Hierarchical Preferences in a Broad-Coverage Lexical Taxonomy

Massimiliano Ciaramita (M.Ciaramita@ISTC.CNR.IT)

Institute for Cognitive Science and Technology

National Research Council, Via Nomentana 56, Roma, 00161 Italy

Steven Sloman (Steven_Sloman@Brown.Edu)

Department of Cognitive and Linguistic Sciences; Box 1978

Mark Johnson (Mark_Johnson@Brown.Edu)

Department of Cognitive and Linguistic Sciences; Box 1978

Eli Upfal (Eli_Upfal@Brown.Edu)

Department of Computer Science; Box 1910

Brown University

Providence, RI 02912 USA

Abstract

We investigate the problem of finding informative superordinates in a broad-coverage taxonomy of nominal concepts. We present results from a study which shows that speakers often exhibit strong preferences on what superordinate is more informative, together with a solid bias for specific classes. We then define the task of identifying the properties that characterize such concepts in the taxonomy as a ranking problem. We identify several such properties which are related to properties of basic concepts. While these properties provide accurate sources of information for identifying the most useful superordinate, their interaction remains obscure.

Introduction

Lexical meaning is often summarized as category membership: a "convertible" is a "car", a "trombonist" is a "musician", "irritation" is a "feeling", etc. Within taxonomic organizations a nominal concept belongs to all its superordinates; e.g., "rattler" belongs to "viper", "snake", "reptile", "vertebrate", "animal", "organism". However, certain superordinates such as "snake" tend to be more important than others. Arguably the relevance is context-dependent; if a "rattler" was found on an asteroid one would probably wonder how an "animal" managed to get there, more than how a "snake" did. Nonetheless, in hierarchical categorization schemes people tend to prefer useful and efficient classes, in terms of information content, i.e., concepts where the trade-off between size of the category and similarity of its members is optimal (Gluck & Corter, 1985). Such concepts are called *basic* (Brown, 1958; Rosch et al. 1976) and are well-studied in humans (Murphy, 2002).

In this paper we address two aspects relevant to basic categories. In the first part we investigate for what fraction of nouns in naturally occurring language there is a corresponding "favorite" superordinate. The goal is to estimate, at least as a first crude approximation, the extent of this phenomenon. We examine this issue with a naming task in which we collected data from English speakers concerning their preferences about different superordinate levels. We used a broad-coverage nominal
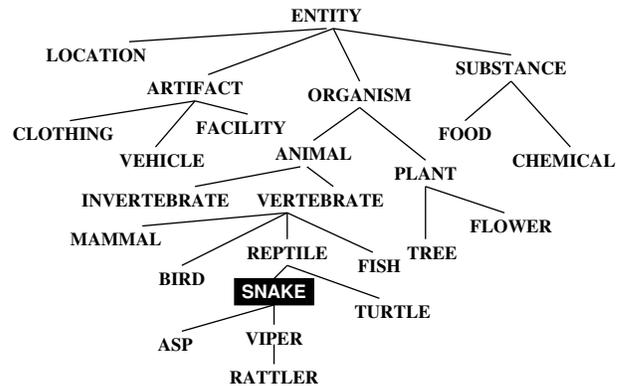

Figure 1. A simplified portion of the Wordnet taxonomy of nominal concepts above the noun "rattler".

taxonomy, and several hundred nouns found in a corpus. We found that for a large fraction of nouns, more than 84%, there is a superordinate which is significantly more informative than the others; participants were mostly unsure about nouns referring to abstract concepts such as relations and states. In the second part of the paper we use the outcome of our study as a gold standard and we investigate the properties that characterize informative superordinates. We frame the problem of choosing the "best" superordinate for a noun as a ranking task. We investigate word-specific properties – those that can be extracted from a word's orthography, from corpus data, and from hierarchical knowledge relating words – to try to characterize informative superordinates.

We found that word length provides the weakest predictor, while entropy and frequency, and especially functions that measure the association between the noun and the superordinate such as mutual information, are more accurate. Finally, most surprisingly, we found that the best predictor is the concept *specificity*; i.e., people have a strong preference for specific concepts. This finding suggests a simple hypothesis: that the data might be explained by a model that combines specificity and other

properties of basic categories. However, we show that the interaction of the different information sources is complex and the choice of the superordinate might depend on subtle semantic and cultural factors.

