Linguistic categorization and complexity

Katya Pertsova
UNC-Chapel Hill
Linguistics Dept, CB 3155
Chapel Hill, NC 27599, USA
pertsova@unc.edu

Abstract

This paper presents a memoryless categorization learner that predicts differences in category complexity found in several psycholinguistic and psychological experiments. In particular, this learner predicts the order of difficulty of learning simple Boolean categories, including the advantage of conjunctive categories over the disjunctive ones (an advantage that is not typically modeled by the statistical approaches). It also models the effect of labeling (positive and negative labels vs. positive labels of two different kinds) on category complexity. This effect has implications for the differences between learning a single category (e.g., a phonological class of segments) vs. a set of non-overlapping categories (e.g., affixes in a morphological paradigm).

1 Introduction

Learning a linguistic structure typically involves categorization. By "categorization" I mean the task of dividing the data into subsets, as in learning what sounds are "legal" and what are "illegal," what morpheme should be used in a particular morphosyntactic context, what part of speech a given words is, and so on. While there is an extensive literature on categorization models within the fields of psychology and formal learning, relatively few connections have been made between this work and learning of linguistic patterns.

One classical finding from the psychological literature is that the subjective complexity of categories corresponding to Boolean connectives follows the order shown in figure 1 (Bruner et al., 1956; Neisser and Weene, 1962; Gottwald, 1971). In psychological experiments subjective complexity is measured in terms of the rate and accuracy of learning an artificial category defined by some (usually visual) features such as color, size, shape, and so on. This finding appears to be consistent with the complexity of isomorphic phonological and morphological linguistic patterns as suggested by typological studies not discussed here for reasons of space (Mielke, 2004; Cysouw, 2003; Clements, 2003; Moreton and Pertsova, 2012). Morphological patterns isomorphic to those in figure 1 appear in figure 2.

The first goal of this paper is to derive the above complexity ranking from a learning bias. While the difficulty of the XOR category is notorious and it is predicted by many models, the relative difference between AND and OR is not. This is because these two categories are complements of each other (so long as all features are binary), and in this sense have the same structure. A memorizing learner can predict the order AND > OR simply because AND has fewer positive examples, but it will also incorrectly predict XOR > OR and AND > AFF. Many popular statistical classification models do not predict the order AND > OR (such as models based on linear classifiers, decision tree classifiers, naive Bayes classifiers, and so on). This is because the same classifier would be found for both of these categories given that AND and OR differ only with respect to what subset of the stimuli is assigned a positive label. Models proposed by psychologists, such as SUSTAIN (Love et al., 2004), RULEX (Nosofsky et al., 1994b), and Configural Cue (Gluck and

Proceedings of the Twelfth Meeting of the Special Interest Group on Computational Morphology and Phonology (SIGMORPHON2012), pages 72–81,
Montréal, Canada, June 7, 2012. ©2012 Association for Computational Linguistics

AFF (affirmation)	*AND*	*OR*	*XOR/↔*
$\boxed{●}$ ▲	$\boxed{●}$ ▲	● $\boxed{▲}$	● $\boxed{▲}$
$\boxed{○}$ △	○ △	$\boxed{○}$ $\boxed{△}$	$\boxed{○}$ △
circle	circle AND black	triangle OR white	(black AND triangle) OR (white AND circle)

Figure 1: Boolean categories over two features, *shape* and *color*: AFF > AND > OR > XOR

affirmation			*AND*			*OR*			*XOR/↔*		
	sg	pl		sg	pl		sg	pl		sg	pl
−part m.	-	$\boxed{\text{-im}}$	−part	$\boxed{\text{-s}}$	-	−*poss*	-	$\boxed{\text{-s}}$	acc.	-	$\boxed{\text{-s}}$
+part m.	-	$\boxed{\text{-im}}$	+part	-	-	+*poss*	$\boxed{\text{-s}}$	$\boxed{\text{-s}}$	nom.	$\boxed{\text{-s}}$	-
Hebrew, verb agreement in pres.			English, verb agreement in pres.			English nouns			Old French, o-stem nouns		

Figure 2: Patterns of syncretism isomorphic to the structure of Boolean connectives

Bower, 1988) also do not predict the order AND > OR fore similar reasons. Feldman (2000) speculates that this order is due to a general advantage of the UP-versions of a category over the DOWN-versions (for a category that divides the set of instances into two uneven sets, the UP-version is the version in which the smaller subset is positively labeled, and the DOWN-version is the version in which the larger subset is positively labeled). However, he offers no explanation for this observation. On the other hand, it is known that the choice of representations can affect learnability. For instance, k-DNF formulas are not PAC-learnable while k-CNF formulas describing the same class of patterns are PAC-learnable (Kearns and Vazirani, 1994). Interestingly, this result also shows that conjunctive representations have an advantage over the disjunctive ones because a very simple strategy for learning conjunctions (Valiant, 1984) can be extended to the problem of learning k-CNFs. The learner proposed here includes in its core a similar intersective strategy which is responsible for deriving the order AND > OR.

