SYNTACTIC CATEGORY LEARNING IN A SECOND LANGUAGE

by

Rachel Wilson

A Dissertation Submitted to the Faculty of the

DEPARTMENT OF PSYCHOLOGY

In Partial Fulfillment of the Requirements
For the Degree of

DOCTOR OF PHILOSOPHY

In the Graduate College

THE UNIVERSITY OF ARIZONA

2002

UMI Number: 3053882

UMI Microform 3053882

ProQuest Information and Learning Company
300 North Zeeb Road
P.O. Box 1346
Ann Arbor, MI 48106-1346

THE UNIVERSITY OF ARIZONA ®
GRADUATE COLLEGE

As members of the Final Examination Committee, we certify that we have read the dissertation prepared by RACHEL WILSON

entitled SYNTACTIC CATEGORY LEARNING IN A SECOND LANGUAGE

and recommend that it be accepted as fulfilling the dissertation requirement for the Degree of Doctor of Philosophy

Dr. LouAnn Gerken — Date 5/2/02

Dr. Janet Nicol — Date 5/2/02

Dr. Rebecca Gomez — Date 5/2/02

Date

Date

Final approval and acceptance of this dissertation is contingent upon the candidate's submission of the final copy of the dissertation to the Graduate College.

I hereby certify that I have read this dissertation prepared under my direction and recommend that it be accepted as fulfilling the dissertation requirement.

Dissertation Director Dr. LouAnn Gerken — Date May 7, 2002

STATEMENT BY AUTHOR

This dissertation has been submitted in partial fulfillment of requirements for an advanced degree at The University of Arizona and is deposited in the University Library to be made available to borrowers under rules of the Library.

Brief quotations from this dissertation are allowable without special permission, provided that accurate acknowledgment of source is made. Requests for permission for extended quotation from or reproduction of this manuscript in whole or in part may be granted by the head of the major department or the Dean of the Graduate College when in her or his judgment the proposed use of the material is in the interests of scholarship. In all other instances, however, permission must be obtained from the author.

SIGNED: Rachel Nelson

ACKNOWLEDGMENTS

Many people played a part in the writing of this dissertation. First, I would like to thank the members of my committee for both their academic and psychological support. Were it not for the dedication of LouAnn Gerken, Janet Nicol and Rebecca Gómez, this dissertation would never have been written.

Likewise, several research assistants helped to collect the data presented here. They are: Casey Given, Michael Steere, Jennie Bentley and Brett Barmore.

Erika Nurmsoo was invaluable in helping me get started on this line of work.

Dr. Del Phillips taught me Russian and took me to Russia ten times. How can I even begin to thank him for that here?

Thanks to my parents for their emotional support and for reminding me that a Ph.D. is a big deal.

A network of close friends kept me going during this process and kept me from going completely crazy. Thanks to Lydia Lester for being patient with my erratic behavior; thanks to Tania Zamuner for the sausages and the frequent calls from the Netherlands to cheer me up. And thanks to the entire circle of friends that supported me: Jon Folstein, Jesse Kirchner, Sonya Bird, Summer Mullins, Will Lewis, Scott Farrar, Peter Norquest, Nastia Snyder, George Torrieri, Jason Barker, Connie Clarke, Patrick Davidson, Celeste Espinoza, Tara Crisp, Billy Manes, Eric and Irina Rook, Aleksandr Aratsky, Anna Murzina and the Baimler family.

Thanks to the Peace and Justice community of Tucson for letting me have the time off to finish this dissertation.

TABLE OF CONTENTS

TABLE OF CONTENTS-- continued

TABLE OF CONTENTS-- continued

TABLE OF CONTENTS-- continued

LIST OF TABLES

ABSTRACT

A central question in the study of language learning is how humans acquire syntactic categorical distinctions among words (e.g. noun, verb, etc.). Past research using miniature artificial grammars suggests that semantic information is not needed for this learning; distributional information alone can provide adequate input for learning. The current experiments extended this finding to a natural language. Adults who had never studied Russian listened to lists of Russian words for seven minutes. The words consisted of a content morpheme and a grammatical ending. The participants were not told the meanings of the words. Next they were tested on a series of legal and non-legal morpheme sequences, including sequences that were not in the training. Results showed that participants were able to distinguish between new legal and non-legal morpheme sequences, provided there were at least two category-markers in the input. This suggests that they were generalizing the words into categories. A corpus study showed that Russian probably contains the kind and quantity of markings required for category learning to take place. Reaction times were also analyzed.

CHAPTER I. INTRODUCTION

A central question in the study of language learning is how humans acquire syntactic categorical distinctions among words. For example, how does a child learning English come to realize that the word 'dog' is a noun? Or that 'sit' is a verb? Or that the word 'swing' can be both a noun and a verb? It is clear that there are certain characteristics of nouns that set them apart from verbs: nouns denote objects, places, or people; nouns can be the subject or object of a sentence; they are able to be modified by adjectives; they can follow articles such as 'a' and 'the,' etc.. Verbs, on the other hand, denote actions and states. They function as predicates of sentences and are modified by adverbs. They can follow words such as 'to.' Thus, it is the case that there is a large amount of converging information available to the child that might play a role in the categorization process. However, it is not at all clear that children make use of all of this information in order to properly form syntactic categories.

The evidence that children can use converging evidence comes from research into the learning of sub-categories, that is, categories within categories, such as noun gender. For a sub-category, the set of available information is much smaller. For example, in some languages, such as Qafar and Hausa, (Corbett, 1981) reliable phonology is the only piece of information available to the child in distinguishing the genders of inanimate nouns. Since children are clearly able to acquire languages with this kind of gender distinction, it must be possible for them to properly categorize on the basis of a very limited set of information, such as phonology alone. Determining the nature of that limited set of information has been the subject of much research in developmental psycholinguistics.

This dissertation examines syntactic category learning in a second language. The participants in all the studies are adults and thus differ in crucial ways from infants learning their first language. For example, it is hypothesized that age of acquisition plays a crucial role in how syntactic structures are ultimately processed in normal language use (Emmorey, Bellugi, Friederici, & Horn, 1995). Additionally, simply because all the participants already know at least one language it is possible that they employ strategies that are unlike those of an infant who has no prior linguistic experience. Furthermore, the participants in these experiments were all college students and therefore had been exposed to formal teaching and learning for quite some time. That prior educational experience may also influence their strategies of language learning. However, despite these limitations, studying adults learning a second language can still provide valuable information about how infants might learn. At the very least, it provides infant researchers with possible directions for work.

This dissertation also seeks to examine second language learning in a relatively new way. The second language acquisition literature is rich with studies in classroom behavior, teaching methods and theories of how adults learn. However, little second language acquisition research has used methodology that is common in cognitive psychology and psycholinguistics. Therefore, this dissertation is part of a growing subfield of second language acquisition literature that draws its methodology from cognitive psychology. These methods include controlled learning experiments and the collection of reaction time data.

To summarize, this dissertation has three goals:

1) to examine how humans learn syntactic categories such as gender;

2) to contribute to the discussion of how first language acquisition takes place;

3) to use a cognitive psychology approach to examine second language learning;

The last section of this dissertation outlines several areas that, while outside the domain of the dissertation, may prove to be fruitful areas of study in the future, such as extending the work described here to infants. The questions examined in this dissertation are too broad for certain answers, but serve as a preliminary step in accomplishing the three aims outlined above.

CHAPTER II. LITERATURE REVIEW

A thorough investigation of linguistic gender comes from Corbett (1991), who points out several important things to keep in mind when examining the topic. First of all, it is important to realize that linguists use the term gender to mean 'kind,' and, therefore, languages can have genders which do not correspond to male and female. For example, Corbett mentions that Bantu languages frequently have between ten and twenty genders. These genders often do not correspond to semantic classes at all, let alone male and female.

Likewise, it is important to remember that not all languages have a gender system. This, of course, causes us to question the purpose of gender. Corbett hypothesizes that it is used in natural language processing to differentiate nouns in a clause. Languages have a variety of other ways of accomplishing this (case systems, verbal agreement, etc.), so it is puzzling why gender exists at all and why it often co-exists with other noun-differentiating systems.

Learning a gender system is an instance of category learning. That is, the learner must figure out that there are several categories of words and that those categories make up a coherent syntactic system. Given that languages are extremely diverse in the kinds of gender systems they have, it seems likely that mechanisms for learning gender in one language are different from the mechanisms of learning gender in another language.

Probably the most traditional account of category learning comes from Steven Pinker (1987), who claims that semantics are what guide categorical distinctions in language learning. Pinker terms this the "Semantic Bootstrapping Hypothesis, " in which semantics serve as a "bootstrap" with which children can "pull themselves up into the syntactic system." According to Pinker, a child would first learn the meaning of a word, say 'ball.' The child would learn that the word 'ball' means a spherical object by means of non-syntactic cues to the word's meaning: discourse, intonation, etc. Then, upon determining that the 'ball' is an object, the child's language-learning mechanisms would tag 'ball' as a noun. This is because, in Pinker's view, certain syntactic categories are universal (such as noun and verb) and those syntactic universals have close parallels with semantic information. In other words, the association of the semantic category 'object' with the syntactic category 'noun' is innate, thanks to a mapping mechanism which serves the express purpose of allowing language acquisition to take place.

However, even if this explanation is viable for category learning, it would seem to be an inadequate explanation of sub-category learning. This is because sub-categories, such as grammatical gender, rarely enjoy a reliable semantic corollary for all nouns.[1] As noted earlier, some languages (such as Qafar and Hausa) rely on strictly phonological markers to differentiate among genders. In cases such as this, knowing the meaning of a

[1] While it is often the case that syntactic gender correlates with real-world gender for animate nouns, inanimate nouns (which comprise the majority of the nouns in a language) do not enjoy such a correlation. This matter is further complicated by the fact that in some languages morphological gender differs from syntactic and semantic gender, as in the case with Russian, which is discussed below.

particular word cannot help the child determine what gender it is, and, hence, how it behaves syntactically. It would be ridiculous to postulate an innate phonological mapping mechanism, since languages differ greatly in phonological distinctions among genders. Of course, some languages do have nouns that overlap with semantic gender, but, as noted before, those nouns are a minority among nouns of those languages.

Pinker explains gender learning by resorting to 1) an ordering of learning processes and 2) a powerful learning mechanism which has the ability to learn classes that are completely arbitrary. In Pinker's view, major category learning (noun and verb) occurs first. After the initial category is learned, learners proceed to learning affixes, attempting to assign each affix a meaning, storing each stem-affix combination in memory. When the learner is confronted by two affixes with the same meaning, she then creates sub-categories (such as masculine or feminine), which is really just a list of words that share the same affix. Those words can be arbitrary in that they are not necessarily related to each other in any way other than that they share an affix. Braine (1987) argues that this would cause words to be segmented incorrectly and would thus cause learners to overgeneralize the categories, since the model has the ability to group together arbitrary content words.

There is evidence to suggest, however, that genders are often not completely arbitrary. Karmiloff-Smith (1976) describes a study of children learning French, which has a complex set of phonological cues to gender which is sometimes overridden by syntactic and semantic irregularities. Karmiloff-Smith's work clearly showed that young children rely on the phonological cues for determining a noun's gender, but that older children rely more on syntactic and semantic cues for determining an unknown noun's gender.

The notion that children access multiple cues is supported by Braine (1987), who reviewed a series of experiments which show that humans are quite sensitive to certain types of information in linguistic input; more specifically, humans are sensitive to the distribution of markers which serve as clues to category membership. This si referred to as "distributional information." The work Braine reviewed used miniature artificial grammars: that is, combinations of semantically-empty nonsense "word" strings. By manipulating the input and the training, it was determined that it was easy for people to learn a) the position of words (first, last) in phrases (Braine, 1963; Smith, 1966), b) the position of words relative to a "function" marker (Braine, 1966, 1971) and c) the co-occurrence of certain morphemes (Braine, 1965).[2]

For example, in Braine (1965), participants were exposed to nonsense strings of the following type:

ax_1b, ax_2b, ax_3b, ax_4b, ax_5b, ax_6b, ax_7b, ax_8b, ax_9b, $ax_{10}b$, $ax_{11}b$, $ax_{12}b$,

px_1q, px_2q, px_3q, px_4q, px_5q, px_6q, $px_{13}q$, $px_{14}q$, $px_{15}q$, $px_{16}q$, $px_{17}q$, $px_{18}q$

---

[2] For a detailed description of these experiments, see Braine (1987). What follows here is a review of the most recent research in artificial grammars because it is most directly relevant to the experiments described here.

where a, b, p, q and $x_{1\text{-}18}$ are nonsense words (a=ane, b=kivil, p=foo, q=slet). They were then administered a series of recall and recognition tests. On the recall tests, the participants mainly listed items they had heard in the training, although they did list some items they had not heard in the training. These were all of the a...b or p...q type. They did not list any items of the p...b or a...q type. The recognition tests showed the same kind of behavior, but more clearly (due to more data points). Participants were able to a) recognize strings that they had heard in the training and b) "recognize" strings that they had not heard before but that conformed to the pattern they had been presented (such as px_8q, for example). This suggests that although the participants had not heard those items, they learned that they adhered to the pattern being presented. In other words, participants were able to generalize the pattern to new material.