Basic Categories and Information

The superordinates of a noun such as "rattler" (viper, snake, reptile, vertebrate, animal, and organism) are not equally useful. The choice of one extreme or the other has both advantages and shortcomings. Broad classes such as "organism" are easy to discriminate, e.g., from "artifacts", but are not very informative because they have very dissimilar subordinates like "animal" and "plant" (cf. Figure 1). In contrast, specific classes such as "viper" contain very similar subordinates but are hard to discriminate; e.g., it is harder to distinguish a "viper" from an "asp" than an "animal" from an "artifact". It seems intuitive that there should be an intermediate level where an optimal balance between discriminative power and similarity of the subordinates is achieved.

A level with such properties is called a *basic level* in human categorization (Brown, 1958; Rosch et al., 1976). Basic categories are intermediate-level classes typically expressed by phonologically simple (short) and frequent words; e.g., "chair" "tree" and "snake". The basic-level is of great importance to human tasks like naming (Rosch et al. 1976), forming mental images (Tversky & Hemenway, 1984), and reasoning about objects' functions and other attributes (Sloman & Ahn, 1999). Basic categories are most useful because they are accurate at predicting distinctive attributes of their members and, at the same time, possess high category resemblance. The basic level provides the most natural contrast between categories and is the most useful for induction. These properties can be expressed with information theoretical, i.e., entropy-based, measures (Gluck & Corter, 1985; Corter & Gluck, 1992) which quantify a category's power to reduce uncertainty about the features of its members – or the *category utility*. Gluck and Corter (1985; Corter & Gluck, 1992) show that this model is consistent with people's performances in category learning experiments. Informativeness does seem to be a partial explanation for the basic-level. The basic-level refers to a hierarchical level that is picked out by a variety of different language-related tasks, in this sense it represents an empirical phenomenon. Informativeness serves as an explanation of the basic-level, almost all theories of the basic-level appeal to informativeness which provides a natural and useful explanatory notion.

The information-theoretical interpretation of the basic level can be formalized precisely and implemented in computational models (cf. (Fisher, 1988) for an implementation of the category utility model). Among other applications, this is relevant to natural language processing systems; e.g., in *language generation*. Basic-level terms are crucial for generating utterances that, not only sound natural, but also obey Gricean maxims of discourse (cf. (Dale & Reiter, 1995)); e.g., the choice of "look at the Canis familiaris" vs. "look at the dog" or "look at the pitbull" has different pragmatic conse-

artifact	34	substance	9	event	4
person	32	attribute	9	possession	4
plant	24	cognition	9	time	3
animal	23	food	8	process	3
act	20	group	8	phenomenon	2
communication	17	body	6	feeling	2
state	11	object	5	relation	2
location	10	quantity	4	shape/motive	1

Table 1. Number of test words per supersense category.

quences. Other applications in principle include all those that involve semantic classification tasks such as lexical disambiguation or named-entity recognition, which currently focus on repertories of, respectively, excessively narrow and excessively broad classes.

Speakers' Hierarchical Preferences

We ran a study in which we collected data about people's preferences within hierarchical classification schemes. The goal was to estimate the fraction of nouns with a consistently preferred superordinate and provide data to determine which nouns showed such consistency and what the properties are of preferred superordinates.

Description of the Data

We found all nouns in the Brown corpus (Francis & Kučera, 1982), a balanced corpus of about a million words, that have an entry in Wordnet (Fellbaum, 1998), a large ontology which contains 115,000 nouns which belong to about 80,000 concepts called *synsets* hierarchically organized according to the "is a kind of" relation (e.g., "pen" is-a "writing implement"). We searched for single words and compounds such as "real estate" or "psychological warfare"[1]. We found 12,218 such nouns, roughly 80% are common nouns. We associated each noun with its most common synset according to Wordnet; e.g., the first synset of rattler is "snake" while the second is "freight-train".[2] Each synset in Wordnet is also tagged with a label corresponding to the categories used by lexicographers to organize the development of the database. There are 26 such categories which we call *supersenses* (cf. (Ciaramita et al., 2003)); e.g., "person", "animal", "plant", "artifact", "location", "feeling" etc. To select a set of words representative of the overall composition of the Wordnet lexicon we binned the word-synset pairs according to their *supersense* label. For each supersense ss_i we randomly selected a number of word senses equal to $\lceil N \frac{|ss_i|}{\sum_j |ss_j|} \rceil$, where N is the total number of test words and $|ss_i|$ is the number of synsets with supersense label ss_i. For example, 14% of the synsets are "artifacts" while 0.05% are "motives", respectively the larger and smaller supersense.[3]

We decided to test 250 common nouns, the number of words that we estimated could be tagged in an hour.

