Vocabulary Alignment via Basic Level Concepts

Rebecca Green
College of Information Studies
University of Maryland

Final Report
2003 OCLC/ALISE Library and Information Science Research Grant Project

Suggested citation:
Green, Rebecca. 2006. "Vocabulary Alignment via Basic Level Concepts." OCLC/ALISE research grant report published electronically by OCLC Research. Available online at: http://www.oclc.org/research/grants/reports/green/rg2005.pdf.

OCLC/ALISE Library and Information Science Research Grant Program

Vocabulary Alignment via Basic Level Concepts
Final Report

Rebecca Green
College of Information Studies
University of Maryland

Abstract: Although significant overlap occurs in the semantic scope of existing thesauri and classification schemes, incompatibility between both the set of concepts they recognize and interrelationships among those concepts stands as a barrier to information exchange and integration, which, contrariwise, would be fostered by points of compatibility between systems. Preliminary research supported the premise that concepts at a hierarchical level called the basic level are more likely to be shared across classificatory systems than concepts at more general or more specific levels. The current research extends the previous research along two dimensions. First, this study has developed a set of linguistically-oriented heuristics for automatically identifying basic-level words in the WordNet noun hierarchy. Second, this study extends and generalizes previous findings by examining the degree of equivalence between closest correspondents of randomly selected concepts across classificatory systems in approximately a dozen subject domains. On the one hand, when examined individually, there are subject domains in which the premise is not supported; on the other hand, across the entire array of subject domains studied, the premise is strongly supported.

1. Introduction

As ever greater amounts of information and literature become available, systems for intellectual organization proliferate. Indeed, not only are there many systems, but there are many *kinds* of systems for organizing cognitive content. For starters, there are thousands of natural languages. On the more synthetic side, numerous bibliographic classification schemes exist, as well as Web search directories, thesauri, and ontologies.

Systems for organizing and expressing the organization of conceptual content—classificatory systems—differ with respect to many overlapping dimensions. For example, every natural language lexicalizes a different set of concepts and interrelates them in a unique manner. Some classification schemes have a general purpose and are of universal scope, while others address a limited and subject-specific domain. Some schemes are geared toward a broad, lay audience, while others target a narrow group of subject specialists. In the past, schemes were intended almost exclusively for human use; increasingly, schemes are designed for computer use and manipulation. Largely as a result of differences in purpose and focus, some schemes are more exhaustive in their coverage of a given subject than others, mostly through the inclusion of more specific concepts. The terms chosen to name a concept may differ across schemes. Schemes may also diverge with respect to how they relate concepts, both in terms of the inventory of relationships recognized and the degree to which syntagmatic relationships are enumerated/pre-coordinated in the scheme (vs. the degree to

which the expression of complex concepts can be generated as the need arises).

Such differences constitute a two-edged sword. As Visser et al. (1997) have noted, heterogeneity at the ontological level—although it often carries with it the advantage of greater efficiency through the tailoring of systems to their intended use—stands in the way of interoperability, a system desideratum that increasingly receives attention. For example, human searchers need the capability of a switching language to gain access to the ever expanding universe of knowledge (Mai 2004). Likewise, the ability to integrate classificatory systems in our networked environments will be required if the Semantic Web aspiration is ever to be achieved.

If it were simply an issue of the number of systems available or of surface differences, the process of integrating systems that organize intellectual content would be unwieldy, but not especially difficult. However, what we often face is some degree of incompatibility between and among our classificatory systems. For example, not only do human languages lexicalize different sets of concepts, but they may also structure (quasi-)equivalent concepts using different relational patterns (Riesthuis, 2001). As a consequence, even multilingual thesauri designed from the outset from the perspective of multiple languages may routinely include situations where corresponding terms are not truly equivalent (Hudon, 1997, 2001).

Among other advantages, the ability to identify concepts on which classificatory systems tend to agree provides a foundation for sharing or reusing information across systems and for realizing benefits from situations where information sources are classed in different ways. The exchange and integration of information across systems is fostered through conceptual equivalence across classificatory structures. For example, to the extent that two classificatory systems are compatible, a user (or information professional) familiar with one classification scheme could have his/her search request, as represented in one scheme, automatically translated into the other scheme, making literature classified only by the second scheme more readily accessible to the user or information professional. The ability to predict points of compatibility across classificatory systems would greatly facilitate such exchanges.

2. E Pluribus Unum

Two major approaches to combining classificatory systems have been used. One option is to build a new classificatory system by merging two or more extant systems. Equivalent concepts in the existing systems would first be identified and unified. After that would come the hard decisions about which parts of the existing systems to re-use. At one extreme, all non-equivalent concepts and all relationships could be retained. Under such a scenario, the number of concepts in the new system would multiply infelicitously, while both types and instances of relationships would also abound. The overall effect would tend toward the topsy-turvy. A more discriminating option would require selecting among the non-equivalent concepts and relationships to be retained. This scenario would be likely to result in a cleaner scheme, but at the price of losing some of the unique benefits associated with the original schemes.