The second goal of the paper is to provide a unified account of learning one vs. several categories that partition the feature space (the second problem is the problem of learning paradigms). The most straight-forward way of doing this – treating category labels as another feature with n values for n labels – is not satisfactory for several reasons discussed in section 2. In fact, there is empirical evidence that the same pattern is learned differently depending on whether it is presented as learning a distinction between positive and negative instances of a category or whether it is presented as learning two different (non-overlapping) categories. This evidence will be discussed in section 3.

I should stress that the learner proposed here is not designed to be a model of "performance." It makes a number of simplifying assumptions and does not include parameters that are fitted to match the behavioral data. The main goal of the model is to predict the differences in subjective complexity of categories as a function of their logical structure and the presence/absence of negative examples.

2 Learning one versus many categories

Compare the task of learning a phonological inventory with the task of learning an inventory of morph-meaning pairs (as in learning an inflectional paradigm). The first task can be viewed as dividing the set of sounds into attested and non-attested ("accidental gaps"). At first glance, the second task can be analogously viewed as dividing the set of stimuli defined by morpho-syntactic features plus an n-ry feature (for n distinct morphs) into possible vs. impossible combinations of morphs and meanings. However, treating morphs as feature values leads to the possibility of paradigms in which

Neutral (AND/ORn)

	f1	$\overline{f1}$
f2	A	B
$\overline{f2}$	B	B

Biased

ANDb

	f1	$\overline{f1}$
f2	A	$\neg A$
$\overline{f2}$	$\neg A$	$\neg A$

ORb

	f1	$\overline{f1}$
f2	$\neg A$	A
$\overline{f2}$	A	A

Table 1: Three AND/OR conditions in Gottwald's study

different morphs are used with exactly the same set of features as well as paradigms with "accidental gaps," combinations of morpho-syntactic feature values that are impossible in a language. In fact, however, morphs tend to partition the space of possible instances so that no instance is associated with more than one morph. That is, true free variation is really rare (Kroch, 1994). Secondly, system-wide rather than lexical "accidental gaps" are also rare in morphology (Sims, 1996). Therefore, I construe the classification problem in both cases as learning a set of non-overlapping Boolean formulas corresponding to categories. This set can consist of just one formula, corresponding to learning a single category boundary, or it can consist of multiple formulas that partition the feature space, corresponding to learning non-overlapping categories each associated with a different label.

3 Effects of labeling on category complexity

A study by Gottwald (1971) found interesting differences in the subjective complexity of learning patterns in figure 1 depending on whether the data was presented to subjects as learning a single category (stimuli were labeled A vs. $\neg A$) or whether it was presented as learning two distinct categories (the same stimuli were labeled A vs. B). Following this study, I refer to learning a single category as "biased labeling" (abbreviated b) and learning several categories as "neutral labeling" (abbreviated n). Observe that since the AND/OR category divides the stimuli into unequal sets, it has two different biased versions: one biased towards AND and one biased towards OR (as demonstrated in table 1). The order of category complexity found by Gottwald was

AFFn, AFFb > ANDb > AND/ORn > ORb, XORb > XORn

These results show that for the XOR category the neutral labeling was harder than biased labeling. On the other hand, for the AND/OR category the neutral labeling was of intermediate difficulty, and, interestingly, easier than ORb. This is interesting because it goes against an expectation that learning two categories should be harder than learning one category. Pertsova (2012) partially replicated the above finding with morphological stimuli (where null vs. overt marking was the analog of biased vs. neutral labeling). Certain results from this study will be highlighted later.

4 The learning algorithm

This proposal is intended to explain the complexity differences found in learning categories in the lab and in the real world (as evinced by typological facts). I focus on two factors that affect category complexity, the logical structure of a category and the learning mode. The learning mode refers to biased vs. neutral labeling, or, to put it differently, to the difference between learning a single category and learning a partition of a feature space into several categories. The effect of the learning mode on category complexity is derived from the following two assumptions: (i) the algorithm only responds to negative instances when they contradict the current grammar, and (ii) a collection of instances can only be referred to if it is associated with a positive label. The first assumption is motivated by observations of Bruner et. al (1956) that subjects seemed to rely less on negative evidence than on positive evidence even in cases when such evidence was very informative. The second assumption corresponds to a common sentiment that having a linguistic label for a category aids in learning (Xu, 2002).