---

[1] For this purpose we used the functions "getindex()" and "morphstr()" from the Wordnet library "wn.h".

[2] This information was compiled by the Wordnet lexicographers based mainly on estimations from the Brown corpus.

[3] We only chose synsets that have no hyponyms, i.e., leaf-nodes, 62870 of the synsets in Wordnet (79%).
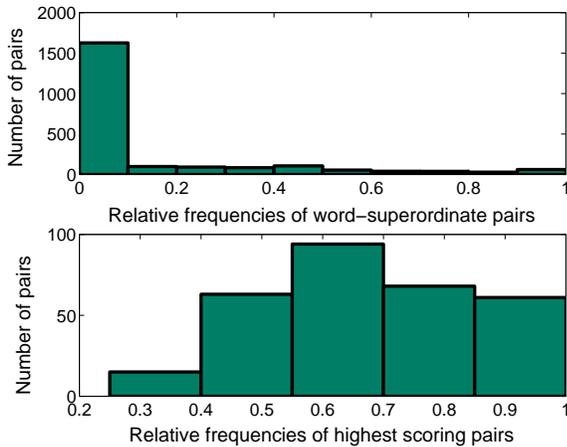
Figure 2. Histograms of relative frequencies from the speakers data for all 2,228 test word-superordinate pairs (above) and for the 301 superordinates that participants chose the most (below).

Table 1 summarizes the number of words per each supersense category generated for this value of N. In addition to the common nouns we selected a set of 51 proper nouns, which are also in Wordnet, for the supersense categories "person" "group" and "location". These three categories alone cover more than 80% of all proper nouns. These 301 common and proper nouns represented our sample of nouns appearing in naturally occurring language. In a pilot study we tested 10 participants on a different set of 213 nouns. We learned that some of the participants didn't know several of the words. During the development of the final data set we excluded obscure nouns and nouns for which the first sense according to Wordnet was clearly not the most frequent sense in current use; e.g., "hot dog" as "exhibitionist" instead of "sandwich".

Description of the Test

Participants were presented one test word at a time on a computer monitor in randomized order. Together they saw a numbered list of candidates, also randomized, the superordinates of the test word. Each superordinate was expressed by one noun, the first in the list for that synset;[4] e.g., one test noun is "turmoil" and the list of candidates is: "state", "disorder" and "disturbance". Participants were asked to choose which term they would use to answer the question "What kind of thing is this?".[5] The total number of candidates, i.e., test word-superordinate pairs was 2228, or 7.4 candidates per word on average. Participants were explicitly told that there were no wrong answers – all candidates provided a correct explanation – and they could imagine a situation in which somebody, who didn't know the meaning of the word, asked the question to which they had to answer using only one of the words in the list.

---

[4]This ordering was also compiled by the lexicographers.

[5]Or "Who is this?" if the test word was a person's name.

Results

We tested 12 Brown graduate and undergraduate students on all 301 nouns and computed the relative frequency of the preferences obtained by each superordinate; e.g., for "turmoil", "state" was chosen twice, P(state)=0.16, "disorder" 4 times, P(disorder)=0.33, and "disturbance" 6 times P(disturbance)=0.5. The upper portion of Figure 2 plots an histogram of the distribution of relative frequencies for all test word-superordinate pairs; a large fraction of superordinates have a relative frequency of 0, i.e., were never selected. In particular, very general candidates such as "entity", "abstraction" or "psychological feature" are by and large ignored. The lower portion plots the distribution of relative frequencies of the 301 candidates that obtained most votes. A large fraction of the favorite superordinates obtained more than half of the votes, about 70% of these have a relative frequency of 0.6 or higher.