This body of work might lead one to believe that humans should be able to categorize on the basis of distributional information. However, one aspect of Braine (1965) brings that idea into question. Although participants generalized to material that they had not heard before, that generalization only went so far. Participants recognized strings of the type aXq or pXb, but only when 'a' was never varied and 'b' never varied. What this suggests is that participants did not perceive that a and p are *categories* of words that can appear first and that q and b are *categories* of words that can appear last. Participants, instead, simply inferred that a and p can occur first and b and q can occur last and that b is dependent on a and q is dependent upon p. Thus, participants were simply sensitive to the co-occurrence of certain morphemes.

This kind of category distinction, however, is crucial in language. Braine (1987) points out that this is exactly the pattern that humans must acquire in order to be able to carry out gender agreement in determiners and nouns. Braine considers the following problem space to be typical of language:

			masculine
Determiner (n_1)	Noun (m_1)	Adj	Gender affix (n_1)
			feminine
Determiner (q_1)	Noun (p_1)	Adj	Gender affix (q_1)

Here, a speaker needs to know which gender a noun belongs to in order to know which determiner to use. Similarly, it is necessary to know the gender of a noun in order to know which ending to attach to the noun. Braine claims (and many would agree) that this is a very common pattern of categories found in language.

Thus, Smith attempted to show that Braine's (1965) experiment would work with categories, not just morphemes (Smith, 1969). Smith exposed participants to 12 bigrams (two-letter combinations) for one minute. Then, the participants were asked to recall which bigrams they had seen. This process was repeated for 12 trials. The letters in the bigrams were divided into four classes: M, N, P, Q. Then the letters were combined such that Class M letters always preceded Class N letters and Class P letters always preceded Class Q letters.

Upon examining the results of the recall trials, Smith discovered that the participants had easily learned that M and P must come first (that is, participants supplied bigrams where M and P always come first), and that N and Q must come last, participants were not able to learn that Class M always precedes Class N and that Class P always precedes Class Q (that is, participants supplied bigrams of the type MQ and PN). Thus, like Braine (1965), Smith showed that participants had difficulty learning these types of co-occurrence restrictions. The participants were overgeneralizing the categories such that, for example, Class M "must come first," when the actual algorithm was Class M "must come first and precede Class N."

Braine (1987), re-examining Smith (1969), hypothesized that it was extremely difficult for people to generalize into classes from distributional information alone. That is, if a subject is given input from a grammar where S--> M + N and S--> P + Q, she will correctly intuit that M and P must come first in the sentence, but she will be unable to further generalize that MQ and PN are ungrammatical. He calls this the MN/PQ problem.

It appears that humans can make use of distributional information only when it comes in conjunction with other types of information. Braine (1987) performed a series of experiments in which an MN/PQ-type miniature grammar was created to simulate real-world gender classes. In addition to the auditory input of the "words" in the grammar, participants also viewed pictures in conjunction with the words. Half of the pictures in each class (that is, half of the "feminine" words and half of the "masculine" words) clearly belonged to a real-world class, i.e. men and women. The other half of the pictures were simply of random inanimate objects which did not form any kind of real-world class. Thus, 25% of the pictures were of women, 25% of the pictures were of men and 50% of the pictures were of neither.

Additionally, the "words" were paired with "numbers," also nonsense words, that were to indicate whether there were one, two, or more of the item present in the pictures. The number words had to agree with the "gender" of the nouns. Thus, there were two words for 'one,' two words for 'two,' etc. One of the words for 'one' was used for the "feminine" nouns, one was used for the "masculine" nouns and the same was true of the words for 'two' and 'more than two.'

Participants were trained on these two-word strings and then tested on both strings they had studied and strings they had not studied. Results showed that participants were correctly able to generalize to cases that they had not studied, even when the unstudied items contained a picture of an inanimate object and not a picture of a person. This is significant because, presumably, participants were using the semantic category of gender to categorize the items. This study showed that not only were the participants able to use gender to categorize correctly, but that they were also able to move beyond gender as the 'marker' of the gender. That is, once the participants learned there were two categories and how those categories worked, the participants did not apply their generalizations only to nouns of the correct gender; they were able to form abstract categories.

Thus, participants were able to correctly classify M and P as being members of different categories, with MQ and PN being ungrammatical strings in the artificial language. Braine speculated that in order to solve the MN/PQ problem it was necessary to give the participants "a way into the system." What he means by that is that a way to solve the MN/PQ problem is to take a three-step approach to the problem. First, assume that many members of N have a feature in common (I will refer to this as *f*). Humans are good at grouping together words that have common features, so these words will form the basis for category N. Next, humans should notice the co-occurrence between M and the all the words that share feature *f*. Lastly, it is necessary for humans to make the leap that since M co-occurs with all the words that share feature *f* and also co-occurs with other words, that the other words must be like the words that share feature *f*. In the case of this experiment, the semantic information served as feature *f* and allowed the critical first step of these three steps to take place.

Brooks, Braine et al. (1993) explored the possibility that non-semantic information could serve the same role. Brooks created two miniature artificial languages composed of 30 words for objects, two sets of three locative suffixes and one agent (the subject of all the sentences). Participants then underwent extensive training in one of the two languages: four to five training sessions lasting approximately 45 minutes each; half the participants were trained on one language, the other half were trained on the other language. In these training sessions, trainers acted out scenarios with toys, using the artificial words for the actions they were performing.

The difference between the two languages was that, in one language (what Brooks refers to as the Experimental Language), a large proportion of the words for the objects contained a common sequence: -oik or -oo. These sequences constitute feature *f*. In the other language, (the Control Language), there were no common sequences. A portion of Brook's stimuli appear below:

Experimental Language (doubly-marked; oik and oo followed by ending)

gender 1

blotoikeef	hoikeef	wernoikeef	bobelleef	zozeeef
blotoikrog	hoikrog	wernoikrog	bobellrog	zozerog
blotoikast	hoikast	wernoikast	bobellast	zozeast

gender 2

getoofoo	elloofoo	snoofoo	kermfoo	brolfoo
getooilg	ellooilg	snooilg	kermilg	brolilg
getootev	ellootev	snootev	kermtev	broltev

<u>Control Language</u> (singly-marked; ending only)

gender 1

blotoikeef	hookeef	wernokeef	bobelleef	zozeeef
blotoikrog	hookrog	wernookrog	bobellrog	zozerog
blotoikast	hookast	wernookast	bobellast	zozeast

gender 2

getelfoo	elloofoo	snoikfoo	kermfoo	brolfoo
getelilg	ellooilg	snoikilg	kermilg	brolilg
geteltev	ellootev	snoiktev	kermtev	broltev

Thus, in the Experimental Language, the sequences -oik and -oo reliably occur with the two genders, while in the Control Language they do not. The segments -oik and -oo do occur in the Control Language, but are not distributed in a way that reliably coincides with the difference in genders.

Brooks' experiment contained additional complexities. First of all, not all of the words occurred during training with the same frequency. Brooks divided her stimuli into high-, medium- and low-frequency items. Additionally, not all the word-suffix combinations were given during the training. These items were specifically not supplied to the participants so that it could be determined whether or not the participants were able to generalize to new material.

Perhaps because the experiment was quite complex, the results are also quite complex. However, there are several results which are relevant to the current investigation. First, participants performed significantly better on the Experimental Language than they did on the Control Language in that they more often recalled items on which they had been trained when they were trained in the Experimental Language. Also, participants trained on the Experimental Language were able to "recall" previously-unheard forms, as long as those forms contained the -oik or -oo segment (that is, they were doubly-marked for gender). There is the suggestion that participants were sometimes able to generalize to previously-unheard forms that did not contain the -oik or -oo marking, but these results are only weakly significant. There was a significant effect of language type (Experimental vs. Control) on these forms ($t(30)=1.77$, $p<.05$, one-tailed), but the effect was just significant and the t-test was not two-tailed.

The significant difference between performance on the Experimental Language and the Control Language suggests that the -oik and -oo segments had an effect. However, it is difficult to say exactly what that effect was. While the participants were able to generalize to new forms which contained -oik and -oo, they had substantial more difficulty in generalizing to new forms that did not contain -oik and -oo (if indeed they were able to generalize to those forms at all). In essence, then, the participants were learning the co-occurrence of morphemes (-oik and -oo plus the suffixes), much in the way that they learned them in Braine (1965). While this work appears to partially solve the MN/PQ problem, the fact that the results were so weak in the critical areas suggests that a true MN/PQ effect remained elusive.

Frigo and McDonald (1998) continued to explore category learning. In their experiments, they told the participants that they would be learning two kinds of greetings: two of them were to be used in the evening and two of them were to be used during the day. The participants were also told that there were two groups of people and that one set of greetings was used for one group and the other set of greetings was used for the other group. Thus, the task for the participants was to correctly categorize which greetings went with which groups of people, a task somewhat analogous to the classification of words into genders. The groups of people were distinguished with phonological markers, whereby 60% of the members of one group shared a morpheme-like sequence of sounds, making this study similar to Brooks (1993).

Like Brooks (1993), Frigo and McDonald found that participants were correctly able to generalize to unstudied items when those unstudied items contained the phonological marker. However, when the unstudied items did not contain the phonological marker, participants performed no better than chance at associating the proper greeting with the person. This again suggests that the MN/PQ problem was not solved.

Frigo and McDonald went on to examine learning on the basis of salience of markers and their redundancy. Participants were not able to generalize to unstudied, unmarked forms unless the markers were salient (at least a syllable in length) and redundant (appeared at the beginning and the end of the word). What is most striking about these results is that participants knew in advance how many categories there were. They were told that there were two kinds of greetings for two groups of people, therefore, given distributional information that correlates with those number of categories, it should have been easy for participants to form the categories and generalize to new cases. However, this was not the case. Despite the information about the type and number of categories, participants were simply not able to learn categories.

Another interesting examination of gender learning in a second language (Russian) was conducted by Kempe and Brooks (2001). Kempe and Brooks suspected that in language directed at children, there was an inordinately high percentage of nouns with diminutive suffixes. This suspicion was confirmed through an analysis of child-directed speech in the Russian portion of the CHILDES database. According to that analysis, between 35% and 40% of the nouns in child-directed speech are diminutives. This differs from adult-directed speech where it was estimated that 2.7% of nouns are diminutives.

Kempe and Brooks hypothesized that these diminutives aid in language learning by providing additional correlational information. Thus, they conducted an experiment whereby participants were trained on the same set of nouns, but one group was given the diminutive versions of the nouns and the other group was given the base forms of the nouns. The participants underwent four training sessions, each lasting approximately 30 minutes. In each session they were trained on two-word strings of Russian words. The strings consisted of a color word (either masculine or feminine) and a noun. The participants viewed pictures of the nouns during the training. As stated before, half of the participants were given nouns with diminutive suffixes and half were given bare noun without diminutive suffixes.

Comparing the performances of both groups, the group trained on the diminutives clearly outperformed the group trained on the non-diminutives. However, in the crucial condition, the condition in which participants were tested on their ability to generalize to new words, there was weak evidence that the participants trained on diminutives were able to perform above chance. This finding, then, is analogous to Brooks (1993), which also had a very weak effect in the relevant condition. Again, it appears that solving the MN/PQ problem is difficult.