4.1 Some definitions

For a finite nonempty set of features F, we define the set of *instances* over these features, $I(F)$, as follows. Let R_f be a set of feature values for a feature f (e.g., $R_{height} = \{high, mid, low\}$). Each instance i is a conjunction of feature values given by the functions $f \rightarrow R_f$ for all features $f \in F$. A category is a set of instances that can be described by some non-contradictory Boolean formula ϕ.[1] Namely, ϕ describes a set of instances X if and only if it is logically equivalent to the disjunction of all instances in X. For instance, in the world with three binary features p, q, w, the formula $p \wedge q$ describes the set of instances $\{\{pqw\}, \{pq\bar{w}\}\}$ (where each instance is represented as a set). We will say that a formula ψ *subsumes* a formula ϕ if and only if the set of instances that ψ describes is a superset of the set of instances that ϕ describes. An empty conjunction $\emptyset$ describes the set of all instances.

The goal of the learner is to learn a set of Boolean formulas describing the distribution of positive labels (in the neutral mode all labels are positive, in the biased mode there is one positive label and one negative label). A formula describing the distribution of a label l is encoded as a set of entries of the form e_{l_i} (an i-th entry for label l). The distribution of l is given by $e_{l_1} \vee \ldots \vee e_{l_n}$, the disjunction of n formulas corresponding to entries for l. Each entry e_{l_i} consists of two components: a maximal conjunction ϕ_{max} and an (optional) list of other formulas EX (for exceptions). A particular entry e with two components, $e[\phi_{max}]$ and $e[EX] = \{\phi_1 \ldots \phi_n\}$, defines the formula $e[\phi_{max}] \wedge \neg(\phi_1 \vee \phi_2 \vee \ldots \vee \phi_n)$. $e[\phi_{max}]$ can intuitively be thought of as a rule of thumb for a particular label and EX as a list of exceptions to that rule. In the neutral mode exceptions are pointers to other entries or, more precisely, formulas encoded by those entries. In the biased mode they are formulas corresponding to instances (i.e., conjunctions of feature values for all features). The algorithm knows which mode it is in because the biased mode contains negative labels while the neutral mode does not. Finally, an instance i is *consistent* with an entry e if and only if the conjunction encoded by i logically implies the formula encoded by e. For example, an instance $\{pqw\}$ is consistent with an entry encoding the formula $\{p\}$.

Note that while this grammar can describe arbitrarily complex patterns/partitions, each entry in the neutral learning mode can only describe what linguistics often refer to as "elsewhere" patterns (more precisely Type II patterns in the sense of Pertsova (2011)). And the $e[\phi_{max}]$ component of each entry by definition can only describe conjunctions. There are additional restrictions on the above grammar: (i) the exceptions cannot have a wider distribution than "the rule of thumb" (i.e., an entry e_l cannot correspond to a formula that does not pick out any instances), (ii) no loops in the statement of exceptions is possible: that is, if an entry A is listed as an exception to the entry B, then B cannot also be an exception for A (a more complicated example of a loop involves a longer chain of entries).

When learning a single category, there is only one entry in the grammar. In this case arbitrarily complex categories are encoded as a complement of some conjunction with respect to a number of other conjunctions (corresponding to instances).

4.2 General description

The general organization of the algorithm is as follows. Initially, each positive label is assumed to correspond to a single grammatical entry, and the ϕ_{max} component of this entry is computed incrementally through an intersective generalization strategy that extracts features invariant across all instances used with the same label. When the grammar overgeneralizes by predicting two different labels for at least one instance, exceptions are introduced. The process of exception listing can also lead to overgeneralizations if exceptions are pointers to other entries in the grammar. When these overgeneralizations are detected the algorithm creates another entry for the same label. This latter process can be viewed as positing homophonous entries when learning form-meaning mappings, or as creating multiple "clusters" for a single category as in the prototype model SUSTAIN (Love et al., 2004), and it corresponds to explicitly positing a disjunctive rule. Note that if exceptions are not formulas for other labels, but individual instances, then exception listing does not lead to overgeneralization and no sub-entries are introduced. Thus, when learning a single category the learner generalizes by using an intersective strategy, and then lists exceptions one-by-one as they are discovered in form of negative evidence.