Statistical Analysis

To estimate the fraction of nouns for which there was one clear favorite superordinate, and evaluate the consistency of the experimental data, we performed a statistical analysis. First we used the K statistic (c.f. Di Eugenio & Glass, 2004). $K = \frac{P(A)-P(E)}{1-P(E)}$, where $P(A)$ is the agreement between participants and $P(E)$ is the expected agreement at chance. Unfortunately K proved inadequate to our case. The data consists of a single row of values for each noun, often characterized by very skewed distributions of votes. These cases yield odd values for K; e.g., if a category has 11 votes $P(A) = 0.83$, $P(E) = 0.84$ and $K = -0.0625$. For values of K close to 0 agreement is assessed as close to chance, while a generally accepted cutoff for "moderate" agreement is 0.67 (Agresti, 1992); although, determining significance levels for K is problematic in itself (Fleiss (1981) for example indicates the interval 0.40-0.75 as an indicator of "fair to good agreement, beyond chance"). In our case the K statistic fails to provide a meaningful assessment of the agreement rate because the expected agreement at chance tend to be unreasonably large especially when the distribution of votes is more skewed; i.e., when there is more agreement. Di Eugenio and Glass (2004) call this the "prevalence" problem with K.[6]

Based on these considerations, we developed an alternative analysis of the data. The aggregated data for each word defines a multinomial random variable which is the result of an experiment consisting of $n = 12$ trials, the number of participants, with k possible outcomes, the number of superordinates. We indicate with k_1 the category that obtained more votes from the data. If there is no agreement among participants one would expect

---

[6]Computing P(E) as $1/k$, where k is the number of categories, thus the theoretical expected agreement at chance, does not solve the problem. The outcomes of K are more meaningful but still suspicious; e.g., in the case of "tabloid" 9 participants chose "print_media" (75%), 2 chose "journalism" and 1 "medium"; this yields a value of K of 0.51 which is still below the commonly accepted threshold for moderate agreement. For completeness we report the average K value on all nouns which was equal to 0.47.

w	k	$k_1/P(k_1)$	$P(H_0)$	Sig
altruism	8	unselfishness/0.42	0.095	NO
forum	4	meeting/0.75	0.002	YES
sidewinder	13	snake/0.917	0	YES

Table 2. Three examples of the results of the significance test: w is the test word, k is the number of candidates, $P(k_1)$ is the probability of the highest-scoring category from the data. $P(H_0)$ is the probability of H_0 estimated with the simulation.

the distribution of votes to be "approximately" uniform. Within this model one can define a null hypothesis according to which the probability of each outcome is the same; i.e., $H_0 : p_1 = p_2 = ... = p_k$. A multinomial experiment under H_0 can be performed by generating one of the k possible categories at random n times. From the outcome of this experiment the probability of the most likely category is computed. If this value is greater or equal to $P(k_1)$ then it is possible, under H_0, to generate a distribution that is consistent with the data. After repeating this experiment, the fraction of times H_0 is consistent yields the significance level at which H_0 can be rejected. We ran this experiment 10,000 times.

Notice that in this model a simulated distribution is consistent with H_0 not just when the distribution is close to uniform, but, more conservatively, when sampling under H_0 it is possible to generate a distribution with one "spike" that is consistent with the experimental data. For example, the following is the data for the noun "apostle": "entity/0", "living_thing/0", "organism/0", "causal_agent/0", "object/0", "advocate/1", "person/3", "believer/3", "supporter/5". This is hardly a uniform distribution of votes, 5 categories out of 9 have no votes and there is one with more than 40% of the votes. However, this is a non-significant case with $P(H_0) = 0.062$, because more than 5% of the times it is possible to generate a distribution where the probability of the category with more votes, whichever it is, is greater or equal to $P(supporter)$. Table 2 illustrates three more cases with different significance levels.

Discussion

For 84.4% of the words H_0 is rejected with $p < 0.05$; i.e., 84.4% of the time there was a superordinate which was chosen as more informative by a large enough number of participants. We found 249 different most informative superordinates, the most frequent are "person" (11), "plant part" (8), "plant" (7), "animal" (4). The words on which there is less agreement all refer to abstract concepts: "act", "cognition", "communication", "quantity", "relation", "state", and "time", with average $P(H_0) = 0.11$. It is possible that the organization of such classes within a taxonomic structure does not reflect how people tend to categorize them. The opposite is true of proper nouns. Participants were, in general, absolutely positive of what kind of thing (or who) each proper noun was; e.g., Baltimore is a "city" (much more than an "urban area", "municipality", "location", "district" etc.), Elvis Presley is a "rock-star" (much more than a "musician", "singer", "performer", "entertainer", etc.), and Scotland Yard is a "law enforcement agency" (much

more than a "police", "organization", "administrative unit" etc.). On average $P(H_0)$ was equal to 0.0317, 0.038 for common nouns alone. This results prove that for a large fraction of nouns, which are instances of very different semantic categories from "artifacts" and "plants" to "substances" and "groups", there is one superordinate which is clearly more informative than the others for a significant number of participants. This means that the capacity for recognizing the most informative taxonomic level has quite broad conceptual coverage. Furthermore, while different people might have different preferred superordinates, participants showed a solid agreement on which superordinate might be the most informative for somebody else, i.e., in a shared context.