The other option is to retain the nomenclature and structure of the existing systems, while also constructing between them a mapping (that is, a set of crosswalks) between their equivalent concepts. This option of aligning classificatory systems preserves the integrity of the input systems, thus retaining the unique benefits of each.

No matter which approach to classificatory system integration is taken, the single most important step in the process is the identification of equivalent concepts. A number of strategies have emerged

for doing this automatically (or at least for suggesting points of articulation for user reaction), as summarized in Euzenat et al. (2004).[1] Among the many kinds of data examined are the names of concepts, their definitions, semantic relationships that concept-related words enter into, attributes of the concepts, the number and identity of instances of the concept, and the taxonomic/structural context in which the concept resides.

The most significant barrier faced in efforts to integrate classificatory systems is the prevalence of "mismatches" between the concepts of one classificatory system and their closest correspondents in another scheme. Knowing a closest correspondent in another scheme may at times be desirable in a bibliographic classification, where a request for literature on one subject will sometimes be satisfied by literature on a closely related subject, either because of its content or because of its links to other parts of the literature. But in situations where search precision is more important than recall, or in systems that give direct access to information and that include a reasoning component, treating a closest correspondent as an exact equivalent may be misleading.

Several classifications of mismatches for ontologies exist (e.g., Visser et al. 1997; Chalupsky, 2000; Klein 2001).[2] Most classifications include mismatches based on terminological differences, where two systems include the same concept, but use different words for the concept (the case of synonymy) or where two systems use the same word, but it refers to different concepts in the two systems (the case of homonymy). While this sometimes makes the identification of equivalence a little tricky, it is not the kind of mismatch of concern here.

Our interest, rather, is in trying to identify equivalent concepts where mismatches operate on the conceptual level. Mismatches can be characterized in various ways:

- If two classificatory systems differ in scope, it is almost a foregone conclusion that concepts that exist within the scope of one system, but outside the scope of the other, will lack exact equivalents across the two systems. Thus, some concepts in one system may be altogether missing from the other.
- One system may cover a particular subject domain in greater specificity (or more generally, at different levels of granularity) than another. Then the closest correspondent that one system has for a concept in the other might be a more general concept. For example, the CAB Thesaurus has a descriptor, *Milk products;* the OECD descriptor that corresponds most closely is the more general *Dairy products*.
- The hierarchical context for concepts whose local characterizations make them appear equivalent may nonetheless reveal a degree of non-equivalence.
    - The concepts may be subordinated to different superordinate concepts. This would imply that the concepts refer to different kinds of things. For example, in the ERIC Thesaurus, as a narrower term of *Organization*, the descriptor *Classification* refers broadly to processes of grouping related phenomena into categories; in the OECD Macrothesaurus, the descriptor *Classification* is a narrower term of *Documentation* and a sibling of *Cataloguing, Filing*, and *Information processing*. The OECD *Classification* is not merely ERIC's general

---

[1]The urgency of the task and the maturity of its development are evident in the emergence of ontology alignment competitions. A prominent example is the Information Interpretation and Integration Conference (I^{3}CON; http://www.atl.external.lmco.com/projects/ontology/i3con.html).

[2]Naturally (!), they are not fully compatible.

Classification applied to the Documentation domain, but carries with it the history, principles, and practices of bibliographic classification.

- The concepts may have different subordinate concepts, which often means that they have different semantic scope. For example, both the Worldwide Political Science Thesaurus and the OECD Macrothesaurus have a descriptor *Social movements*. The scope note in the OECD Macrothesaurus for *Social movements* reads "collective efforts to transform some given [sic]"; narrower terms include *Labor movements, National liberation movements, Peasant movements*, and *Student movements.* A subtly different scope note for *Social movements* is given in the OECD Macrothesaurus: "Collective activities with spontaneous origins that acquire institutional structure over time; they generally emerge in the context of some form of perceived social injustice and center around specific causes and/or charismatic leaders." Narrower terms of *Social movements* in the OECD Macrothesaurus include *Environmental Movements, Feminism, Human Rights Movements, Nativistic Movements*, and *Protest Movements.* That the two concepts are not exactly equivalent is perhaps best seen in the fact that *Labor movements* is a narrower term of *Social movements* in the OECD Macrothesaurus, but in the Worldwide Political Science Thesaurus *Labor movements* is a narrower term of *Movements* and a sibling of *Social movements*.