However, Gerken, Gómez, and Nurmsoo (1999) have discovered that participants are able to solve the MN/PQ problem (forming categories) on the basis of phonological information alone, provided that the phonological information converges in a reliable way. Using the "paradigm completion" method, participants were trained on non-word "sentences". The sentences are composed of "function" word and "content" word pairs from the following set, where W, X, Y and Z are content words and a_1 and a_2 are function words:

a_1 W	a_1 X	a_1 Y	a_1 Z
a_2 W	a_2 X	a_2 Y	

Participants were given instances of all the content and function word combinations, except for the combination in the last cell of the set. These word combinations were interspersed with word combinations from the following set:

b_1 L	b_1 M	b_1 N	b_1 P
b_2 L	b_2 M	b_2 N	

Crucially, all the (a) function words were open syllables with lax vowels; the corresponding content words (W, X, Y, and Z) were closed syllables which also contained lax vowels. The (b) function words, on the other hand, were open syllables with tense vowels and their corresponding content words were closed syllables with tense vowels. For example, a sequence in the first category might be 'pæ l″ g' while a sequence in the second category might be 'so púg.'

During training, participants were given instances of all the grammatically possible function-content-word combinations, except for the last cell in each set. At test, however, participants were able to successfully determine that the items occupying the last cells (a_2Z and b_2P) were grammatical, and that items such as b_2Z and a_2P (formed by combining the wrong function word with a content word) are ungrammatical. This shows that the participants had correctly deduced that Z belongs to the same category as W, X, and Y and that P belongs to the same category as L, M, and N. Thus, it appears that category learning is possible without semantic or categorical information; phonological cues alone appear to be sufficient for category induction to take place. Plus, it should be noted that participants must have based their learning on abstract phonological similarity; the actual vowels in the function words never matched the vowels in the content words. The two matched only in terms of vowel quality (tense/lax).

Interestingly, Gerken, et al. (1999) noticed that the subjects appeared to be unaware that they were learning anything. This finding is contrary to other work on implicit learning in the adult second language acquisition literature. There is much debate in that literature about the nature of implicit learning in second language acquisition (Gregg, 1984; Krashen, 1981; Schmidt, 1990), but only recently have researchers conducted controlled studies. The consensus from those studies is that implicit learning is necessary for language to be used fluently but that it is difficult to obtain evidence of implicit learning actually occurring (Alanen, 1992; Doughty, 1991; Ellis, 1993; Robinson, 1997).

One second language learning study examines precisely the problem at issue here: the implicit learning of syntactic categories (DeKeyser, 1995). DeKeyser created a miniature artificial language that consisted of 98 lexical items, most of which were verbs and nouns. Both the verbs and nouns were inflected. Thus, one of the tasks with which the participants were confronted was the division of the verbs and nouns into categories and sub-categories. The participants each participated in 20 learning sessions (of about 25 minutes in duration). During the learning sessions, the participants were shown pictures on a computer screen which were described by sentences in the artificial language appearing in written form below the pictures. The participants were divided into two groups. One group of participants received occasional explicit instruction in the grammar rules of the artificial language. The other group received no such instruction. After the training sessions, the participants were tested on their knowledge of the artificial language by means of a task which required them to supply captions for both pictures they had seen in the training and new pictures. Participants who had received instruction in the grammatical rules performed well on the test. Subject who had not received instruction performed at chance on every structure that was to be learned (including the categories noun and verb and their subcategories). DeKeyser states the results unequivocally, "no knowledge of rules was acquired implicitly and...no implicit knowledge was available...for categorical rules" (pg. 397).

The Gerken, et. al. study, however, suggests that implicit learning in second language acquisition is possible when there is a sufficient amount of converging information. Given the success of the artificial grammar paradigm-completion method for investigating category learning, it makes sense to attempt to apply that framework to a natural language situation, where factors such as phonological regularity are the cues that real language learners must discern. Russian was selected as the research language for this dissertation because of the richness of its morphological system. Russian nouns are declined into six cases, singular and plural, most with a distinct morphological marking. Thus, it might be the case that Russian nouns are marked with enough convergent information for learners of Russian to be able to solve the MN/PQ problem.

According to Corbett (1991), the Russian gender system is morphological, as opposed to phonological or semantic. This means that in order to determine the gender of a particular noun, it is necessary to compare two declensions of the same noun (nominative singular and genitive singular, for example). A phonological system, on the other hand, would permit gender determination on the basis of one form of the noun only. A semantic system would require accessing the meaning of the noun in order to determine its gender. However, Corbett notes that no gender systems are purely morphological, Russian included. The exceptions to the morphological rules which determine gender in Russian are all easily resolved with semantics. For example, the noun 'dedushka' has a feminine declension pattern, but takes masculine modifiers. This is because 'dedushka' means *grandfather*, a semantically masculine noun.

It might be the case that Russian children employ a combination of distributional morphological information and semantic information much in the same way as was demonstrated by Braine (1987), described above. However, empirical investigation suggests that this is unlikely. Studies of Russian child language acquisition indicate that children are late in acquiring the correct gender of exceptional semantically-determined forms (such as 'dedushka,' above) (Popova, 1958). This implies that in the early stages of acquisition children are relying on distribution of morphological information alone.

In fact, Gvozdev (1969) hypothesized, from his nine-year detailed recordings of his son's speech, that grammatical categories are acquired before morphological details. Thus, while a child may not be marking gender correctly (that is, the child uses gender markings that do not coincide with the real-world sex of a noun), she nonetheless has a gender system and correctly identifies the number of categories in that gender system (in the case of Russian, masculine, feminine and neuter), as evidenced by the fact that young children use the correct number of gender morphemes (three), even though they may use them on the wrong words.

Slobin (1966) claims that "grammatical gender is responsible for what is perhaps the most difficult and drawn-out linguistic learning of the Russian-speaking child, although it is almost unequivocally marked phonetically" (pg.142), citing the difficulty

Russian children have in associating morphological endings with the correct semantic gender of nouns. Thus, in Russian, semantic cues provide contradictory information for the child learner. But it is important to note that the difficulty in learning the system comes from the inability to match up grammatical gender with the few nouns that have actual gender. This is an entirely different problem than discovering that there are distinct categories that are handled differently syntactically. The focus of the present work is the latter.

Russian thus provides an ideal test case for the paradigm-completion method since it is hypothesized that Russian children must determine gender primarily through morphological information. Additionally, the paradigm-completion method allows a more fine-grained investigation of implicit learning in second language acquisition. For example, if Braine is correct in asserting that the MN/PQ problem is insoluble without additional information (a 'way into the system'), it could be the case that previous implicit learning studies failed because the input did not provide the kind of information that is necessary for category learning to take place.

Thus, the experiments to be discussed here have four aims:

1) to determine whether artificial grammar test paradigms and results generalize to natural language. Since artificial grammar experiments use grammars which are *artificial*, it might be the case that such controlled situations do not exist in natural language.

2) to determine if it is possible to categorize gender on the basis of morphophonological information

3) to determine whether adults are capable of making this categorization

4) to probe the possibilities of implicit learning in second language acquisition/teaching.

CHAPTER III. UNSUCCESSFUL ATTEMPTS TO FIND DISTRIBUTIONAL LEARNING

Experiment 1.

Method and Results

The experimental method and procedure closely follows that of Gerken, Gómez and Nurmsoo (1999), described above, and employs a within-participants design.

Participants. Eleven undergraduate students participated in this experiment for course credit. None had any previous exposure to Russian or any other Slavic language.

Materials. Two sets of Russian words were created. One set was composed of feminine words in the dative and accusative cases; the other was comprised of masculine words in the genitive and instrumental. These cases were chosen because there is no overlap in the endings; each ending is unique (in this experiment). One cell in each set is empty. The item that would occupy that cell, were it filled, was labeled as a "critical item." Thus, the participants heard minimal pairs of words with alternating endings except for one pair-less item.

The two sets represent the grammatical set of Russian words.

grammatical feminine set

devushke	dache	knige	korove
devushku	dachu	knigu	

grammatical masculine set

dekana	mal'chika	vora	brata
dekanom	mal'chikom	vorom	

Next, two sets were created such that masculine endings were added to feminine words and vice versa. These sets represent the ungrammatical set of Russian words.[3]

ungrammatical feminine set

devushka	dacha	kniga	korova
devushkom	dachom	knigom	korovom

ungrammatical masculine set

dekane	mal'chike	vore	brate
dekanu	mal'chiku	voru	bratu

All of the items were recorded into a Macintosh computer using SoundEdit 16 v. 2.0. A series of four sets of training stimuli was created. Only the grammatical items were included in the training sessions. The grammatical items were ordered pseudo-randomly[4] with two seconds of silence in between each item. Each token of each item was the same in all the training sessions. So, for example, the pronunciation of 'mal'chika' was exactly the same in all four training sessions. This was done to eliminate variability in the input.[5] The critical items were not (by definition) included in the

[3] Most of the items in the ungrammatical sets are actually grammatical words of Russian; that is, they are in legal morphological combinations. However, given the "grammatical" input that the participants receive, there is no way for them to know that they are actually grammatical.

[4] The randomness was manipulated such that two content words never followed each other. For example, mal'chika was never followed by mal'chikom.

[5] In a real language-learning situation, the tokens would never be identical. However, it was felt that in order to investigate the main questions of this line to research it was necessary to limit the amount of variability in the training and the test. It is an open question whether variable tokens would produce the same results as those obtained in the research presented here.

training session. The four training sessions lasted a total of approximately four minutes. Each training session contained fourteen items.

Next, a post-training test was created. The test consisted of a pseudo-random auditory list of all the grammatical and ungrammatical items. Included in this list were the two critical items that the participants had never before heard. Note that they had also not heard any of the ungrammatical items before; thus, the crucial distinction at test was whether the participants could differentiate between the critical items and the ungrammatical items.

To increase the number of data points available for analysis, the test contained two instances of the critical items and their ungrammatical counterparts. Thus, the entire test contained 36 items: 14 grammatical items on which the participants were trained; 2 (x2) critical grammatical items on which the participants were not trained; 14 new (ungrammatical) items on which the participants were not trained; and 2 (x2) critical ungrammatical items on which the participants were not trained.

Procedure. Participants were seated in front of a tape-recorder which was equipped with a foot pedal. After a brief training to familiarize them with the apparatus, participants were instructed to listen to the training lists described above. After hearing each word, they were to stop the tape via the foot pedal, say the word aloud, and then press the foot pedal again to resume listening to the taped material. The participants were instructed to pronounce the Russian words as best they could, but not to worry too much about their pronunciation.

After the training sessions, the participants were asked to take the test described above. On an answer sheet, the participants were to circle "+" if they considered the item to be "a grammatically correct word of Russian" and they were to circle "-" if they considered the item to be "grammatically incorrect in Russian." If participants asked for more explanation, they were told, "You will now hear a series of words, some of which you heard in the training and some of which you did not. Please decide whether or not the words are well-formed words of Russian."

Scoring. Each "+" response was assigned a value of 1. Each "-" response was assigned a value of 0. The critical comparison here is the likelihood that participants will answer 'yes' on the grammatically correct items as compared to the likelihood that they will answer 'yes' on the grammatically incorrect items.

There are two relevant predictions. First, participants should be able to distinguish between regular (non-critical) grammatical and ungrammatical words; they heard only grammatical words during training, so the ungrammatical words should be unfamiliar to the participants. Thus, we would expect a large difference in means for the regular grammatical and regular ungrammatical items. Second, if participants are able to form categories based on the training, then the critical items should have significantly higher means than both the regular ungrammatical items and the ungrammatical critical items. In sum, if participants are able to categorize, we would expect high means for both grammatical conditions (regular and critical) and low means for both ungrammatical conditions (regular and critical).

Results. Mean values for each condition are given in Table 1.1 below, along with idealized scores.

Table 1.1

Mean Proportions of "Grammatical" responses for all Word Types for Experiment 1 (N=11) and idealized scores

Idealized Predicted Scores		based on		
Word Type	**Mean Score**	**Error**	**familiarity**	**learning**
GR (grammatical/regular)	.8766	.0321	1	1
UR (ungrammatical/regular)	.4610	.0612	0	0
GC (grammatical critical)	.4545	.0880	0	1
UC (ungrammatical critical)	.4773	.0920	0	0

A 2x2 Analysis of Variance was performed. There was a significant effect of grammaticality ($F(1,10) = 14.562$, $p < 0.003$) and of item type (regular vs. critical) ($F(1,10) = 7.651$, $p < 0.020$) and a significant interaction ($F(1,10) = 21.910$, $p < .001$). The nature of the interaction was that GR items were correct more frequently than any of the other three item types. Note that the mean scores for Grammatical Critical and Ungrammatical Critical are virtually identical, and the small difference goes in the opposite direction of the learning prediction.