Ranking Superordinates

We now investigate the properties that characterize the most informative superordinates of nouns.

Preliminaries

One way of formalizing the task of recognizing the most informative categories in a taxonomy is as a *ranking* problem. We adapt here a notation used for the problem of re-ranking parse trees (Collins, 2000). In our case there are $n = 301$ test words, for each word w_i there are k_i possible superordinates of concept c_i, the leaf-concepts w_i belongs to. We define as c_{ij} the jth superordinate of c_i and c_{i1} the highest scoring candidate according to the results of the experiment; i.e., the most informative superordinate. The goal in a ranking problem is to find good scoring functions $F(c_{ij})$; i.e., functions that assign a score to c_{i1} that is higher than the score for the other candidates c_{ij}. More precisely we are interested in functions that minimize a ranking *loss function*, which counts the number of the time a candidate $c_{ij} \neq c_{i1}$ is scored by $F(.)$ higher than c_{i1}:

$$RankLoss = \sum_{i}^{n} \sum_{j \geq 2}^{k_i} [\![F(c_{i1}) < F(c_{ij})]\!] \qquad (1)$$

where $[\![.]\!]$ is the indicator function.

Basic scoring functions

We define a set of scoring functions based on known properties of basic categories. In particular we are interested in properties that can be extracted from corpus data or from the taxonomy. Basic categories are typically expressed by short frequent words (Rosch et al., 1976). They dominate many subordinate categories very similar to each other (Gluck & Corter, 1985), hence if we consider a concept as a random variable whose possible outcomes are its children concepts then this variable should be characterized by high entropy. For example, a concept like "tree", which dominates several kinds of trees, many of which have similar frequencies (oak, pine, elm, redwood, etc.), has a higher entropy than "entity" which dominates very dissimilar things. The information-theoretical interpretation of the basic level leads also to a characterization of good superordinates as those which provide the greatest reduction in

	Ranking function					
Score	L	H	FR	LR	PMI	SD
RA	44.4	59.1	61.6	70.4	79.0	**87.0**
EM	12.0	24.9	26.2	35.9	46.5	**56.1**

Table 3. Results of all the ranking functions.

uncertainty about the target noun, or that are strongly correlated with the noun. Thus, we also introduce two functions that implement this intuition. Finally we noticed from the participants' data that preferred superordinates are often low-level classes and we define a feature also for this notion. We denote with w_i a test word, with c_i its word sense, with c_{ij} one of the superordinates of c_i and with w_{ij} the first noun in the synset c_{ij}. The following are the basic features we used to build scoring functions:

1. FR(c_{ij}): frequency of w_{ij}

2. L(c_{ij}): number of characters of w_{ij}

3. H(c_{ij}): entropy of c_{ij} calculated as $-\sum_k P(c_{ijk}) \log P(c_{ijk})$; c_{ijk} is a child of c_{ij}, and $P(c_{ijk}) = \frac{counts(c_{ijk})}{\sum_k counts(c_{ijk})}$

4. PMI(c_{ij}): point-wise mutual information between w_i and w_{ij}

5. LR(c_{ij}): likelihood ratio $-2 \log \frac{L(H_1)}{L(H_2)}$

6. SD(c_{ij}): length of the shortest path from c_i to c_{ij}

Frequencies in (1) are collected using "Yahoo!". For (3) we used frequencies from the Brown corpus for all words in Wordnet (plus a smoothing count of 1) and added the counts of each word to all its superordinates. Functions (4) and (5) are designed to capture how much the frequencies of two words are correlated. Point-wise mutual information measures how much information one word contains about another word, it is computed as $I(w_{ij}, w_i) = \log \frac{P(w_i, w_{ij})}{P(w_i)P(w_{ij})}$. Function (5) formulates a log-likelihood chi-squared statistic comparing the hypothesis that the distributions of w_i and w_{ij} are independent (H_1) against the hypothesis that they are dependent (H_2). The frequencies for (4) and (5) were also computed from "Yahoo!". For (6) we measured the distance in edges of the shortest path between c_i and c_{ij}.[7] Each of these functions implicitly defines a ranking by assigning a score to each concept. For example, PMI(garbage,waste) = 2.9, while PMI(garbage,material) = 1.4. Using PMI "waste" is ranked higher than "material" with respect to the noun "garbage".