• Few classificatory systems are so completely and perfectly faceted as to include reference to a concept from all conceivable perspectives. Consequently, the exact concept sought may not exist in a classificatory system, but a closely related concept of a different semantic type does. For example, while the European Education Thesaurus contains the descriptor *Nomads,* its closest correspondent in the Unesco IBE Education Thesaurus is *Nomadism*.

The identification of true equivalences plays an important role, no matter which approach to the combining of classificatory systems is undertaken. On the one hand, if a new scheme is built by merging existing schemes, it is crucial that the closest points of articulation between the schemes be identified at the outset as these will guide the remainder of the merger. The quality of the final product is constrained by the skeletal structure created by the first set of mappings. If these are incorrect, everything else will go at least minimally awry. On the other hand, if classificatory system integration takes the form of building a set of crosswalks between the existing schemes, then the identification of points of equivalence is even the more crucial, since it is precisely at these points that the crosswalks will be built.

3. Hierarchical Level and Conceptual Equivalence

Our efforts to increase the level of interoperability between systems would be facilitated and improved if we could predict where points of equivalence are likely to be found. Research findings from the social sciences suggest that hierarchical level is likely to play a significant role. For example, ethnobiological data show that folk classifications of flora are more likely to agree at a hierarchical level known as the basic level than they are at superordinate or subordinate levels (Berlin, 1992).

The basic level concept emerges from a context where specific things belong simultaneously to multiple classes, some of which are at different hierarchical levels. When we refer to a specific entity, we usually do so by using a label for a class the entity is a member of. For example, we seldom bother to refer to an automobile by its vehicle identification number (VIN), which would

enable us to name it uniquely, but, given a neutral context, we will typically refer to it as a *car*. But the class of cars is not the only class that could be used for referring to a specific automobile; both more specific (e.g., *sedan*) and more general (e.g., *vehicle*) classes/names are also available. Similarly, we are more likely to refer to *apples, shoes,* and *chair*s, which are basic level categories, than to their superordinates *(fruit, footwear, furniture)* or subordinates *(Granny Smith apples, sneakers, recliners).*

The hierarchical level we choose when we refer to entities turns out not to be random (Brown, 1958; Rosch et al., 1976). Indeed, a variety of processes by which we interact with objects in our world converge on this level (Lakoff, 1987, 46-47). Several of these processes are linguistic in nature. In addition to the reference process already mentioned, names for basic level categories tend to be shorter than the names for superordinate or subordinate categories. Words for basic level categories tend to enter the language earlier and to be learned by children earlier than words naming more general and more specific classes. Other processes that privilege the basic level of the hierarchy concern perception, function, and knowledge organization. Relative to perception, the basic level is the highest level in a hierarchy where humans can normally form a single mental image of a class and where they can identify the class by its average shape. Relative to function, the basic level tends to be the highest level in a hierarchy where humans interact with entities with a relatively constant motor program. Relative to knowledge organization, when people are asked to list all that they know about a certain category, the biggest increase in number of statements, over and above what one can say about the next-most-general superordinate class, comes at the basic level, thus implying that more information is stored at (is specific to) the basic level than at (to) any other hierarchical level.

Basic level categories are the highest level categories that enjoy a significant degree of homogeneity within a universe awash in heterogeneity. In a cluster analysis, basic level categories would correspond to clusters whose members are all more like each other than any of them is like a member of another class. The distinctiveness of this grouping increases the likelihood that people will perceive the set of entities to be members of a single class.

According to adherents of basic level theory, basic level categories are fundamental. Superordinate categories are formed by merging basic level categories, and subordinate categories are formed by dividing basic level categories. Thus, at the same time that basic level categories have a privileged status that argues for a high level of agreement on their essence and membership, the derivative nature of superordinate and subordinate categories suggests a lesser degree of agreement on the essence and membership of such classes. Likewise, both superordinate and subordinate categories rely on a degree of expertise not needed to apprehend basic level categories.

Human-built classificatory systems may be presumed to mirror human classificatory behaviors to a significant degree. Given this presupposition, we arrive at the premise that basic level concepts are more likely to be included in classificatory systems and thus that classificatory systems are more likely to have basic level concepts in common than would be the case for superordinate or subordinate concepts.

4. Methodology

The present study consisted of two phases. In the first phase a set of objective criteria for postulating which level within a given conceptual hierarchy is the basic level was developed; these criteria build on the characteristics of basic level categories outlined above. These criteria were used

to generate a set of basic level concepts for the nouns in WordNet,[3] a general lexical database for English words, which is structured hierarchically.

In the second phase, the hypothesis that concepts at the basic level are more likely to be shared across classificatory systems than concepts at more general or more specific levels was investigated across approximately a dozen subject domains. This exploration aimed at determining how widely the finding of greater universality for concepts at the basic level generalizes.