Discussion

Participants in the study appeared to remember the training items, but failed to generalize to new grammatical items. It was hypothesized that participants in Experiment 1 were simply not given enough training. Thus, in the second experiment additional training sessions were included.

Experiment 2.

Method and Results

Participants. Seven undergraduate students participated in the experiment for course credit.

Materials. Same as in Experiment 1, except that two extra training sessions were created. These resembled the previous training sessions. Total training time was approximately seven minutes.

Procedure. Same as above.

Scoring. Same as above.

Results. Mean values for each condition are given in Table 2.1 below. As in Experiment 1, participants were able to distinguish grammatical from ungrammatical items (although not as well as in Expt. 1) but were not able to distinguish grammatical critical items from ungrammatical items.

Table 2.1

Mean Proportions of "Grammatical" responses for all Word Types for Experiment 2 (N=7) and idealized scores

Word Type	**Mean Score**	**Error**	Idealized Predicted Scores based on **familiarity**	**learning**
GR (grammatical/regular)	.8265	.0377	1	1
UR (ungrammatical/regular)	.6327	.0550	0	0
GC (grammatical critical)	.3214	.0899	0	1
UC (ungrammatical critical)	.3929	.0922	0	0

A 2x2 Analysis of Variance showed that there was a significant effect of grammaticality $F(1,6)= 5.053$, $p<0.066$) and of item type ($F(1,6)= 46.746$, $p<0.001$). The interaction between item type and grammaticality ($F(1,6)= 4.442$, $p>0.080$) was nearly significant. The difference between means for the GR condition and the UR condition is significant ($F(1,6)= 8.734$, $p< .025$), while the difference between means between the GC condition and the UC condition was not ($F(1,6)=1.000$, $p>0.356$).

Discussion

Comparing the means of the results of Experiment 1 with Experiment 2, it is clear that there is a difference between the UR condition and both the GC and UC conditions. In Experiment 1, all the means for the UR, GC and UC conditions were around chance. However, in Experiment 2, the mean for the UR condition is much higher than the means for the GC and UC conditions, which are in fact lower than chance. What the latter result suggests is that the increase in training, while still not enabling the participants to form categories, had the effect of polarizing the grammaticality ratings. This could be because the participants were simply less familiar with the critical items; the content word corresponding to the critical item was used only once per training session. Therefore, the participants only heard that word six times during the entire training. In contrast, the non-critical words each appeared in two variants so they are heard twice as often in the training session and are therefore more familiar. Hence the difference between the experimental results could be due to the fact that the regular items are simply more familiar to the participants and that difference in familiarity is exaggerated with increased training. Additionally, observations conducted while participants were being trained suggested that the participants were having difficulty perceiving the difference between the -e ending (from the first set) and the -a ending of the second set.[6] If participants were misperceiving the items, then it would be impossible for them to correctly categorize. To deal with this problem, a new set was created for Experiment 3.

Experiment 3.

<u>Method and Results</u>

Participants. Ten undergraduate students participated in the experiment for course credit.

---

[6] The observations consisted of listening to the participants' productions. Although production is not necessarily an accurate gauge of a participant's ability to perceive (Green & Zampini, 1998), there are independent theoretical reasons to suspect that the participants were mis-perceiving the endings. Both endings are unstressed in Russian and therefore both of the vowels are reduced. Although there is still a

Materials. Same as in Experiment 2, except that the materials were re-designed. Instead of using the -e suffix in the first set, the instrumental suffix -oj was used. It was hypothesized that -oj would be more effectively differentiated from -a. Thus, the new experimental sets were as follows[7]:

grammatical feminine set

devushkoj	ruchkoj	knigoj	korovoj
devushku	ruchku	knigu	

grammatical masculine set

dekana	mal'chika	vora	brata
dekanom	mal'chikom	vorom	

Ungrammatical items were constructed so that feminine endings appear incorrectly on the items in the second set and vice versa. The ungrammatical sets were the following:

ungrammatical feminine set

devushka	ruchka	kniga	korova
devomshkom	ruchkom	knigom	

ungrammatical masculine set

dekanoj	mal'chikoj	voroj	bratoj
dekanu	mal'chiku	voru	

Additionally, because of the differences among the UR, GC and UC conditions between Experiment 1 and Experiment 2, the content word corresponding to the critical item in the set was now repeated during each training session to increase the familiarity of

---

phonetic difference between the two sounds, it is possible that this difference is difficult for native-English speakers to perceive since, in English, the two sounds are allophones of the phoneme /a/.

[7] Because 'oj' is not a possible ending for the lexical item dacha, a new word, 'ruchka,' was substituted.

the content word of the critical item. Thus, a sample training session would now be as follows (content words corresponding to critical items are boldfaced):

vorom, mal'chika, devushkoj, **brata**, ruchku, dekanom, **brata**, devushku, knigu, dekana, **korovoj**, mal'chikom, vora, knigoj, dekana, **korovoj**

Procedure. Same as above.

Scoring. Same as above.

Results. Mean values are given in Table 3.1.

Table 3.1

Mean Proportions of "Grammatical" responses for all Word Types for Experiment 3 (N=10) and idealized predicted scores

Word Type	Mean Score	Error	Idealized Predicted Scores based on familiarity	Idealized Predicted Scores based on learning
GR (grammatical/regular)	.8429	.0581	1	1
UR (ungrammatical/regular)	.4286	.0522	0	0
GC (grammatical critical)	.2000	.0816	0	1
UC (ungrammatical critical)	.3750	.1070	0	0

A 2x2 Analysis of Variance was performed which showed no significant effect of grammaticality ($F(1,9)= 3.258$, $p> 0.105$), but did show a significant effect of item type ($F(1,9)=25.568$, $p< 0.001$). There was also a significant interaction between item type and grammaticality ($F(1,9)= 22.387$, $p< 0.001$). The interaction shows that item type was dependent on grammaticality, which would not be the case if the participants were learning the categories. Pairwise comparisons showed there to be a significant difference between the GR and UR conditions ($F(1,9)= 21.084$, $p<0.001$) but no significant difference between the GC and UC conditions ($F(1,9)= 3.645$, $p>0.089$). The latter difference went in the opposite direction predicted if the participants had been learning.

Discussion

It is clear that in its present incarnation, the experimental paradigm described above is unable to induce category learning in participants. At this point one might suspect that category learning on the basis of distributional information alone is simply not possible. However, recall that Gerken, et al. (1999) were able to induce category learning on the basis of distributional evidence alone, provided there was enough convergent information in the input. Thus, a more likely explanation for the failure of the above experiments to induce category learning is that the set of training words lacked the convergent information necessary for learning. We explore this possibility in the studies in the next chapter.

CHAPTER IV. EXPERIMENTS WHICH SHOWED DISTRIBUTIONAL LEARNING

It is hypothesized that the experimental design used in the experiments so far results in a set of information which is unlearnable. Recall Braine's (1987) MN/PQ problem, where participants are correctly able to intuit that M and P must come first in word strings, but are unable to further generalize that MQ and PN are ungrammatical. This is the same situation we see here: although participants can distinguish words they have heard before from ones they haven't heard before, they are unable to learn which endings are allowed to appear on which words.

Braine (1987) has theorized (and Gerken, et al. (1999) have shown) that the solution to the MN/PQ problem lies in increasing the amount of convergent information that participants are given during training. However, a crucial difference between the previous studies (Braine and Gerken, et al.) and the current work is that the earlier studies employed artificial languages and the present experimental paradigm uses a natural language. Therefore it is necessary to examine the natural language-learning environment of Russian children to determine what kind of regularities might exist naturally in the input they receive. Whatever regularities exist might be employed in the acquisition process.

Braine has termed the additional distributional information required to solve the MN/PQ problem as providing "the way into the system." What it is necessary to find, then, is the Russian child's "way into the system" that enables category learning of gender to take place. As described above, that additional correlational information is unlikely to be semantics, since semantic information correlates only loosely with gender (masculine words such as *papa* 'dad' and male names such as *Misha* decline like feminine nouns, even though they are syntactically and semantically masculine). However, it is possible that additional morphological information might be enough to allow learners to intuit the gender sub-categories.

Morphological markings are abundant in Russian, with multiple derivational affixes frequently combining with lexical items prior to inflection. So, while we have been examining a morphological instance of inflection in this experiment, it is possible that additional derivational markings might provide further distributional evidence for the formation of sub-categories. For example, all nouns which end in the derivational affix -*tel'* (a rough equivalent of the English suffix -er, as in teacher) are masculine. In contrast, the suffix -*k* (used to denote female persons) is always feminine. The derivational suffix always immediately precedes the inflectional ending on Russian words; this means that, in essence, a great many Russian words are actually *doubly* marked for gender. First, they acquire a gendered derivational suffix and then they acquire a gendered inflectional ending. Thus, it may be that language learners focus on the combination of derivational and inflectional suffixes to establish categories.

A way to test this idea experimentally would be to follow the design of Braine (1987). Recall that this is the experiment where Braine accompanied his auditory stimuli with pictures. Half of the feminine nouns were accompanied by pictures of women; the other half were pictures of inanimate objects. The analogous situation was true of the masculine nouns. Given this additional categorical information, the participants were correctly able to categorize. This same idea could be employed with derivational suffixes: half of the feminine nouns would share a common derivational suffix. The other half would have different suffixes and the masculine nouns would be set up in the same way. It is possible that this additional morphological information might be enough to cue the participants that there are two categories.

Experiment 4 tests the addition of derivational suffixes as the possible "way into the system" required for category learning. Of course, in the experience of the participant, the information is not truly morphological. Because neither the words nor their suffixes are connected to any semantic information, it is imprecise to call the information morphological. Rather, the information is more phonological in nature, consisting of a series of common phones. However, this information is different from the correlational phonological information used in Gerken, et al.; that study used feature harmony to supply the necessary information. This study instead uses segment co-occurrences. I will continue to refer to this information as 'morphological' in order to differentiate it from the kind of information used in Gerken, et. al.[8]

---

[8] It is not clear yet that this is a meaningful distinction to make, but it nonetheless seems prudent to differentiate these two types of informational systems at the outset of this research.

This study more closely resembles (Brooks et al., 1993), in terms of experimental stimuli. Recall that Brooks (1993) was able to achieve a weak MN/PQ effect. However, there are crucial differences between that work and the current study. First, the training in the Brooks study was significantly longer; second, Brooks used a production task to evaluate performance; third, the set of words and morphemes was larger; and fourth, the participants learned the meanings of the artificial words. The current experiment follows the pattern of Experiments 1-3, but uses a set of stimuli that resembles Brooks (1993). Additionally, confidence ratings were collected to determine how sure participants were of their responses.

Experiment 4.

Method and Results

Participants. Sixteen undergraduate students participated in this experiment for course credit.

Materials. The set of materials was similar to that of Experiments 1-3. However, in Experiment 4 half of the feminine nouns shared a common derivational suffix (-k) and half of the masculine nouns shared a common derivational suffix (-tel'). In order to increase the salience of the words which share the common derivational suffixes, the set size[9] was also increased. Thus, the complete set of grammatical items is as follows (derivational suffixes are underlined):

---

[9] Set size refers to the number of items in the training stimuli. A "larger set size" means that the number of items to be learned is larger.

grammatical feminine set

polkoj	rubashkoj	ruchkoj	vannoj	knigoj	korovoj
polku	rubashku		vannu	knigu	korovu

grammatical masculine set

uchitelya	stroitelya	zhitelya	kornya	tramvaya	pisarya
uchitelem	stroitelem		kornem	tramvayem	pisarem

The gap in the paradigm (the critical item) falls within the subset of words containing the common derivational suffix.

Procedure. Same as in Experiments 1-3, with the addition of a new task. In addition to circling a "+" for grammatical items and a "-" for ungrammatical items, participants were asked to rate how sure they were about their responses, on a scale from 1-6 (1 being very unsure and 6 being very sure). Thus the experiment contained two tasks at test: a grammaticality judgment task, and a confidence rating task. It was felt that perhaps the grammaticality judgment task was simply not sensitive enough to detect the category learning. With a larger range of scores, therefore, it is conceivable that the confidence ratings might be able to detect learning that the grammaticality judgments are too blunt to perceive.

Scoring. Same as in Experiments 1-3, except that the confidence ratings were also separately scored. The confidence ratings were scored in the following way: when a participant circled "+," the confidence rating was multiplied by +1. When a participant circled "-," the confidence rating was multiplied by -1. Thus the scale from ungrammatical to grammatical ranged from -6 to +6.