The problem of learning Boolean formulas is known to be hard (Dalmau, 1999). However, it is plausible that human learners employ an algorithm that is not generally efficient, but can easily handle certain restricted types of formulas under certain

[1] The set of Boolean formulas is obtained by closing the set of feature values under the operations of conjunction, negation, and disjunction.

simple distributions of data. (Subclasses of Boolean formulas are efficiently learnable in various learning frameworks (Kearns et al., 1994).) If the learning algorithm can easily learn certain patterns (providing an explanation for what patterns and distributions count as simple), we do not need to require that it be in general efficient.

4.3 Detailed description

First I describe how the grammar is updated in response to the data. The update routine uses a strategy that in word-learning literature is called cross-situational inference. This strategy incrementally filters out features that change from one instance to the next and keeps only those features that remain invariant across the instances that have the same label. Obviously, this strategy leads to overgeneralizations, but not if the category being learned is an affirmation or conjunction. This is because affirmations and conjunctions are defined by a single set of feature values which are shared by all instances of a category (for proof see Pertsova (2007) p. 122). After the entry for a given label has been updated, the algorithm checks whether this entry subsumes or is subsumed by any other entry. If so, this means that there is at least one instance for which several labels are predicted to occur (there is competition among the entries). The algorithm tries to resolve competition by listing more specific entries as exceptions to the more general ones.[2] However there are cases in which this strategy will either not resolve the competition, or not resolve it correctly. In particular, the intermediate entries that are in competition may be such that neither subsumes the other. Or after updating the entries using the intersective strategy one entry may be subsumed by another based on the instances that have been seen so far, but not if we take the whole set of instances into account. These cases are detected when the predictions of the current grammar go against an observed stimulus (step 11 in the function "Update" below). Finally, exception listing fails if it would lead to a "loop" (see section 4.1). The XOR pattern is an example of a simple pattern that will lead to a loop at some point during learning. In general this happens whenever the distribution of the two labels are intertwined in such a way that neither can be stated as a complement of the invariant features of the other.

The following function is used to add an exception:

```
**AddException**(expEntry, ruleEntry):
1. **if** adding $expEntry$ to $ruleEntry[EX]$ leads
   to a loop **then** FAIL
2. **else** add $expEntry$ to $ruleEntry[EX]$
```

The routine below is called within the main function (presented later); it is used to update the grammar in response to an observed instance x with the label l_i (the index of the label is decided in the main function).

```
**Update**
**Input:** G (current grammar); $x$ (an observed instance), $l_i$ (a label for this instance)
**Output:** newG
 1: newG ← G
 2: **if** $\exists e_{l_i} \in newG$ **then**
 3:     $e_{l_i}[\phi_{max}] \leftarrow e_{l_i}[\phi_{max}] \cap x$
 4: **else**
 5:     add the entry $e_{l_i}$ to $newG$ with values
        $e_{l_i}[\phi_{max}] = x$; $e_{l_i}[EX] = \{\}$.
 6: **for all** $e_{l'_j} \in newG$ ($e_{l'_j} \neq e_{l_i}$) **do**
 7:     **if** $e_{l'_j}$ subsumes $e_{l_i}$ **then**
 8:         AddException($e_{l_i}, e_{l'_j}$)
 9:     **else if** $e_{l_i}$ subsumes $e_{l'_j}$ **then**
10:         AddException($e_{l'_j}, e_{l_i}$)
11: **if** $\exists e_{l'_j} \in newG$ ($l' \neq l$) such that $x$ is consistent
    with $e_{l'_j}$ **then**
12:     AddException($e_{l_i}, e_{l'_j}$)
```

Before turning to the main function of the algorithm, it is important to note that because a grammar may contain several different entries for a single label, this creates ambiguity for the learner. Namely, in case a grammar contains more than one entry for some label, say two A labels, the learner has to decide after observing a datum (x, A), which entry to update, e_{A_1} or e_{A_2}. I assume that in such cases the learner selects the entry that is most similar to the

[2] This idea is familiar in linguistics from at least the times of Pāṇini. In Distributed Morphology, it is referred to as the Subset Principle for vocabulary insertion (Halle and Marantz, 1993). Similar principles are assumed in rule-ordering systems and in OT (i.e., more specific rules/constraints are typically ordered before the more general ones).

current instance, where similarity is calculated as the number of features shared between x and $e_{A_i}[\phi_{max}]$ (although other metrics of similarity could be explored).