Phase 1

As previously noted, the characteristics of concepts from several perspectives—linguistic, perceptual, functional, cognitive—generally converge at the basic level. Some number of these characteristics are not readily identified apart from experimental results involving only a small number of specific concepts. What we need is a more general and objective approach for identifying the basic level of a hierarchy. Phase 1 of the research therefore consists of an analysis of the characteristics of pre-identified basic level concepts within WordNet to produce a general set of criteria for identifying such concepts. Specifically, phase 1 included three steps:

- identifying noun hierarchies in WordNet whose basic level had been previously identified;
- identifying data elements available within WordNet that correspond to known characteristics of basic level categories; and
- devising a scheme to score the levels of the WordNet hierarchies from step 1, based on the elements from step 2, that would assign the highest score within any specific hierarchy to the known basic level category.

Step 1. The seminal empirical study on basic level categories, Rosch et al. (1976), is based on six nonbiological superordinate categories, Musical instrument, Fruit, Tool, Clothing, Furniture, and Vehicle. Three basic level categories were identified for each superordinate category, and two subordinate categories were identified for each basic level category, thus yielding thirty-six (6 x 3 x 2) tri-level hierarchies, a relatively sparse amount of data to work with.

Figure 1 records the hierarchies used in Rosch et al. (1976); in this figure the words/phrases that occur in WordNet have been bolded. As can be seen, only one-third of the most specific terms are found in WordNet, which leaves only twelve complete tri-level hierarchies in WordNet for which the basic level has been previously established. Furthermore, five of the six hierarchies are quite similar in nature, being manufactured entities.

It is not intuitively clear that the characteristics shared by basic level terms in the seminal study will be predictive of basic level nouns generally. While Hoffmann and Ziessler (1983) posit that a basic level exists in every lexical hierarchy, we lack the empirical data to support this premise. Furthermore, we do not know whether the criteria that tend to converge for basic level categories in physical object hierarchies would also tend to converge for basic level categories in other kinds of hierarchies.

---

[3]http://www.cogsci.princeton.edu/~wn

Superordinate	Basic level	Subordinate
Musical instrument	**Guitar** **Piano** **Drum**	Folk guitar, classical guitar **Grand piano, upright piano** Kettle drum, **bass drum**
Fruit	**Apple** **Peach** **Grapes**	Delicious apple, Mac[k]intosh apple Freestone peach, Cling peach **Concord grapes**, Green seedless grapes
Tool	**Hammer** **Saw** **Screwdriver**	**Ball-peen hammer, claw hammer** Hack hand saw, cross-cutting hand saw **Phillips screwdriver,** regular screwdriver
Clothing	**Pants** **Socks** **Shirt**	**Levis,** double knit pants Knee socks, ankle socks **Dress shirt,** knit shirt
Furniture	**Table** **Lamp** **Chair**	**Kitchen table,** dining room table **Floor lamp,** desk lamp Kitchen chair, living room chair
Vehicle	**Car** **Bus** **Truck**	**Sports car,** four door sedan car City bus, cross country bus Pick up truck, tractor-trailer truck

Figure 1. Noun hierarchies with established basic levels
(**bolded** terms exist in WordNet)

Step 2. The information available within WordNet that touches on characteristics associated with basic level categories include: length and structure of lexical units (i.e., is the lexical unit a phrase, a compound word, or a simple word?), height within the WordNet tree structure, the number of links to parts, and frequency of usage. Other information available in WordNet with possible relevance to the identification of basic level categories includes: the total number of links to other concepts[4] and the number of children/subordinate concepts, both immediate (one level down) and overall (no matter how many levels down; this quantity is referred to here as "tree size").

Step 3. Analysis of these characteristics for the 12 complete hierarchies in WordNet data led to the following generalizations:

- If a word is longer than 15 characters, it is unlikely that it names a basic level category.
- If a lexical unit is a phrase (e.g., *dress shirt, sports car*), it is unlikely that it names a basic level category. Further, if a lexical unit includes more than two words, it may be safely assumed that it does not name a basic level category.
- If the name of a concept is included within the name of a more specific concept, it probably names the basic level category.

[4]Carol Bean has suggested (personal communication) that basic level concepts participate in more relationships than do other concepts.

- If, according to SEMCOR[5] data, the frequency of occurrence for one concept is greater than the frequency of occurrence for both its next broadest and its next most specific concept, the concept probably names a basic level category.
- If a concept has more than four levels beneath it, it probably names a superordinate level category.
- If a concept has more than one part listed, it probably names a basic level category.
- If a concept has no children, it is unlikely to be basic level, but is much more likely to be a subordinate concept.