Results. Mean values for the grammaticality judgments in Experiment 4 are given in Table 4.1 and mean values for the confidence ratings are given in Table 4.2.

Table 4.1

Mean Proportion of "Grammatical" responses for all Word Types for Experiment 4 (N=16) and idealized scores

			Idealized Predicted Scores based on	
Word Type	**Mean Score**	**Error**	**familiarity**	**learning**
GR (grammatical/regular)	0.7760	.0514	1	1
UR (ungrammatical/regular)	0.3333	.0543	0	0
GC (grammatical critical)	0.7969	.0469	0	1
UC (ungrammatical critical)	0.3750	.0685	0	0

A 2x2 Analysis of Variance showed that there was a significant effect of grammaticality ($F(1,15)= 22.469$, $p< 0.001$), but not of item type ($F(1,15)= 0.547$, $p< 0.471$). Also, there was not a significant interaction between the two factors ($F(1,15)=0.068$, $p> 0.797$). Pairwise comparisons of the Grammatical Critical (GC) and the Ungrammatical Critical (UC) revealed a significant difference ($F(1,15)= 17.330$, $p<.001$).

Table 4.2

Mean Confidence Ratings for all Word Types for Experiment 4 (N=16) and idealized scores

Word Type	Mean Score	Error	Idealized Predicted Scores based on familiarity	learning
GR (grammatical/regular)	3.0625	.5379	6	6
UR (ungrammatical/regular)	-1.2344	.4542	-6	-6
GC (grammatical critical)	2.8438	.4808	-6	6
UC (ungrammatical critical)	-1.1719	.6232	-6	-6

Turning to confidence ratings, a 2x2 Analysis of Variance showed a significant effect of grammaticality ($F(1,15)=22.257$, $p<0.001$), but not of item type ($F(1,15)= 0.049$, $p>0.828$). There was also no interaction ($F(1,15)= 0.186$, $p>0.672$). Pairwise comparisons showed that the difference between the Grammatical Critical (GC) and the Ungrammatical Critical (UC) was significant ($F (1,15)= 18.228$, $p<.001$).

The results of Experiment 4, for both the grammaticality judgment task and the confidence rating task, clearly pattern with the idealized outcome based on learning.

Discussion

Given the success of the design of Experiment 4, it appears that category learning is possible if enough distributional information is present. Also, it appears that morphological information alone (provided there is enough of it) is able to induce category learning. However, it is important to point out that there were two changes made from Experiment 3 to Experiment 4. First, as discussed, derivational morphemes[10] were added to increase the amount of distributional information available to the learner. Second, the set size was increased to highlight the salience of that convergent information. Thus, there are several possible explanations for the results obtained in Experiment 4. It is possible that the category learning which took place had nothing, in fact, do to with the derivational morphemes. Instead, it is possible that merely increasing the set size resulted in category learning.

It is easy to see how this might be the case. Because the number of words the participants would be exposed to in an increased set size would be greater, the need for categories might also be greater. Also, it would increase the likelihood that stimulus words of the same category would be presented in sequence (thus, for example, a participant might hear *devushkoj* followed by *knigoj*, where both words have the same ending). If participants are able to categorize with the increased set size, that suggests that the MN/PQ problem is not actually insoluble, as Braine has claimed. Instead it

[10] I use the term 'morpheme' for expositional simplicity, but, since they lack meaning, these aren't true morphemes. They are simply strings of sounds/

would point to the need for a specific distribution of information. In order for the participant to form a category, it may simply entail hearing two "content" words successively.

In order to determine whether the increase in set size was the determining factor in Experiment 4, the following experiment repeated Experiment 3 but expanded the set size.

Experiment 5.

<u>Method and Results</u>

Participants. Sixteen undergraduate students participated in this experiment for course credit.

Materials. Same as in Experiment 3 except that new items were added, thus increasing the set size to the same size as Experiment 4. These items do not share common derivational suffixes.[11]

grammatical feminine set

devushkoj	ruchkoj	knigoj	lapoj	vannoj	korovoj
devushku	ruchku	knigu	lapu	vannu	

grammatical masculine set

dekana	mal'chika	vora	shkafa	plakata	brata
dekanom	mal'chikom	vorom	shkafom	plakatom	

[11] Actually, two of the feminine items do share a common derivational suffix: devushkoj and ruchkoj. However, these two items were also present in Experiment 3, where category learning did not take place. In order to properly contrast between Experiments 3 and 4, it was necessary to change as few of the variables as possible between Experiments 3 and 5—such that the only change was the increase in set size.

Procedure. Same as Experiment 3, with the addition of the confidence ratings, as described in Experiment 4.

Scoring. Same as Experiment 3, with the addition of the confidence ratings, as described in Experiment 4.

Results. Mean values for the grammaticality judgments in Experiment 5 are given in Table 5.1 and mean values for the confidence ratings are given in Table 5.2.

Table 5.1

Mean "Grammatical" responses for all Word Types for Experiment 5 (N=16) and idealized scores

Word Type	Mean Score	Error	Idealized Predicted Scores based on familiarity	Idealized Predicted Scores based on learning
GR (grammatical/regular)	0.8281	.0279	1	1
UR (ungrammatical/regular)	0.3177	.0333	0	0
GC (grammatical critical)	0.2969	.0654	0	1
UC (ungrammatical critical)	0.3594	.0644	0	0

The mean for the Grammatical Regular items was much higher than the means for the Ungrammatical Regular, Grammatical Critical and Ungrammatical Critical items. A 2x2 Analysis of Variance showed that there was a significant effect of grammaticality ($F(1,15)=15.016$, $p< 0.001$) and of item type ($F(1,15)=34.552$, $p<0.001$). As in Experiments 1-3, there was also an interaction between type and grammaticality ($F(1,15)= 31.532$, $p< 0.001$). There was a significant difference between GR and UR ($F1,15)= 94.034$, $p<0.001$) but not between the UC and GC conditions ($F(1,15)= 0.429$, $p<0.523$). Additionally, a 2x2x2 Analysis of Variance was performed to examine the difference between this experiment and Experiment 4 (which contained the double-marking). There was a significant difference between the experiments (F(1,30)= 1211.861, p<0.001) but no interaction between experiment (double-marking vs. no double-marking) and grammaticality (grammatical vs. ungrammatical) (F(1,30)= 3.723, p<0.063). This suggests that it was indeed the double-marking which was responsible for the category learning in Experiment 4.

Table 5.2

Mean Confidence Ratings for all Word Types for Experiment 5 (N=16) and idealized scores

Word Type	Mean Score	Error	Idealized Predicted Scores based on familiarity	Idealized Predicted Scores based on learning
GR (grammatical/regular)	3.5417	0.2972	6	6
UR (ungrammatical/regular)	-1.7187	0.3433	-6	-6
GC (grammatical critical)	-2.0313	0.6907	-6	6
UC (ungrammatical critical)	-1.4063	0.6653	-6	-6

A 2x2 Analysis of Variance on confidence ratingsshowed that there was a significant effect of both grammaticality (F(1,15)= 17.989, p<0.001) and item type (F(1,15)= 27.903, p<0.001). There was also a significant interaction between the two (F(1,15)=33.353, p< 0.001). There was a significant difference between the GR and UR conditions (F(1,15)=87.006, p<0.001) but not between the GC and UC conditions (F (1,15)= 0.489, p> 0.495).

A 2x2x2 ANOVA was performed on the confidence ratings of Experiment 4 and Experiment 5. There was a main effect of experiment (F(1,30)=15.282,p<0.001) and a nearly significant interaction between experiment and grammaticality ($F(1,30)=3.145$, $p<0.086$).

Another interesting phenomenon is that in all the experiments where category learning failed to take place (Experiments 1-3 and 5), the means for the GC condition are always lower than that of other conditions. This difference is never statistically significant, but it occurs in every study. It is unclear why this should be the case. UR, GC and UC items are all 'new' to the participant since she was trained on only the GR items. Therefore, there is no immediate explanation for why the GC items should consistently score lower than the UR and UC items.

Examining the means of the UR, GC and UC conditions in experiments where category-learning did not take place could shed some light on these unanswered questions. I set aside these issues for future research.

Discussion

It appears that the additional morphological information was indeed the determining factor in the success of Experiment 4. However, there is another possibility for the success of that experiment. It is possible that the ungrammatical items (both regular and critical) simply sound ungrammatical. For example, prosody might give enough evidence that an item is ungrammatical. Given the failure of Experiments 1-3 and Experiment 5 to induce category learning, this is unlikely, since there is little reason to believe that the ungrammatical items in those experiments sound any more grammatical than the ungrammatical items in Experiment 4. However, just to be sure, participants in Experiment 6 were tested on Experiment 4's test items, but without going through the training. Thus, if there was something about the ungrammatical items that made them sound ungrammatical to the naive listener, that should become apparent in the results of Experiment 6.

Experiment 6

Method and Results

Participants. Sixteen undergraduate students participated in this experiment for course credit.

Materials. The test from Experiment 4 was the only material for this experiment.

Procedure. Participants were seated in front of the tape recorder and were instructed in its use. They were told that they would be asked to judge the acceptability of some Russian words, even though they had never studied Russian before. They were told that it was perfectly acceptable to guess in instances when they were unsure (which was expected to be most of the time). As in Experiment 4, they were also asked to rate how confidently they felt about their decisions.

Scoring. Same as in Experiment 4.

Results. Mean values for the grammaticality judgments in Experiment 6 are given in Table 6.1 and mean values for the confidence ratings are given in Table 6.2.

Table 6.1

Mean "Grammatical" responses for all Word Types for Experiment 6 (N=16) and idealized scores

Word Type	**Mean Score**	**Error**	**Predicted Score if ungrammatical items are simply obviously ungrammatical**	**Predicted score if participants are simply guessing**
GR (grammatical/regular)	0.5573	.0329	1	.5
UR (ungrammatical/regular)	0.5156	.0306	0	.5
GC (grammatical critical)	0.6406	.0644	1	.5
UC (ungrammatical critical)	0.5938	.0786	0	.5

A 2x2 ANOVA showed that there was no significant effect of grammaticality ($F(1,15)=1.088$, $p>0.313$) or of item type ($F(1,15)= 1.525$, $p>0.236$); nor was there an interaction ($F(1,15)=0.004$, $p>0.949$). Although the grammatical items scored slightly higher than the ungrammatical items, this difference was not statistically significant. Pairwise comparisons showed the difference between the GC and UC items was not significant ($F(1,15)=.458$, $p>.509$).

Additionally, a 2x2x2 ANOVA was performed to examine the difference between Experiment 4 (training) and Experiment 6 (no training). There was a significant interaction between amount of training (training vs. no training) and grammaticality (grammatical vs. ungrammatical) [$F(1,30)= 14.881$, $p<0.001$], which is to be expected if differentiating grammatical from ungrammatical were based on training.

Table 6.2

Mean Confidence Ratings for all Word Types for Experiment 6 (N=16) and idealized scores

Word Type	Mean Score	Error	Predicted Score if ungrammatical items are simply obviously ungrammatical	Predicted score if participants are simply guessing
GR (grammatical/regular)	0.6979	.2509	6	0
UR (ungrammatical/regular)	0.0885	.2042	-6	0
GC (grammatical critical)	1.2031	.4821	6	0
UC (ungrammatical critical)	0.7344	.5685	-6	0

A 2x2 ANOVA on confidence ratings showed no significant effect of grammaticality ($F(1,15)=3.251$, $p>0.092$) or of item type ($F(1,15)= 1.462$, $p<0.245$); nor was there any interaction ($F(1,15)=0.052$, $p>0.822$). The scores on the confidence ratings also pattern with a predicted score had the participants simply been guessing. Thus, the differences in means observed in Experiment 4 were not simply an artifact of the test itself, but instead reflected learning on the part of the participants.

Discussion

It would appear that converging morphological information is enough to solve Braine's MN/PQ problem. However, a closer look at Experiment 4 reveals that this is not necessarily the case. The items for Experiment 4 are repeated here:

grammatical feminine set

polkoj	rubashkoj	ruchkoj	vannoj	knigoj	korovoj
polku	rubashku		vannu	knigu	korovu

grammatical masculine set

uchitelya	stroitelya	zhitelya	kornya	tramvaya	pisarya
uchitelem	stroitelem		kornem	tramvayem	pisarem

Because the critical items in Experiment 4 were doubly marked (that is, contained both the derivational and inflectional morphology), it is possible that participants were simply picking up on the co-occurrence of the derivational and the inflectional morpheme, and not categorizing at all. That is, at test, participants might simply have learned that 'tel' occurs with 'em,' for example, but had not learned that 'em' marks a particular category and 'u' marks another. Recall that Braine (1965) showed that participants can easily learn the co-occurrence of morphemes but that learning that morphemes can co-occur does not solve the MN/PQ problem. Without solving the MN/PQ problem, the hypothesis that distributional information can contribute significantly to learning is greatly weakened.