Finally, I would like to note that the value of an entry $e_l(x)$ can change even if the algorithm has not updated this entry. This is because the value of some other entry that is listed as an exception in $e_l(x)$ may change. This is one of the factors contributing to the difference between the neutral and the biased learning modes: if exceptions themselves are entries for other labels, the process of exception listing becomes generalizing.

```
Main
Input: an instance-label pair $(x, l)$, previous hypothesis $G$ (initially set to an empty set)
Output: newG (new hypothesis)
1: set $E$ to the list of existing entries for the label $l$ in $G$
2: $k \leftarrow |E|$
3: if $E \neq \{\}$ then
4:     set $e_{l_{curr}}$ to $e_{l_i} \in E$ that is most similar to $x$
5:     $E \leftarrow E - e_{l_{curr}}$
6: else
7:     $curr \leftarrow k + 1$
8: if $l$ is positive and $(\neg\exists e_{l_{curr}} \in G$ or $x$ is not consistent with $e_{l_{curr}})$ then
9:     if $update(G, x, l_{curr})$ fails then
10:        goto step 3
11:    else
12:        $newG \leftarrow update(G, x, l_{curr})$
13: else if $l$ is negative and there is an entry $e$ in G consistent with $x$ (positive label was expected) then
14:    add $x$ to $e[EX]$ and minimize $e[EX]$ to get $newG$
```

Notice that the loop triggered when $update$ fails is guaranteed to terminate because when the list of all entries for a label l is exhausted, a new entry is introduced and this entry is guaranteed not to cause $update$ to fail.

This learner will succeed (in the limit) on most presentations of the data, but it may fail to converge on certain patterns if the crucial piece of evidence needed to resolve competition is seen very early on and then never again (it is likely that a human learner would also not converge in such a case).

This algorithm can be additionally augmented by a procedure similar to the selective attention mechanism incorporated into several psychological models of categorization to capture the fact that certain hard problems become easy if a subject can ignore irrelevant features from the outset (Nosofsky et al., 1994a). One (not very efficient, but easy) way to incorporate selective attention into the above algorithm is as follows. Initially set the number of relevant features k to 1. Generate all subsets of F of length k, select one such subset F_k and apply the above learning algorithm assuming that the feature space is F_k. When processing a particular instance, ignore all of its features except those that are in F_k. If we discover two instances that have the same assignment of features in F_k but that appear with two different labels, this means that the selected set of features is not sufficient (recall that free variation is ruled out). Therefore, when this happens we can start over with a new F_k. If all sets of length k have been exhausted, increase k to $k + 1$ and repeat. As a result of this change, patterns definable by smaller number of features would generally be easier to learn than those definable by larger number of features.

5 Predictions of the model for learning Boolean connectives

We can evaluate predictions of this algorithm with respect to category complexity in terms of the proportion of errors it predicts during learning, and in terms of the computational load, roughly measured as the number of required runs through the main loop of the algorithm. Recall that a single data-point may require several such runs if the update routine fails and a new sub-category has to be created.

Below, I discuss how the predictions of this algorithm compare to the subjective complexity ranking found in Gottwald's experiment. First, consider the relative complexity order in the neutral learning mode: AFF > AND/OR > XOR.

In terms of errors, the AFF pattern is predicted to be learned without errors by the above algorithm (since the intersective strategy does not overgeneralize when learning conjunctive patterns). When learning an AND/OR pattern certain orders of data presentation will lead to an intermediate overgeneralization of the label associated with the disjunctive category to the rest of the instances. This will happen if the OR part of the pattern is processed before the AND part. When learning an XOR pattern, the learner is guaranteed to overgeneralize one of the labels on any presentation of the data. Let's walk through the learning of the XOR pattern, repeated below for convenience.

	f1	$\overline{f1}$
f2	A	B
$\overline{f2}$	B	A

Suppose for simplicity that the space of features includes only f1 and f2, and that the first two examples that the learner observes are $(A, \{f1, f2\})$ and $(A, \{\overline{f1}, \overline{f2}\})$. After intersecting $\{f1, f2\}$ and $\{\overline{f1}, \overline{f2}\}$ the learner will overgeneralize A to the whole paradigm. If the next example is $(B, \{f1, \overline{f2}\})$, the learner will partially correct this overgeneralization by assuming that A occurs everywhere except where B does (i.e., except $\{f1, \overline{f2}\}$). But it will continue to incorrectly predict A in the remaining fourth cell that has not been seen yet. When B is observed in that cell, the learner will attempt to update the entry for B through the intersection but this attempt will fail (because the entry for B will subsume the entry for A, but we can't list A as an exception for B since B is already listed as an exception for A). Therefore, a new sub-entry for B, $\{\overline{f1}, f2\}$, will be introduced and listed as another exception for A. Thus, the final grammar will contain entries corresponding to these formulas: $B : (\overline{f1} \wedge f2) \vee (f1 \wedge \overline{f2})$ and $A : \neg((\overline{f1} \wedge f2) \vee (f1 \wedge \overline{f2}))$.

