossary

Taylor, A.G. (1995). On the subject of subjects. *Journal of Academic Librarianship 21*(6), 484-491.

Webster's Third New International Dictionary, Unabridged. Merriam-Webster, 2002. Retrieved January 30, 2005 from http://unabridged.merriam-webster.com

Yang, K, Jacob, E., Loehrlein, A., Lee, S., Yu, N. (2004). *Organizing the Web: Semi-automatic construction of a faceted scheme*. IADIS International Conference WWW/Internet, 2004.