Application of these findings to the WordNet noun network started with identifying every leaf node in the network (i.e., synsets that lack more specific synsets) and the hierarchical structure above it. Each set of noun synsets constituting a complete hierarchy was analyzed to identify the most likely basic level synset, using a modified majority voting scheme (Van Halteren, Zavrel, & Daelemans, 1998). This analysis involved three components. The first component considered two characteristics—complexity of the lexical unit and height of the node—that disqualify a synset. If the lexical unit associated with a synset[6] consists of more than two units (where units are divided by spaces or hyphens), the corresponding synset was disqualified. On an absolute level, if a synset has more than five levels below it, it was disqualified; on a relative level, if a synset is in the top half of its hierarchy, it was disqualified.[7] The second component involved three characteristics that favor a synset's being at the basic level, namely, if the lexical unit associated with the synset is the shortest of all such lexical units for the hierarchical chain, if the number of synsets listed as being parts of the synset is the highest for all synsets in the chain, or if the ratio of frequency of occurrence of the synset to its treesize is greatest for all synsets in the chain. All of these characteristics cast a vote for a synset. The third component of the identification algorithm involved characteristics that disfavor a synset's being at the basic level, namely, if the synset has fewer children than does any other synset in its chain or if the synset has fewer relationships with other synsets that does any other synset in its chain. Both of these characteristics cast a vote against a synset. After this voting, the synset receiving the most votes was hypothesized to be at the basic level.

This analysis was implemented for 59,692 hierarchies and identified 7,168 basic level synsets. Note that the number of hierarchies analyzed is much greater than the number of basic level synsets identified. This results from the basis for counting hierarchies and basic level synsets: On the one hand, the hierarchical chain above every leaf node in the WordNet noun network is treated separately in the enumeration of hierarchies, as each hierarchy differs from every other hierarchy, at least at the

---

[5]SEMCOR is a semantic concordance that includes 186 short texts from the Brown Corpus in which all open-classed words have been tagged with their WordNet senses.

[6]In accordance with its name, a synset ("synonym set") typically has multiple lexical units associated with it. For purposes of identifying the most likely basic level synset in a hierarchical chain, the lexical unit that is most commonly associated with a synset in SEMCOR has been used. If the synset does not occur in SEMCOR, then the first lexical unit listed for the synset is used.

[7]It is not uncommon for very general concepts, for example, the one named by *thing*, to have many of the linguistic characteristics associated with basic level categories. Such synsets must therefore be filtered out by a height criterion.

leaf node level. On the other hand, if a given synset is identified as a basic level synset for more than one hierarchy, it still counts only once. That there are so many more hierarchies than basic level synsets indicates how commonly a given synset has been identified as a basic level synset for more than one hierarchy (on average basic level synsets operate over 8.3 hierarchies).

This is as it should be. For example, given that *guitar, apple, hammer, shirt, chair,* and *truck* are basic level terms, we would expect all nodes for specific kinds of guitars to have the same basic level synset, i.e., the one that includes *guitar* (which occurs in only 1 synset), all nodes for specific kinds of apples to have the same basic level synset, and so forth. On a more general level, if we take any two sibling leaf nodes—that is, two nodes that have the same parent node (and indeed the same "ancestry" all the way up their respective hierarchical chains), but no children nodes—we would expect them to have the same basic level node. Indeed, once we have identified a basic level node for any one hierarchy, we would expect that node to represent the basic level in all hierarchies for all synsets subordinate to it. Furthermore, we would also expect (all of) its siblings to be identified as basic level synsets.

The method used for identifying basic level synsets supports such expectations in part. A synset that is disqualified for basic level consideration in one hierarchy on the grounds that its primary lexical unit consists of more than two units will be disqualified on the same grounds in all hierarchical chains that it participates in.

However, all the other criteria that go into the determination of basic level synsets take hierarchical context into account. Where a given synset that is identified in any one context as a basic level synset has an atypical value on those criteria, e.g., if the primary lexical unit for the synset is much shorter than the average word or the synset has very many synsets related to it as parts, it is more likely to be identified as a basic level synset in other hierarchical contexts. But where a synset has more typical values on these criteria and wins basic level designation in one hierarchy by a close vote, it is less likely to win the vote in all hierarchies that it participates in.

The set of basic level categories generated in phase 1 has been used in phase 2 of the research.

Phase 2

Preliminary research (Green, Bean, & Hudon, 2002) addressed whether a concept's hierarchical level affects the likelihood of a concept's inclusion in multiple classificatory schemes. Specifically, the study investigated the validity of the hypothesis that concepts at the *basic level* are more universal—that is, more likely to co-occur across classificatory systems—than either more general or more specific concepts.[8] The hypothesis of the greater universality of basic level concepts was examined and validated (at $\alpha = .001$) in three contexts: across languages (specifically, between corresponding terms of the two languages of a bilingual thesaurus in the social sciences), across vocabularies (e.g., between terms of different medical vocabularies mapping to the same concept within the Unified Medical Language System® [UMLS]; see related work by Bodenreider & Bean, 2001), and across ontologies (e.g., between most nearly equivalent nodes of ThoughtTreasure and

[8]The issue is not whether the same *terms* co-occur across classification schemes, nor whether concepts are expressed by the same term, but whether the same *concepts* co-occur.