Overall the error pattern predicted by the learner is consistent with the order AFF $>$ AND/OR $>$ XOR.

I now turn to a different measure of complexity based on the number of computational steps needed to learn a pattern (where a single step is equated to a single run of the main function). Note that the speed of learning a particular pattern depends not only on the learning algorithm but also on the distribution of the data. Here I will consider two possible probability distributions which are often used in categorization experiments. In both distributions the stimuli is organized in blocks. In the first one (which I call "instance balanced") each block contains all possible instances repeated once; in the second distribution ("label balanced") each block contains all possible instances with the minimum number of repetitions to insure equal numbers of each label. The distributions differ only for those patterns that have an unequal number of positive/negative labels (e.g., AND/OR). Let us now look at the minimum and maximum number of runs through the main loop of the algorithm required for convergence for each type of pattern. The minimum is computed by finding the shortest sequence of data that leads to convergence and counting the number of runs on this data. The maximum is computed analogously by finding the longest sequence of data. The table below summarizes min. and max. number of runs for the feature space with 3 binary features (8 possible instances) and for two distributions.

	Min	Max	Max
		(instance)	(label)
AFF	4	7	7
AND/OR	4	8	11
XOR	7	9	9

Table 2: Complexity in the neutral mode

The difference between AFF and AND/OR in the number of runs to convergence is more obvious for the label balanced distribution. On the other hand, the difference between AND/OR and XOR is clearer for the instance balanced distribution. This difference is not expected to be large for the label balanced distribution, which is not consistent with Gottwald's experiment in which the stimuli were label balanced, and neutral XOR was significantly more difficult to learn than any other condition.

We now turn to the biased learning mode. Here, the observed order of difficulty was: AFFb $>$ ANDb $>$ ORb, XORb. In terms of errors, both AFFb and ANDb are predicted to be learned with no errors since both are conjunctive categories. ORb is predicted to involve a temporary overgeneralization of the positive label to the negative contexts. The same is true for XORb except that the proportion of errors will be higher than for ORb (since the latter category has fewer negative instances).

The minimum and maximum number of runs required to converge on the biased categories for two types of distributions (instance balanced and label balanced) is given below. Notice that the minimum numbers are lower than in the previous table because in the biased mode some categories can be learned from positive examples alone.

	Min	Max	Max
		(instance)	(label)
AFFb	2	7	7
ANDb	2	8	8
ORb	4	16	22
XORb	6	16	16

Table 3: Complexity in the biased mode

The difference between affirmation and conjunction is not very large which is not surprising (both are conjunctive categories). Again we see that the two types of distributions give us slightly different predictions. While ANDb seems to be learned faster than ORb in both distributions, it is not clear whether and to what extent ORb and XORb are on average different from each other in the label balanced distribution. Recall that Gottwald found no significant difference between ORb and XORb (in fact numerically ORb was harder than XORb). Interestingly, in a morphological analogue of Gottwald's study in which the number of instances rather than labels was balanced, I found the opposite difference: ORb was easier to learn than XORb (the number of people to reach learning criterion was 8 vs. 4 correspondingly) although the difference in error rates on the testing trials was not significant (Pertsova, 2012). More testing is needed to confirm whether the relative difficulty of these two categories is reliably affected by the type of distribution as predicted by the learner.[3]

Finally, we look at the effect of labeling within each condition. In the AFF condition, Gottwald found no significant difference between neutral labeling and biased labeling. This could be due to the fact that subjects were already almost at ceiling in learning this pattern (median number of trials to convergence for both conditions was ≤ 5). In the AND/OR condition, Gottwald observed the interesting order ANDb $>$ AND/OR $>$ ORb. This order is also predicted by the current algorithm. Namely, the neutral category AND/OR is predicted to be harder than ANDb because (1) ANDb requires less computational resources (2) on some distributions of data overgeneralization will occur when learning an AND/OR pattern but not an ANDb category. The AND/OR $>$ ORb order is also predicted and is particularly pronounced for label balanced distribution. Since two labels are available when learning the AND/OR pattern, the AND portion of the pattern can be learned quickly and subsequently listed as an exception for the OR portion (which becomes the"elsewhere" case). On the other hand, when learning the ORb category, the conjunctive part of the pattern is initially ignored because it is not associated with a label. The learner only starts paying attention to negative instances when it overgeneralizes. For a similar reason, the biased XOR category is predicted to be harder to learn than the neutral XOR category. This latter prediction is not consistent with Gottwald's finding, who found XORn not just harder than other categories but virtually impossible to learn: 6 out of 8 subjects in this condition failed to learn it after more than 256 trials. In contrast to this result (and in line with the predictions of the present learner), Pertsova (2012) found that the neutral XOR condition was learned by 8 out of 12 subjects on less than 64 trials compared to only 4 out of 12 subjects in the biased XOR condition.