Evaluation

The accuracy of each ranking function is evaluated by computing its ranking error on the experimental data.

---

[7]Wordnet's nominal taxonomy is not a tree since there exists many concepts with more than one parent, to find the shortest path we used Dijkstra's algorithm.

Since each word can have a different number of categories and there are a few cases in which there are two correct answers, i.e., two highest scoring classes, we computed the ranking error rate with the following variant of Equation 1 which is used for multi-label ranking problems (Schapire & Singer, 2000):

$$E_R = \frac{1}{n} \sum_{j \geq 2}^{k_i} \frac{1}{Z_i} [\![F(c_{i1}) \leq F(c_{ij})]\!] \qquad (2)$$

where $|c_{i1}|$ is the number of correct answers for word w_i and $Z_i = |c_{i1}|(k_i - |c_{i1}|)$. For word length (2), and distance (6), we used the negative of these values because we prefer short words and low classes. In Table 3 we report *ranking accuracy*, RA $= (1 - E_R)$*100, and also the fraction of times the correct answer is the top scoring superordinate, or *exact match* (EM).

Length (L) provides the worst ranking function. Participants don't always prefer short nouns as their answers. Entropy (H) and frequency (FR) are considerably more accurate, 59.1% and 61.6% ranking accuracy, 24.9% 26.2% exact match accuracy. Likelihood ratio (LR), RA=70.4%, EM=35.9%, and particularly mutual information (PMI), RA=79%, EM=46.5%, capture even more robust regularities in the data. A good superordinate is likely to be characterized by a strong distributional correlation with the noun and, even more, by the reduction in uncertainty which provides with respect to the target noun. Finally, the evaluation shows that distance (SD) is the most reliable ranking function: RA=87%, EM=56.1%. People often find the most informative levels to be among the most specific superordinates, 82% of the times the preferred superordinates is within two categories above the noun. This is even more surprising considering that participants saw a list of superordinate terms in randomized order.

Discussion

The most informative superordinates are characterized by well-known properties of basic levels; 83.3% of the time the output of at least one of the ranking functions is the same as the correct answer. There might be a model which combines the individual sources of information and fits the experimental data more accurately. A simple way of combining different ranking functions is by defining a new function which combines the predictions of the individual functions as a weighted sum yielding the posterior probability of a superordinate c_{ij} given a set of scoring functions $\mathcal{F} = \{FR, L, H, PMI, LR, SD\}$ (Florian & Yarowsky, 2002):

$$P(c_{ij}|\mathcal{F}) = \frac{\sum_{F \in \mathcal{F}} \lambda_F rank_F(c_{ij})}{\sum_j \sum_{F \in \mathcal{F}} \lambda_F rank_F(c_{ij})} \qquad (3)$$

where λ is a vector of parameters adjusted to weigh each function individually in order to maximize accuracy. We adjusted λ with a line search using as development data the results for the 213 words of the pilot study. Unfortunately this method doesn't work, RA=86.8%, EM=58.6%, which is comparable to the distance measure but not better. The problem is that the different

functions are all strongly correlated. This fact can be verified by comparing the individual functions outcomes; i.e., success or failure in predicting the true superordinate of a noun. The pairs of functions with highest coefficient of correlation are SD and PMI ($\rho = 0.538$), FR and LR ($\rho = 0.527$), L and FR ($\rho = 0.4227$) etc. Because of this high degree of correlation, and because there is a strong bias for specificity, the simple combination model is forced to put so much weight on the distance feature that the others rarely make a difference.[8]