WordNet; see related work by Hovy, 2002).[9]

The study reported here aimed at determining whether this finding of greater universality for concepts at the basic level generalizes to other contexts, by examining approximately a dozen subject domains for which two or more online thesauri were available. Subject domains analyzed include agriculture, education, engineering, the environment, graphic materials, health, information science, legislation, political science, population science, and water sciences.

Conduct of the study included the following steps:

- Two or more online thesauri were identified for each subject domain. (Specific thesauri used are given in the Appendix.)
- Ten descriptors were randomly selected from each thesaurus.
- Each randomly selected descriptor was expanded using the relational structure of the thesaurus to identify an entire conceptual hierarchy in which the descriptor occurs.
- For each concept in these hierarchical expansions, the closest correspondent in the other thesauri for the subject domain was identified.
- Each pair of descriptors—with one descriptor drawn from the hierarchical expansion of a randomly selected concept from one thesaurus and the other descriptor being the closest correspondent in another thesaurus—was investigated to determine if the two concepts were exactly equivalent or not.
- Basic level concepts for each conceptual hierarchy were determined. Where possible, these were taken from the results of phase 1. Where a basic level concept could not be established from those results, the basic level was chosen on the basis of known characteristics of basic level categories.[10]

---

[9]The hypothesis was specifically confirmed as follows:

1. Across the three case studies combined, terms for basic level concepts were more likely to have equivalents in other knowledge organization and representation tools than were terms for non-basic level concepts (subordinate and superordinate concepts combined) ($\chi^2 = 25.24047$, df = 1, $\alpha = .001$).

2. Across the three case studies combined, terms for basic level concepts were more likely to have equivalents in other knowledge organization and representation tools than were terms for subordinate concepts ($\chi^2 = 21.45006$, df = 1, $\alpha = .001$).

3. Across the three case studies combined, terms for basic level concepts were more likely to have equivalents in other knowledge organization and representation tools than were terms for superordinate concepts ($\chi^2 = 22.41596$, df = 1, $\alpha = .001$).

[10]There were circumstances under which no helpful phase 1 results could be applied to the identification of the basic level within the conceptual hierarchies. First, there were many cases where the hierarchies established in a particular thesaurus had no equivalent in the hierarchical structure of WordNet. WordNet is, after all, 'just' another classificatory system, and, like other classificatory systems, is unique and likely to be found incompatible with other classificatory systems. Second, thesauri often introduce descriptors that are not recognized as standard phrases of English. Thus, many thesaurus descriptors do not occur in WordNet. Consequently, zero, one, or more descriptors from a conceptual hierarchy might match basic level synsets as selected in phase 1.

- The results of the equivalence analysis were subjected to statistical analysis (using χ^2/goodness-of-fit tests), by specific subject domain, by general subject domain (e.g., humanities, social sciences, natural sciences), and overall.

Table 1 presents the results of phase 2. For each subject domain is shown the number (and percentage for the subject domain) of: descriptor pairs at the basic level that are exactly equivalent, descriptor pairs at the basic level that are not exactly equivalent, descriptor pairs not at the basic level that are exactly equivalent, and descriptor pairs not at the basic level that are not exactly equivalent. Counts and percentages (in parentheses) are summarized within general subject domains. Additionally, as the number of thesauri and the size of hierarchical expansions varied across subject domains, percentages for general subject domains were also computed with each specific subject domain's results being weighted equally (these percentages are bolded in the table).

The final column of Table 1 shows the probability that the observed values differ from the expected values. In two cases (information science, population science) there are insufficient data to compute a reliable χ^2 score. In several other cases the analysis shows either that it is highly unlikely (graphic arts, political science) or just unlikely (legislation, engineering, the environment) that the observed values differ from expected values. In such cases, no distinction between concepts at the basic level and concepts not at the basic level is supported by the analysis, and the premise under investigation is not validated. In other cases (education, agriculture, health, and water sciences), it is highly probable that the observed values differ from expected values. In all of these cases, the closest correspondents for basic level concepts are more likely to be exact equivalents than not to be exact equivalents; in contrast, the closest correspondents for concepts not at the basic level are more likely not to be exact equivalents than to be exact equivalents. In these cases, the χ^2 test supports the premise that basic level concepts in one thesaurus are significantly more likely (at α = .001) to have exact equivalents in another thesaurus than are concepts not at the basic level,

The data are summed over general subject domains as well as being summed over all the data. Except for the case of the humanities, in which only one subject domain was found with two or more online thesauri, the hypothesis of greater universality for basic level concepts is supported at all levels of generalization (at α = .01 for the social sciences and at α = .001 for the natural sciences and technology and overall).