In order to show that category learning has really taken place, the critical items must be from among those that are not doubly-marked.

Experiment 7.

Method and Results

Participants. Sixteen undergraduate students participated in this experiment for course credit.

Materials. Same as in Experiment 4, except that the singleton did not contain the double-marking of derivational plus inflectional ending:

grammatical feminine set

polkoj	rubashkoj	ruchkoj	vannoj	knigoj	korovoj
polku	rubashku	ruchku	vannu	knigu	

grammatical masculine set

uchitelya	stroitelya	zhitelya	kornya	tramvaya	pisarya
uchitelem	stroitelem	zhitelem	kornem	tramvayem	

Procedure. Same as Experiment 4.

Scoring. Same as Experiment 4.

Results. Mean values for the grammaticality judgments in Experiment 7 are given in Table 7.1 and mean values for the confidence ratings are given in Table 7.2.

Table 7.1

Mean "Grammatical" responses for all Word Types for Experiment 7 (N=16) and idealized scores

			Idealized Predicted Scores based on	
Word Type	**Mean Score**	**Error**	**familiarity**	**learning**
GR (grammatical/regular)	0.9010	0.0277	1	1
UR (ungrammatical/regular)	0.2500	0.0791	0	0
GC (grammatical critical)	0.6250	0.0456	0	1
UC (ungrammatical critical)	0.1719	0.0440	0	0

A 2x2 ANOVA showed that there was a significant effect of grammaticality ($F(1,15)= 144.794$, $p<0.001$) and of item type ($F(1,15)= 11.942$, $p<0.004$), but there was no significant interaction between the two (F(1,15)= 2.834, p> 0.113). Pairwise comparisons showed the difference between Grammatical Critical (GC) and the Ungrammatical Critical (UC) conditions to be significant (F (1,15)= 25.903, p<.001).

Table 7.2

Mean Confidence Ratings for all Word Types for Experiment 7 (N=16) and idealized scores

			Idealized Predicted Scores based on	
Word Type	Mean Score	Error	familiarity	learning
GR (grammatical/regular)	4.2656	0.3225	6	6
UR (ungrammatical/regular)	-2.0625	0.4542	-6	-6
GC (grammatical critical)	1.1875	0.6663	-6	6
UC (ungrammatical critical)	-2.7344	0.4351	-6	-6

A 2x2 ANOVA on confidence ratings showed a significant effect of grammaticality (F(1,15)= 101.589, p<0.001) and of item type (F(1,15)= 19.415,p<0.001). The analysis also showed an interaction (F=(1,15)= 4.741, p<0.046). In the pairwise comparisons, the difference between the Grammatical Critical (GC) and the Ungrammatical Critical (UC) conditions was significant (F (1,15)= 21.994, p<0.001).

Thus, the results for Experiment 7 pattern with the idealized outcome based on learning. The interaction between type and grammaticality for the confidence ratings is unexpected, but that is likely due to the fact that familiarity was also playing a role in the items.

Discussion

Because the critical items for this experiment were not ones which contained the double-marking, it cannot be the case that the participants were simply noticing the legal sequence of morphemes and generalizing off of that. Instead, participants had to learn that there were categories. This is strong evidence in favor of the hypothesis that distributional information in language aids learning.

However, the finding that morphological information alone can induce category learning may be overstated. A *post hoc* examination of the items in Experiment 4 and 7 revealed that the items contain a subtle phonological[12] cue to gender categories. The items for Experiment 7 are repeated here:

grammatical feminine set

polkoj	rubashkoj	ruchkoj	vannoj	knigoj	korovoj
polku	rubashku	ruchku	vannu	knigu	

grammatical masculine set

uchitelya	stroitelya	zhitelya	kornya	tramvaya	pisarya
uchitelem	stroitelem	zhitelelem	kornem	tramvayem	

All of the consonants which precede the inflectional ending in the grammatical feminine set are non-palatalized. In contrast, all of the consonants preceding the inflectional ending in the masculine set are palatalized. The palatalization of the last consonant affects the

[12] In keeping with the terminology used throughout this discussion, 'phonological' refers to the features of a phone.

shape of the inflectional ending in Russian, thus providing additional correlational information for the learner.

It is not clear that this additional phonological information is simply an accidental artifact of the current experimental design. While it is the case that both masculine and feminine nouns can have palatalized consonants preceding the inflectional ending, there is not complete freedom on which consonants are allowed to be palatalized in the two genders (for example, all nouns ending in -j are masculine. This includes the noun 'tramvaya/travaem' found in the experimental set).

It is beyond the scope of this paper to examine the exact distribution of palatialized and non-palatalized consonants in Russian nouns. Therefore, it is safer to say that a combination of morphology and phonology (rather than morphology alone) is responsible for the adequate distributional information in this experiment. The important point is that word meanings are not needed in order to learn categories; that account is still preserved despite the existence of a phonological cue.

However, a further examination of Experiment 7 reveals another problem with the test items which is related to the palatalization issue described above. For ungrammatical feminine items in the test (in which masculine endings were paired with feminine nouns), the final consonants were palatalized, yielding test items such as: ruchkya, vannya, polkem and knigem. The ungrammatical masculine items, however, were not de-palatalized, yielding test items such as stroitelyu, zhitelyoj, kornyu, and uchitelyoj.

Thus, the endings 'yu' and 'yoj' were new to the participants at test. It is therefore possible that participants were simply responding to the familiarity of the ending and not indicating that they had learned the categories.

To examine this possibility, another *post hoc* analysis was performed to determine whether there was a difference between the two genders and to see if the observed effect was simply a result of participants responding to the masculine items (with the unfamiliar endings) or if participants were able to learn the feminine items (which did not contain unfamiliar endings).

A 2x2x2 ANOVA was performed to examine the effects and interactions of item type (regular vs. critical), grammaticality (grammatical vs. ungrammatical) and gender (masculine vs. feminine). The analysis showed a significant effect of item type ($F(1,15)=11.942$, $p< 0.004$), and of grammaticality ($F(1,15)=144.794$, $p< 0.001$), but not of gender ($F(1,15)= 2.834$, $p>0.113$). Nor were there any significant interactions. This suggests that although the ungrammatical masculine test items contained an unfamiliar ending, participants were responding to both the feminine and masculine items in a way consistent with category learning.

Nevertheless, another experiment was performed to correct the problem with the masculine items. In this case, the masculine test items were de-palatalized, such that test items which had been stroitelyu, zhitelyoj, kornyu, and uchitelyoj were now stroitelu, zhiteloj, kornu, and uchiteloj. To make the test as contrastive as possible, the empty cell was changed in both sets such that the critical item in the feminine was now 'vannu' and the critical item in the masculine was now 'kornem.' This is because the stems of both 'vannu' and 'kornem' end in n. This creates the best test of the critical items. Therefore, the grammatical critical items were 'vannu' and 'kornem' and the ungrammatical critical items were 'vannem' and 'kornu'-- that is, a non-palatalized n in kornu. Thus, the new test contained no unfamiliar endings.

Experiment 8.

Method and Results

Participants. Sixteen undergraduate students participated in this experiment for course credit.

Materials. Same as for Experiment 7, except for the changes mentioned above. The new experimental set is as follows:

grammatical feminine set

polkoj	rubashkoj	ruchkoj	vannoj	knigoj	korovoj
polku	rubashku	ruchku		knigu	korovu

grammatical masculine set

uchitelya	stroitelya	zhitelya	kornya	tramvaya	pisarya
uchitelem	stroitelem	zhitelem		tramvayem	pisarem

Procedure. Same as Experiment 7.

Scoring. Same as Experiment 7.

Results. Mean values for the grammaticality judgments in Experiment 8 are given in Table 8.1 and mean values for the confidence ratings are given in Table 8.2.

Table 8.1

Mean Proportions of "Grammatical" responses for all Word Types for Experiment 8 (N=16) and idealized scores

Word Type	Mean Score	Error	Idealized Predicted Scores based on familiarity	learning
GR (grammatical/regular)	.9062	.0236	1	1
UR (ungrammatical/regular)	.4531	.0761	0	0
GC (grammatical critical)	.5156	.0457	0	1
UC (ungrammatical critical)	.3594	.0718	0	0

A 2x2 ANOVA showed that there was a significant effect of grammaticality ($F(1,15)= 31.810$, $p< 0.001$), and a significant effect of item type ($F(1,15)= 16.807$, $p<0.001$). Additionally, there was a significant interaction between item type and grammaticality ($F(1,15)=15.869$, $p<0.001$). Pairwise comparisons showed a significant difference between the GR and UR conditions ($F(1,15)= 84.538$, $p<0.001$) but only a marginal difference between the GC and UC conditions ($F(1,15)= 3.947$, $p<0.066$).

Table 8.2

Mean Confidence Ratings for all Word Types for Experiment 8 (N=16) and idealized scores

Word Type	Mean Score	Error	Idealized Predicted Scores based on familiarity	learning
GR (grammatical/regular)	4.5990	.2695	6	6
UR (ungrammatical/regular)	-0.4583	.4543	-6	-6
GC (grammatical critical)	0.6250	.7042	-6	6
UC (ungrammatical critical)	-1.0313	.6762	-6	-6

In the analysis of the confidence ratings, there was a significant effect of grammaticality ($F(1,15)= 37.100$, $p< 0.001$) and of item type ($F(1,15)=18.089$, $p<0.001$). There was also a significant interaction between the two ($F(1,15)= 25.730$, $p< 0.001$). Pairwise comparisons showed a significant difference between the GR and UR conditions ($F(1,15)=94.441$, $p<0.001$) and a significant difference between the GC and UC conditions ($F(1,15)= 4.886$, $p<0.043$).

Discussion

The presence of an interaction in this experiment suggests that participants might not have been learning the categories, while the pairwise comparisons suggest that they were. This is a surprising result, given that the *post hoc* analysis of Experiment 7 seemed to show that the incorrect palatalization on masculine nouns was not playing a role in the success of the experiment. There were, however, two changes made between Experiment 7 and Experiment 8: the problem with the masculine items was fixed and the empty cell in the paradigm was changed. Thus, participants had to be able to tell that "kornem" was grammatical and "vannem" was not. Perhaps those two words are too similar to each other to enable participants to make the correct decision. Therefore, the following experiment changed the empty cell back to where it had been in Experiment 7 (that is, "korovoj" and "pisarya" were singletons).

Experiment 9.

Method and Results

Participants. Sixteen undergraduate students participated in this experiment for course credit.

Materials. Same as for Experiment 7, except that the palatalization confound was removed. The experimental set is exactly the same as in Experiment 7:

grammatical feminine set

polkoj	rubashkoj	ruchkoj	vannoj	knigoj	korovoj
polku	rubashku	ruchku	vannu	knigu	

grammatical masculine set

uchitelya	stroitelya	zhitelya	kornya	tramvaya	pisarya
uchitelem	stroitelem	zhitelem	kornem	tramvayem	

Procedure. Same as Experiment 7.

Scoring. Same as Experiment 7.

Results. Mean values for the grammaticality judgments in Experiment 9 are given in Table 9.1 and mean values for the confidence ratings are given in Table 9.2.

Table 9.1

Mean "Grammatical" responses for all Word Types for Experiment 9 (N=16) and idealized scores

Word Type	**Mean Score**	**Error**	Idealized Predicted Scores based on **familiarity**	**learning**
GR (grammatical/regular)	0.9167	0.0295	1	1
UR (ungrammatical/regular)	0.4323	0.0642	0	0
GC (grammatical critical)	0.6719	0.0496	0	1
UC (ungrammatical critical)	0.3125	0.0625	0	0

A 2x2 ANOVA on the showed that there was a main effect of grammaticality ($F(1,15)= 46.096$, $p<0.001$ and of item type ($F(1,15)= 21.292$, $p<0.001$), but there was no interaction between the two ($F(1,15)= 2.030$, $p< 0.175$). Pairwise comparisons showed that the difference between the GC and UC conditions ($F(1,15)= 17.752$, $p<0.001$) was significant.