To conclude this section, almost all complexity rankings discussed in this paper are predicted by the proposed algorithm. This includes the difficult to model AND $>$ OR ranking which obtains in the biased learning mode. The only exception is the neutral XOR pattern, which was really difficult to learn in Gottwald's non-linguistic experiment (but not in Pertsova's morphological experiment), and which is not predicted to be more difficult than biased XOR. Further empirical testing is needed to clarify the effect of labeling within the XOR condition.

[3] Another possible reason for the fact that Gottwald did not find a difference between ORb and XORb is this: if selective attention is used during learning, it will take longer for the learner to realize that ORb requires the use of two features compared to XORb especially when the number of positive and negative examples are balanced. In particular, a one feature analysis of ORb can explain 5/6 of the data with label balanced stimuli, while a one feature analysis of XORb can only explain 1/2 of the data, so it will be quickly abandoned.

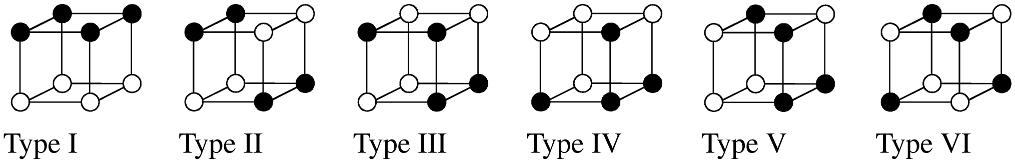

Figure 3: Shepard et. al. hierarchy

6 Other predictions

Another well-studied hierarchy of category complexity is the hierarchy of symmetric patterns (4 positive and 4 negative instances) in the space of three binary features originally established by Shepard et. al (1961). These patterns are shown in figure 3 using cubes to represent the three dimensional feature space.

Most studies find the following order of complexity for the Shepad patterns: I > II > III, IV, V > VI (Shepard et al., 1961; Nosofsky et al., 1994a; Love, 2002; Smith et al., 2004). However, a few studies find different rankings for some of these patterns. In particular, Love (2002) finds IV > II with a switch to unsupervised training procedure. Nosofsky and Palmeri (1996) find the numerical order I > IV > III > V > II > VI with intergral stimulus dimensions (feature values that are difficult to pay selective attention to independent of other features, e.g., hue, brightness, saturation). More recently Moreton and Persova (2012) also found the order IV > III > V, VI (as well as I > II, III, > VI) in an unsupervised phonotactics learning experiment.

So, one might wonder what predictions does the present learner make with respect to these patterns. We already know that it predicts Type I (affirmation) to be easier than all other types. For the rest of the patterns the predictions in terms of speed of acquisition are II > III > IV, V > VI in the neutral learning mode (similar to the typical findings). In the biased learning mode, patterns II through VI are predicted to be learned roughly at the same speed (since all require listing four exceptions). If selective attention is used, Type II will be the second easiest to learn after Type I because it can be stated using only two features. However, based on the error rates, the order of difficulty is predicted to be I > IV > III > V > II > VI (similar to the order found by Nosofsky and Palmeri (1996)). No errors are ever made with Type I. The proportion of errors in other patterns depends on how closely the positive examples cluster to each other. For instance, when learning a Type VI pattern (in the biased mode) the learner's grammar will be correct on 6 out of 8 instances after seeing any two positive examples (the same is not true for any other pattern, although it is almost true for III). After seeing the next instance (depending on what it is and on the previous input) the accuracy of the grammar will either stay the same, go up to 7/8, or go down to 1/2. But the latter event has the lowest probability. Note that this learner predicts non-monotonic behavior: it is possible that a later grammar is less accurate than the previous grammar. So, for a non-monotonic learner the predictions based on the speed of acquisition and accuracy do not necessarily coincide.

There are many differences across the categorization experiments that may be responsible for the different rankings. More work is needed to control for such differences and to pin down the sources for different complexity results found with the patterns in figure 3.

7 Summary

The current proposal presents a unified account for learning a single category and a set of categories partitioning the stimuli space. It is consistent with many predictions about subjective complexity rankings of simple categories, including the ranking AND > OR, not predicted by most categorization models, and the difference between the biased and the neutral learning modes not previously modeled to my knowledge.