5. Conclusion and Future Research

The practical benefit of this study is the direction it provides for building crosswalks between classificatory systems. If basic level concepts are significantly more likely to have exact equivalents across systems than concepts not at the basic level, then developing mappings across classificatory systems should emphasize basic level concepts, as these will be the cleanest mappings. Since, among their other characteristics, basic level concepts carry more information than other concepts, mappings involving these descriptors will also provide more benefit from an information-seeking perspective.

There are a number of directions in which this line of inquiry should be further expanded:

1. Most of the work on basic level concepts has focused on concrete entities, an emphasis reflected in restricting phase 1 to nouns within WordNet. Further investigation is warranted on the degree to which the basic level concept also applies to abstract entities, states, processes, attributes, and relationships.

2. As the criteria for identifying basic level synsets used in phase 1 is based on such a small number of hierarchies and as they tend toward homogeneity, human assessments of the phase 1 results is called for.

3. The study should be repeated on thesauri for other subject domains in the humanities, to see if the anomalous result for the graphic arts generalizes across the humanities or is itself anomalous.

4. This study limits its investigation to (monolingual) thesauri, while the preliminary investigation also examined the universality-of-basic-level-concepts hypothesis across languages (in the context of a multilingual thesaurus) and across ontologies. Both of these contexts warrant further research.

5. The implications of the universality-of-basic-level-concepts hypothesis for complex or precoordinate classes has received as yet essentially no attention. Many documents have been classified using, for example, both the Library of Congress and Dewey Decimal classification schemes, where many of the classes reflect complex concepts. Such concepts do not fit neatly into the basic level concept mold, although the constituent concepts that make up the complex concept do. Extension of this research into the realm of these standard bibliographic subject access tools is especially called for.

References

Berlin, Brent. (1992). *Ethnobiological classification: Principles of categorization of plants and animals in traditional societies.* Princeton: Princeton University Press.

Bodenreider, Olivier & Bean, Carol A. (2001). Relationships among knowledge structures: Vocabulary integration within a subject domain. In C. A. Bean & R. Green (Eds.), *Relationships in the organization of knowledge*, 81-98. Dordrecht: Kluwer Academic Publishers.

Brown, Roger. (1958). How shall a thing be called? *Psychological Review* 65: 14-21.

Chalupsky, Hans. (2000). OntoMorph: A translation system for symbolic knowledge. *Principles of knowledge representation and reasoning: Proceedings of the Seventh International Conference on Knowledge Representation and Reasoning.*

Green, Rebecca, Bean, Carol A., & Hudon, Michèle. (2002). Universality and basic level concepts. In María J. López-Huertas (Ed.), *Challenges in knowledge representation and organization for the 21st Century : Integration of knowledge across boundaries : Proceedings of the Seventh International ISKO Conference, 10-13 July 2002, Granada, Spain*, 311-317. Würzburg: Ergon.

Euzenat J., Le Bach T., Barrasa J., Bouquet P., De Bo J., Dieng R., Ehrig M., Hauswirth M., Jarrar M., Lara R., Maynard D., Napoli A., Stamou G., Stuckenschmidt H., Shvaiko P., Tessaris S., Van Acker S. & Zaihrayeu I. (2004). State of the art on ontology alignment. Available: <http://www.starlab.vub.ac.be/publications/kweb-223.pdf>[21 January 2005].

Hoffmann, J. & C. Ziessler. (1983). Objektidentifikation in künstliche Begriffshierarchien. *Zeitschrifte für Psychologie*, 191/4: 135-167.

Hovy, Eduard. (2002). Comparing sets of semantic relations in ontologies. In R. Green, C. A. Bean, & S. H. Myaeng (Eds.), *The semantics of relationships: An interdisciplinary perspective*, 91-110. Dordrecht: Kluwer Academic Publishers.

Hudon, Michèle. (1997). Multilingual thesaurus construction: Integrating the views of different cultures in one gateway to knowledge and concepts. *Knowledge Organization* 24/2: 84-91.

Hudon, Michèle. (2001). Relationships in multilingual thesauri. In C. A. Bean & R. Green (Eds.), *Relationships in the organization of knowledge*, 67-80. Dordrecht: Kluwer Academic Publishers.

Klein, Michel. (2001). Combining and relating ontologies: An analysis of problems and solutions. *Workshop on Ontologies and Information Sharing, IJCAI '01*.

Lakoff, George. (1987). *Women, fire, and dangerous things: What categories reveal about the mind.* Chicago: University of Chicago Press.

Mai, Jens-Erik, (2004). The future of general classification. *Cataloging and Classification Quarterly,* 37 (1/2): 3-12.