Table 9.2

Mean Confidence Ratings for all Word Types for Experiment 9 (N=16) and idealized scores

Word Type	Mean Score	Error	Idealized Predicted Scores based on familiarity	learning
GR (grammatical/regular)	4.6250	.3701	6	6
UR (ungrammatical/regular)	-0.1198	.5915	-6	-6
GC (grammatical critical)	1.9844	.4230	-6	6
UC (ungrammatical critical)	-1.0313	.5405	-6	-6

As with the grammaticality judgments, there was a significant effect of grammaticality ($F(1,15)= 45.304$, $p<0.001$) and of item type ($F(1,15)= 24.366$, $p<0.001$). There was also a significant interaction ($F(1,15)= 6.080$, $p<0.026$), which is not surprising since these are confidence ratings. Pairwise comparisons showed that the difference between the GR and UR conditions was significant ($F(1,15)= 53.113$, $p<0.001$) as was the difference between the GC and UC conditions ($F(1,15)= 18.685$, $p<0.001$).

Discussion

Clearly, the results of this experiment pattern with the idealized outcome based on learning, just as the results of Experiment 7 did. The fact that the results of Experiment 8 are not nearly as strong as the results of Experiments 7 and 9 is strange and anomalous. I set aside this issue for future research, but acknowledge that any serious work on these questions in the future will need to explain the results of Experiment 8.

Recall that the experiments described here were designed to meet four aims. These aims are re-stated below:

1) to determine whether successful artificial grammar experiments can generalize to natural language

2) to determine if it is possible to categorize words on the basis of morphophonological information alone

3) to determine whether adults are capable of making this categorization

4) to probe the possibilities of implicit learning in second language acquisition

First, these experiments clearly show that work in artificial grammar can be generalized to a natural language, in this case, Russian. This is an important finding, because for many years it was assumed that artificial grammar research would generalize to natural language, but that assumption was not tested. These results, then, allow researchers to be more confident about the generalizability of the results of artificial grammar experiments.

Second, the results demonstrate that morphological information is sufficient for category learning, provided that there is enough morphological information. This confirms Braine's (1987) assertion that in order for category learning to take place, there must be a sufficient amount of converging information. In these experiments, the sequence of two co-occurring strings (the derivational and inflectional endings) were the necessary requirements to induce category learning.

Third, it is clear that adults are capable of making this category distinction. This raises the question of what the results might be if children or infants were employed in this experiment, as opposed to university undergraduates. Current research suggests that 17-month-old infants are capable of making this distinction (Gerken, Wilson & Lewis, in prep.).

Lastly, this experiment successfully showed (contra DeKeyser, 1995 and Robinson, 1997) that implicit learning is possible in second language learning, given the right kind of input. Admittedly, these experiments contained no formal means of assessing whether or not the participants were aware of having learned the categories.

However, observations of and conversations with participants suggests that they were unaware of any learning having taken place.[13] This is important because, although it is widely assumed that second language learners learn some kinds of grammatical rules implicitly, no research has been able to demonstrate this experimentally. It is clear that the lack of success in the previous experiments was simply due to insufficient information in the input. This suggests that while implicit learning is possible, it must be guided by very specific types and amounts of converging information in the input. This finding is a first step in determining what kinds of input are necessary for implicit learning to take place.

Additionally, it should be noted that for Experiments 4-7, the experimental stimuli closely resemble those of Brooks (1993). However Brooks was only able to achieve a weak MN/PQ effect. It is difficult to say what the exact source of difference is. Brooks' methodology differed substantially from the paradigm used here even if her stimuli did not. There are several possible explanations for the fact that her study did not really show an MN/PQ effect. One possibility is that the test task was too difficult. Production tasks are more difficult than judgment tasks and may simply be unable to pick up the kind of learning that has taken place here. Another, more interesting possibility, is

---

[13] Examples of observations include: raised eyebrows when explained what they would be doing on the test, comments such as, "You've got to be kidding," and snickers and laughs. Conversations after the experiment, when the aim of the experiments was explained to the participants, indicated that the participants were unaware that they had learned anything.

However, in tests of implicit learning it is often the case that, instead of a 'grammaticality judgment' type of task, participants are asked whether they remember hearing an item or not. The idea that because the participants were exposed to the meanings of the words, the automatic mechanisms usually employed for this type of learning were unable to operate. This makes sense if we consider that it is likely that infants do not know the meanings of a great number of words, but that they have considerable knowledge of grammar. This might suggest that language learning mechanisms first focus in on regularities in the input and set aside the question of meaning until later in the language acquisition process.

---

behind this is that if the participants have learned the material, they will 'remember' having heard it even if they have never heard it. It is unclear what the results would be had that method been employed.

CHAPTER V. FOLLOW-UP EXPERIMENTS

Since it has been shown that participants are capable of learning categories based on distributional information alone, provided there is enough convergent information in the input, the question remains as to how much information is actually necessary. The previous experiments have demonstrated a need for correlated cues, but have not specified how much of a correlation needs to be present in order for learning to take place.

It might be possible that the correlational information (in this case, the double-marking) is actually only necessary for one of the categories, as the double-marking itself may provide enough information for learning the categories. The following experiments examine this question.

Experiment 10.

Method and Results

Participants. Sixteen undergraduate students participated in this experiment for course credit.

Materials. Same as for Experiment 7, except that now only the masculine set contains double-marking (the derivational suffix 'tel' plus the inflectional ending). As in previous experiments, only half of the masculine items have the double-marking. Thus, only a quarter of the entire experimental set contains double-markings, and those double-markings occur in only one category:

grammatical feminine set

malinoj	rubashkoj	lapoj	vannoj	knigoj	korovoj
malinu	rubashku	lapu	vannu	knigu	

grammatical masculine set

uchitelya	stroitelya	zhitelya	kornya	tramvaya	pisarya
uchitelem	stroitelem	zhitelem	kornem	tramvayem	

Procedure. Same as Experiment 7.

Scoring. Same as Experiment 7.

Results. Mean values for the grammaticality judgments in Experiment 10 are given in Table 10.1 and mean values for the confidence ratings are given in Table 10.2.

Table 10.1

Mean "Grammatical" responses for all Word Types for Experiment 10 (N=16) and idealized scores

			Idealized Predicted Scores based on	
Word Type	Mean Score	Error	familiarity	learning
GR (grammatical/regular) Fem	0.7917	0.0579	1	1
GR (grammatical/regular) Masc	0.9271	0.0402	1	1
UR (ungrammatical/regular) Fem	0.4792	0.0524	0	0
UR (ungrammatical/regular) Masc	0.3854	0.0658	0	0
GC (grammatical critical) Fem	0.5938	0.0938	0	1
GC (grammatical critical) Masc	0.5625	0.1106	0	1
UC (ungrammatical critical) Fem	0.3125	0.1008	0	0
UC (ungrammatical critical) Masc	0.4063	0.1138	0	0

A 2x2x2 (item type, grammaticality and gender) ANOVA showed a significant effect of grammaticality ($F(1,15)= 41.068, p<0.001$) and of item type ($F(1,15)= 8.687, p<0.010$). There was no effect of gender ($F(1,15)= 0.131, p>0.723$). Additionally, the ANOVA showed an interaction ($F(1,15)= 5.952, p<0.028$), between item type and grammaticality. Pairwise comparisons showed that the difference between the the GC and UC conditions ($F (1,15)= 23.710, p< 0.001$) was significant.

Table 10.2

Mean Confidence Ratings for all Word Types for Experiment 10 (N=16) and idealized scores

			Idealized Predicted Scores based on	
Word Type	Mean Score	Error	familiarity	learning
GR (grammatical/regular) Fem	3.5833	0.5567	6	6
GR (grammatical/regular) Masc	4.5104	0.4766	6	6
UR (ungrammatical/regular) Fem	-0.0729	0.5284	-6	-6
UR (ungrammatical/regular) Masc	-0.7917	0.5653	-6	-6
GC (grammatical critical) Fem	1.0313	0.8484	-6	6
GC (grammatical critical) Masc	0.5938	0.9652	-6	6
UC (ungrammatical critical) Fem	-1.9688	0.8222	-6	-6
UC (ungrammatical critical) Masc	-0.1563	1.0173	-6	-6

The analysis of confidence ratings showed that there was a significant effect of grammaticality ($F(1,15)=46.066, p<0.001$) and of item type ($F(1,15)=12.852, p<0.003$). Additionally, there was an interaction between the two ($F(1,15)= 13.618, p<0.002$) and a three-way interaction among the three variables ($F(1,15)= 4.652, p<0.048$). The difference between the GC and UC conditions ($F (1,15)= 20.393, p<0.001$) was significant.

The results of both the grammaticality judgment and the confidence rating tasks in Experiment 10 pattern with the idealized outcome based on learning, but this evidence is weak at best and might instead be interpreted as a lack of ability to learn categories.

Discussion

The results of Experiment 10 suggest that double-marking is necessary in only one of the categories in order for category induction to take place. Considering how little convergent information is actually present in the input (recall that the doubly-marked items account for only one-fourth of the total input), this would imply that the learning mechanisms involved in category induction are actually quite strong, given the right kind of input.

The double-marking on the masculine nouns in Experiment 10 was quite salient in that the derivational suffix in those nouns comprised an entire syllable, 'tel.' It is possible that the salience of the marking actually compensated for the low frequency of converging information. Experiment 11 seeks to examine whether a less salient marking will produce similar results to Experiment 10, despite the fact that Experiment 10 yielded weak and equivocal results.

Experiment 11.

Method and Results

Participants. Sixteen undergraduate students participated in this experiment for course credit.

Materials. Same as for Experiment 7, except that now only the feminine set contains double-marking (the derivational suffix ‘k’ plus the inflectional ending). Note that the suffix ‘k’ is likely not as salient as the masculine suffix ‘tel,’ given that it comprises only one segment and not an entire syllable.

grammatical feminine set

polkoj	rubashkoj	ruchkoj	vannoj	knigoj	korovoj
polku	rubashku	ruchku	vannu	knigu	

grammatical masculine set

uchitelya	iyulya	medvedya	kornya	tramvaya	pisarya
uchitelem	iyulem	medvedem	kornem	tramvayem	

Procedure. Same as Experiment 7.

Scoring. Same as Experiment 7.

Results. Mean values for the grammaticality judgments in Experiment 11 are given in Table 11.1 and mean values for the confidence ratings are given in Table 11.2.

Table 11.1

Mean Proportions of "Grammatical" responses for all Word Types for Experiment 11 (N=16) and idealized scores

Word Type	Mean Score	Error	Idealized Predicted Scores based on familiarity	learning
GR (grammatical/regular) Fem	0.7708	0.0453	1	1
GR (grammatical/regular) Masc	0.8437	0.0515	1	1
UR (ungrammatical/regular) Fem	0.3854	0.0622	0	0
UR (ungrammatical/regular) Masc	0.2292	0.0625	0	0
GC (grammatical critical) Fem	0.7188	0.0909	0	1
GC (grammatical critical) Masc	0.625	0.107	0	1
UC (ungrammatical critical) Fem	0.1875	0.0625	0	0
UC (ungrammatical critical) Masc	0.3125	0.0774	0	0

A 2x2x2 ANOVA showed an effect of grammaticality ($F(1,15)= 64.560$, $p<0.001$) and of item type ($F(1,15)= 4.947$, $p<0.042$), but showed no effect of gender ($F(1,15)= 0.080$, $p>0.781$) nor any interactions, either between type and grammaticality ($F(1,15)= 0.735$, $p>0.405$), or a three-way interaction ($F(1,15)= 3.571$, $p>0.078$), suggesting that participants had learned the categories Pairwise comparisons showed that the difference between the GC and UC conditions ($F(1,15)= 39.194$, $p< 0.001$) was significant.