Given the specificity bias, an interesting question is when do people decide to use a more general class? Interestingly a measure that is correlated with the participants' use of a more general superordinate is when the different functions predict conflicting categories. This uncertainty can be measured using entropy; e.g., for the noun "remembrance" the six functions predict 5 times "memory" (the correct class) and once "ability", thus the entropy is low (0.65), in this case participants chose the most specific superordinate. Entropy of the predictions and presence of a generalization are positively correlated ($\rho = 0.18$). Uncertainty in this case indicates that there is no obvious most informative candidate. In such cases, often the preferred superordinate can be associated with at least one of the identified properties. However, the specific choice seems to involve complex semantic and cultural inferences triggered by the superordinate term. For example, "person" is used for nouns such as "aborigine" and "alcoholic", it is possible that participants in these cases avoid the use of terms with strong cultural connotations such as "primitive" or "drunkard" (but "trouble shooter" is a "repairman"). General terms are also used when the alternatives presuppose some technical knowledge; e.g., "plant" instead of "rhododendron" for "azalea" (but "flower" for "chrysanthemum") and "animal" instead of "placental" for "anteater" (but "bear" for "grizzly"). Further research will be necessary to gain a better understanding of these issues.

Conclusion

We investigated the coverage and properties of informative superordinates using statistics from a corpus of word frequencies and associations. We found that for a vast fraction of nouns humans have converging preferences about which superordinate is more informative. In little more than half of the cases the most informative class is the most specific one. Informative superordinates are characterized by properties of basic concepts such as frequency and association measures, however a complete understanding of the determinants of the basic-level will require appealing to principles beyond orthography, frequency, and hierarchical structure. For example, causal knowledge and pragmatic principles are also relevant.

Acknowledgments

We would like to thank Elie Bienenstock and Fulvio Domini for helpful discussions and comments.

---

[8]Experiments with a more sophisticated ranking function based on Boosting (Collins, 2000) improved only slightly over the distance model.

References

Agresti, A. (1992). Modeling Patterns of Agreement and Disagreement. *Statistical Methods in Medical Research*, 1.

Brown, R. (1958). How Shall a Thing be Called?. *Psychological Review*, 65, 14-21;

Ciaramita, M. & Johnson, M. (2003). Supersense Tagging of Unknown Nouns in WordNet. In *Proceedings of the 2003 Conference on Empirical Methods in Natural Language Processing*.

Collins, M. (2000). Discriminative Reranking for Natural Language Parsing. In *Proceedings of the 17th ICML*.

Corter, J. & Gluck, M. (1992). Explaining Basic Categories: Feature Predictability and Information. *Psychological Bulletin*. 111(2).

Dale, R. & Reiter, E. (1995). Computational Interpretation of the Gricean Maxims in the Generation of Referring Expressions, *Cognitive Science*, 19.

Di Eugenio, B. & Glass, M. (2004). The Kappa Statistics: A Second Look. *Computational Linguistics*, 30(1).

Fellbaum, C. (1998). *WordNet: An Electronic Lexical Database*. Cambridge, MA: MIT Press.

Fisher, D.H. (1988). A Computational Account of Basic Level and Typicality Effects. In *Proceedings of AAAI 1988*.

Fleiss, J.L (1981). *Statistical Methods for Rates and Proportions*, (2nd ed.). New York, NY: Wiley.

Florian, R. & Yarowsky, D. (2002). Modeling Consensus: Classifiers Combination for Word Sense Disambiguation. In *Proceedings of the 2002 Conference on Empirical Methods in Natural Language Processing*

Francis, W & Kučera, H (1982). *Frequency Analysis of English Usage: Lexicon and Grammar*. Boston, MA: Houghton Mifflin.

Gluck, M. & Corter, J. (1985). Information, Uncertainty and the Utility of Categories. In *Proceedings of the 7th Annual Conference of the Cognitive Science Society*.

Murphy, G.L. (2002). *The Big Book of Concepts*. Cambridge, MA: MIT Press..

Rosch, E. & Mervis, C.B & Gray, W.D. & Johnson, D.M. & Boyes-Braem, P. (1976). Basic Objects in Natural Categories, *Cognitive Psychology* 8.

Schapire, R.E. & Singer, Y. (2000). BoosTexter: A Boosting-Based System for Text Categorization. *Machine Learning*, 39.

Sloman, S. & Ahn, W. (1999). Feature Centrality: Naming versus Imagining. *Journal of Memory and Cognition*, 27(3), 526-37.

Tversky, B. & and Hemenway, K. (1984). Objects, Parts, and Categories. *Journal of Experimental Psychology: General*, 113(2), 169-97.