Riesthuis, Gerhard J. A. (2001). Information languages and multilingual subject access. *Subject retrieval in a networked environment: Papers presented at an IFLA satellite meeting, Dublin, Ohio, USA, 14-16 August 2001*..

Rosch, Eleanor, Mervis, Carolyn, Gray, Wayne, Johnson, David, & Boyes-Braem, Penny. (1976). Basic objects in natural categories. *Cognitive Psychology* 8: 382-439.

Van Halteren, Hans, Zavrel, Jakub, & Daelemans, Walter. (1998). Improving data-driven wordclass tagging by system combination. *Proceedings of the 36th Annual Meeting of the Association for Computational Linguistics and the 17th International Conference on Computational Linguistics*, 491-497.

Visser, P.R.S., D.M. Jones, T.J.M. Bench-Capon and M.J.R. Shave (1997). An Analysis of Ontology Mismatches; Heterogeneity versus Interoperability, Working notes of the AAAI 1997 Spring Symposium on Ontological Engineering, Stanford University, California , USA, pp.164-172.

Discipline	Basic level Exact equivalence	Basic level Other	Not basic level Exact equivalence	Not basic level Other	Probability
Humanities	2 (.01)	12 (.06)	22 (.12)	154 (.81)	0.269
Graphic arts	2 (.01) **(.01)**	12 (.06) **(.06)**	22 (.12) **(.12)**	154 (.81) **(.81)**	0.269 **0.269**
Social sciences	83 (.16) **(.18)**	53 (.10) **(.10)**	167 (.33) **(.36)**	210 (.41) **(.36)**	0.999 **0.992**
Education	48 (.16)	33 (.11)	92 (.30)	132 (.43)	0.995
Information science	6 (.08)	3 (.04)	22 (.31)	40 (.56)	*
Legislation	11 (.22)	8 (.16)	16 (.31)	16 (.31)	0.415
Political science	9 (.21)	7 (.16)	15 (.35)	12 (.28)	0.035
Population science	9 (.21)	2 (.05)	22 (.51)	10 (.23)	*
Natural sciences and technology	79 (.13) **(.14)**	93 (.16) **(.15)**	115 (.19) **(.20)**	311 (.52) **(.52)**	0.999 **0.999**
Agriculture	16 (.20)	10 (.13)	18 (.23)	35 (.44)	0.98
Engineering	7 (.09)	13 (.17)	15 (.19)	43 (.55)	0.564
Environment	15 (.10)	35 (.24)	34 (.23)	62 (.42)	0.489
Health	20 (.13)	14 (.09)	18 (.12)	104 (.67)	0.999
Water sciences	21 (.16)	13 (.10)	29 (.22)	67 (.52)	0.974
Total	164 (.13) **(.14)**	158 **(.12)**	304 **(.26)**	676 **(.47)**	0.999 **0.999**

Table 1. Closest correspondent concepts across thesauri: Raw counts (Percentages of raw counts) **(Equally weighted percentages)**

Appendix: Thesauri used in the study, grouped by discipline
Note: OECD = Organisation for Economic Co-operation and Development

- Humanities
  - Graphic arts
    - ARTbibliographies Modern Thesaurus (ABM)
    - Iconclass
    - Thesaurus for Graphic Materials I: Subject Terms (TGM I)
- Social sciences
  - Education
    - OECD, Macrothesaurus 6 (education, training)
    - Unesco IBE [International Bureau of Education] Education Thesaurus
    - European Education Thesaurus (EET)
    - ERIC [Educational Resources Information Center] Thesaurus
  - Information science
    - ASIS Thesaurus of Information Science
    - INFODATA Thesaurus
  - Legislation
    - Legislative Indexing Vocabulary (LIV)
    - Global Legal Information Network (GLIN)
  - Political science
    - OECD, Macrothesaurus 4.04 (politics)
    - Worldwide Political Science Abstracts Thesaurus
  - Population science
    - OECD, Macrothesaurus 14 (demography, population)
    - POPIN Thesaurus (Population Multilingual Thesaurus)
- Natural sciences and technology
  - Agriculture
    - CAB [Center for Agriculture and Biosciences] Thesaurus
    - OECD, Macrothesaurus 7 (agriculture)
  - Engineering
    - Copper Data Center Thesaurus of Terms
    - Metallurgical Thesaurus / Engineered Materials Thesaurus
  - Environment
    - Pollution Thesaurus
    - EnVoc Multilingual Thesaurus of Terms
    - General Multilinguial Environmental Thesaurus (GEMET)
  - Health
    - Health and Ageing Thesaurus
    - Public Health Information Thesaurus
    - Life Sciences Thesaurus
  - Water sciences
    - ASFA [Aquatic Sciences and Fisheries] Thesaurus
    - Water Resources Thesaurus
    - OECD. Macrothesaurus 17.5 (hydrology, water)