References

Jerome S. Bruner, Jacqueline J. Goodnow, and George A. Austin. 1956. *A study of thinking*. John Wiley and Sons, New York.

George N. Clements. 2003. Feature economy in sound systems. *Phonology*, 20(3):287–333.

Michael Cysouw. 2003. *The paradigmatic structure of person marking*. Oxford studies in typology and linguistic theory. Oxford University Press, Oxford.

Víctor Dalmau. 1999. Boolean formulas are hard to learn for most gate bases. In Osamu Watanabe and Takashi Yokomori, editors, *Algorithmic Learning Theory*, volume 1720 of *Lecture Notes in Computer Science*, pages 301–312. Springer Berlin / Heidelberg.

Jacob Feldman. 2000. Minimization of Boolean complexity in human concept learning. *Nature*, 407:630–633.

Mark A. Gluck and Gordon H. Bower. 1988. evaluating an adaptive network model of human learning. *Journal of memory and language*, 27:166–195.

Richard L. Gottwald. 1971. Effects of response labels in concept attainment. *Journal of Experimental Psychology*, 91(1):30–33.

Morris Halle and Alec Marantz. 1993. Distributed morphology and the pieces of inflection. In K. Hale and S. J. Keyser, editors, *The View from Building 20*, pages 111–176. MIT Press, Cambridge, Mass.

Michael Kearns and Umesh Vazirani. 1994. *An introduction to computational learning theory*. MIT Press, Cambridge, MA.

Michael Kearns, Ming Li, and Leslie Valiant. 1994. Learning boolean formulas. *J. ACM*, 41(6):1298–1328, November.

Anthony Kroch. 1994. Morphosyntactic variation. In Katharine Beals et al., editor, *Papers from the 30th regional meeting of the Chicago Linguistics Society: Parasession on variation and linguistic theory*. Chicago Linguistics Society, Chicago.

Bradley C. Love, Douglas L. Medin, and Todd M. Gureckis. 2004. SUSTAIN: a network model of category learning. *Psychological Review*, 111(2):309–332.

Bradley C. Love. 2002. Comparing supervised and unsupervised category learning. *Psychonomic Bulletin and Review*, 9(4):829–835.

Jeff Mielke. 2004. *The emergence of distinctive features*. Ph.D. thesis, Ohio State University.

Elliott Moreton and Katya Pertsova. 2012. Is phonological learning special? Handout from a talk at the 48th Meeting of the Chicago Society of Linguistics, April.

Ulrich Neisser and Paul Weene. 1962. Hierarchies in concept attainment. *Journal of Experimental Psychology*, 64(6):640–645.

Robert M. Nosofsky and Thomas J. Palmeri. 1996. Learning to classify integral-dimension stimuli. *Psychonomic Bulletin and Review*, 3(2):222–226.

Robert M. Nosofsky, Mark A. Gluck, Thomas J. Palmeri, Stephen C. McKinley, and Paul Gauthier. 1994a. Comparing models of rule-based classification learning: a replication and extension of Shepard, Hovland, and Jenkins (1961). *Memory and Cognition*, 22(3):352–369.

Robert M. Nosofsky, Thomas J. Palmeri, and Stephen C. McKinley. 1994b. Rule-plus-exception model of classification learning. *Psychological Review*, 101(1):53–79.

Katya Pertsova. 2007. *Learning Form-Meaning Mappings in the Presence of Homonymy*. Ph.D. thesis, UCLA.

Katya Pertsova. 2011. Grounding systematic syncretism in learning. *Linguistic Inquiry*, 42(2):225–266.

Katya Pertsova. 2012. Logical complexity in morphological learning. In *Proceedings of the 38th Annual Meeting of the Berkeley Linguistics Society*.

Roger N. Shepard, C. L. Hovland, and H. M. Jenkins. 1961. Learning and memorization of classifications. *Psychological Monographs*, 75(13, Whole No. 517).

Andrea Sims. 1996. *Minding the Gaps: inflectional defectiveness in a paradigmatic theory*. Ph.D. thesis, The Ohio State University.

J. David Smith, John Paul Minda, and David A. Washburn. 2004. Category learning in rhesus monkeys: a study of the Shepard, Hovland, and Jenkins (1961) tasks. *Journal of Experimental Psychology: General*, 133(3):398–404.

Leslie G. Valiant. 1984. A theory of the learnable. In *Proceedings of the sixteenth annual ACM symposium on Theory of computing*, STOC '84, pages 436–445, New York, NY, USA. ACM.

Fei Xu. 2002. The role of language in acquiring object kind concepts in infancy. *Cognition*, 85(3):223 – 250.