Table 11.2

Mean Confidence Ratings for all Word Types for Experiment 11 (N=16) and idealized scores

Word Type	Mean Score	Error	Idealized Predicted Scores based on familiarity	learning
GR (grammatical/regular) Fem	2.9063	0.4335	6	6
GR (grammatical/regular) Masc	3.9583	0.5285	6	6
UR (ungrammatical/regular) Fem	-0.6979	0.5017	-6	-6
UR (ungrammatical/regular) Masc	-2.2188	0.6379	-6	-6
GC (grammatical critical) Fem	1.8125	0.7904	-6	6
GC (grammatical critical) Masc	1.5313	0.9709	-6	6
UC (ungrammatical critical) Fem	-2.1563	0.5395	-6	-6
UC (ungrammatical critical) Masc	-1.1875	0.7099	-6	-6

The analysis of confidence ratings showed that there was a significant effect of grammaticality ($F(1,15)=43.498$, $p<0.001$) and of item type ($F(1,15)=8.401$, $p<0.011$) but again, there were no interactions, either between type and grammaticality ($F(1,15)= 3.423$, $p>0.084$) or a three-way interaction ($F(1,15)=4.048$, $p>0.063$). The difference between the GC and UC conditions was significant ($F(1,15)= 21.111$, $p<0.001$).

The results of both the grammaticality judgment and the confidence rating tasks in Experiment 11 pattern with the idealized outcome based on learning.

Discussion

The results of this experiment point again to a strong learning mechanism in the face of a small amount of converging information. The double-marking in this case came from the derivational suffix 'k,' a single segment, suggesting that it is not necessarily the salience of the suffix that matters, but the fact that it regularly co-occurs with another suffix that is the crucial factor in these experiments. It is interesting that the results of Experiment 11 were much stronger than Experiment 10. The results of Experiment 10, as with the results of Experiment 8, warrant an explanation, but that lies outside the scope of this dissertation.

It should be noted that Experiment 3 was functionally equivalent to Experiment 11, except that the set size was much smaller. A *post hoc* examination of the items in Experiment 3 shows that, as in Experiment 11, half of the feminine nouns were doubly marked with the 'k' suffix. The set for Experiment 3 is given again below:

grammatical feminine set

devushkoj	ruchkoj	knigoj	korovoj
devushku	ruchku	knigu	

grammatical masculine set

dekana	mal'chika	vora	brata
dekanom	mal'chikom	vorom	

Recall that Experiment 3 produced results that did not pattern with category learning; the participants were simply responding to the familiarity of the items. Although there were a smaller number of participants in Experiment 3 than in Experiment 11, it is clear that the results are dramatically different and would be unlikely to change if there were six more participants in Experiment 3. Thus, these experiments suggest that set size also appears to be a crucial, but not the only, factor.[14]

Another way to examine category learning is through reaction times to stimuli. For example, it might be that participants have shorter reaction times to grammatical items than ungrammatical items if they have learned the categories. Finding another measure of category learning is important to ensure that the findings reported here are not simply an artifact of the experimental method. Thus, the following experiment was performed to explore the issue of reaction times in these types of experiments.

Experiment 12

Method and Results

Participants. Eighteen undergraduate students participated in this experiment for course credit.

Materials. As with previous experiments, two sets of Russian words were created. However, due to a need for many data points, each set contained a large number of items. As with previous experiments, half the items in each set were doubly-marked for gender. However, unlike previous experiments, this time one of the markings preceded the noun, simulating a determiner. The Russian language does not have a wide variety of determiners, so artificial ones were created. The gendered suffixes used in previous experiments (-k and -tel') were used again in this experiment. The case endings were

---

[14] Although set size is important, increased set size alone will not induce category learning, c.f. Experiment 5.

removed from the ends of the words. Thus, the resulting sets were semi-artificial words of Russian. As in previous experiments, some cells were left empty. These items comprise the "critical items." The sets were as follows:

set 1

na	rubashk	devushk	ruchk	muk	morkovk	chernik		trub	korov	rezin	pugovits	pomad
gi	rubashk	devushk	ruchk	muk	morkovk	chernik	zim	trub	korov	rezin	pugovits	

set 2

po	uchitel'	stroitel'	zhitel'	zritel	lyubitel'	pisatel'		vor	koren'	brat	sosed	karandash
su	uchitel'	stroitel'	zhitel'	zritel	lyubitel'	pisatel'	aptekar'	vor	koren'	brat	sosed	

Thus, during training, participants heard determiner-noun combinations such as 'na rubashk' and 'su vor.'

Next, two post-training tests were created. Each test contained 10 items from the training (grammatical items), 10 ungrammatical items (that is, words from the training paired with an incorrect determiner, such as 'su devushk,' for example), 2 grammatical critical items, 2 ungrammatical critical items and 24 fillers. The fillers consisted of a determiner from the training paired with a noun that never occurred in the training. The tests were counter-balanced such that participants were never tested on the same noun twice (that is, they never heard a noun in a grammatical *and* an ungrammatical combination with the determiner).

Seven training sessions were created from the sets described above using Superlab, a program which synchronizes the display code with the video raster, allowing accurate reaction time (RT) measurement. The tests were also imported into the Superlab format.

Procedure. Participants were seated in front of a computer and were told that they would be learning some words of Russian. They were briefly instructed in the use of the computer. Participants were then instructed to listen to the training sessions described above. After hearing each word combination they were to say the combination aloud and then press the space bar to resume listening to the training material. The participants were instructed to pronounce the Russian words as best they could, but not to worry too much about their pronunciation. The duration of the training sessions lasted 15-20 minutes.

Next the participants were given a brief practice to familiarize them with the two tasks involved in the test. In the test, participants heard a determiner-noun combination. They were first asked whether the combination contained a word that was new to them. If so, they were to press the 'P' key on the computer. If the combination did not contain a new word, they were to press the 'Q' key. The computer keyboard was outfitted with a mask that only left the 'Q,' 'P,' and space keys available to the participants.

Next, participants were instructed to perform a grammaticality judgment on the combination, as in all the previous experiments described here, and to record their decisions on a separate answer sheet. When they had finished both of these tasks, they were to press the space bar to hear the next test item.

Scoring. Reaction Times were recorded for the first task (pressing 'P' or 'Q'), as were errors. The grammaticality judgments were scored as in previous experiments.

There are several predictions. First, given that half of the items in each set were doubly-marked, participants should have been able to learn the two categories in this experiment. Thus the mean grammaticality judgment scores for the Regular Grammatical items and the Critical Grammatical items should be high; the mean scores for the Regular Ungrammatical items and the Critical Ungrammatical items should be low.

Second, if reaction time is a valid measure of learning, there should be shorter RTs to the grammatical items, both Regular and Critical. RTs to Ungrammatical items (Regular and Critical) should be significantly longer.

Results. Mean Reaction Times for each condition are given in Table 12.1. Mean scores for the grammaticality judgments are given in Table 12.2, along with idealized scores.

Table 12.1

Mean Reaction Times for all Word Types for Experiment 12 (N=18)

Word Type	Mean RT	Standard Error
GR (grammatical/regular)	1695.2	186.3
UR (ungrammatical/regular)	2204.8	256.8
GC (grammatical critical)	1783.1	178.3
UC (ungrammatical critical)	2259.7	330.4

A 2x2 ANOVA showed that there was a significant effect of grammaticality ($F(1,17)= 7.973$, $p<0.012$) but there was no effect of item type ($F(1,17)= 0.176$, $p<0.680$) and no interaction between the two ($F(1,17)= 0.013$, $p<0.909$). The difference between the GC and UC conditions was marginal ($F(1,17)= 3.982$, $p< 0.062$).

Table 12.2

Mean Proportions of "Grammatical" responses for all Word Types for Experiment 12 (N=18) and idealized scores

			Idealized Predicted Scores based on	
Word Type	**Mean Score**	**Error**	**familiarity**	**learning**
GR (grammatical/regular)	.8850	.0264	1	1
UR (ungrammatical/regular)	.7400	.0426	0	0
GC (grammatical critical)	.8000	.0669	0	1
UC (ungrammatical critical)	.6750	.0750	0	0
Fillers	.3563	.0510	.5	.5

A 2x2 ANOVA showed a significant effect of grammaticality ($F(1,17)=8.761$, $p<0.008$) but not of item type ($F(1,17)= 2.395$, $p>0.138$), nor was there a significant interaction between the two ($F(1,17)= 0.054$, $p>0.819$). The difference between the GC and UC conditions was not significant ($F(1,17)= 2.436$, $p>0.135$).

The scores for the ungrammatical items are unusually high in this experiment. This is likely due to the presence of the fillers. Participants had heard neither the ungrammatical nor the filler items before, but the two are not equal in terms of familiarity. For the grammatical items, participants had heard the components of the items in the training, but the filler items were completely new words. This also probably explains why the scores for the fillers were so low (that is, lower than the .5 expected by chance).

Discussion

The fact that Experiment 12 showed strong effects with both RTs and grammaticality judgments suggests that RT is a valid measure of learning. This adds to the accumulated evidence supporting the idea that it is the double-marking that helps participants to learn the categories.

The evidence presented in this chapter suggests that participants are able to learn syntactic categories on the basis of morphophonological information alone. While this is in itself interesting, the real world implications of this are severely weakened if it cannot be shown that the Russian language is significantly rich with the requisite morphophonological markings. Thus, the last experiment consisted of a corpus search to determine the distribution of such markers in Russian.

Experiment 13

Method and Results

This experiment sought to determine whether or not double-marking is a reliable cue in Russian.

Procedure. Using a search engine created by W. Lewis (2002), Russian-language sites on the World Wide Web were searched for instances of the morphological markers used in the experiments discussed above (-tel'- and -k-). The program is similar to commercial search engines such as *Google* and *Yahoo!* in that it "spins a web" of URLs from a seed site. Unlike *Google* and *Yahoo!*, however, the search engine does not retrieve websites, but instead retrieves linguistically relevant tokens such as sentences, words, morphemes, etc.

Two kinds of searches were conducted. First, given the total number of instances of the case-markings used in the experiments (for feminine nouns -oj and -u and for masculine nouns -em and -ya), a search was conducted to determine the proportion of those instances which also included the derivational marker (tel or k). The second search considered the question in reverse. That is, given the total number of instances of -k- and -tel'-, what is the proportion of those occurrences that are followed by the case-markings used in these experiments (oj and u; em and ya). So for the feminine nouns, for example, that meant counting all of the times that k was followed by a case-marking ending and then determining how many of those were the case-ending oj.

It is necessary to examine the two searches simultaneously in order to gain an appreciation of how useful these morphemes could be to the learner. For example, it might be the case that X has a high co-occurrence rate with Y, but that Y is so frequent without X that the co-occurrence of X and Y is almost meaningless. The reverse could also be true: that is, that Y is very infrequent, but that in the few times that Y occurs, X always co-occurs with it. Thus, the proportions were set up as follows:

for feminine nouns:

$$\frac{\text{\# of instances of -koj}}{\text{\# of instances of -oj}} \quad \text{and} \quad \frac{\text{\# of instances of -ku}}{\text{\# of instances of -u}}$$

$$\frac{\text{\# of instances of -oj}}{\text{\# of instances of -k-}} \quad \text{and} \quad \frac{\text{\# of instances of -u}}{\text{\# of instances of -k-}}$$

for masculine nouns:

$$\frac{\text{\# of instances of -telem}}{\text{\# of instances of -em}} \quad \text{and} \quad \frac{\text{\# of instances of -telya}}{\text{\# of instances of -ya}}$$

$$\frac{\text{\# of instances of -em}}{\text{\# of instances of -tel'-}} \quad \text{and} \quad \frac{\text{\# of instances of -ya}}{\text{\# of instances of -tel'-}}$$

The searches were designed to only search word-final strings so that no extraneous lexical items were included (such as *mojka* or *telefon*).

For these counts, the program searched approximately 13,540 Russian-language websites and included approximately 1,600,000 lexical items. Each search is different because the World Wide Web is constantly in flux, rendering a constantly changing corpus. Some searches are larger than others because the initial seed site happened to yield websites which contained more valid links to other sites. Despite the differences in total number of pages searched, the two searches should be comparable since it is the percentages which are compared and not the raw counts.

Results. The searches revealed the following counts:

for feminine nouns:

instances of -koj: 1508 / instances of -oj: 7489	and	instances of -ku: 871 / instances of -u: 5188
# of instances of -koj: 3359 / # of instances of -k-: 37,061	and	# of instances of -ku: 1499 / # of instances of -k-: 37,061

This results in the following percentages:

Percentage of instances of -oj that are also instances of -koj:	20%
Percentage of instances of -u that are also instances of -ku:	16.7%
Percentage of instances of -k that are also instances of -koj:	9%
Percentage of instances of -k that are also instances of -ku:	4%

for masculine nouns:

instances of -telem: 60 / instances of -em: 2818	and	instances of -telya: 195 / instances of -ya: 35314
instances of -telem: 127 / instances of -tel'-: 2509	and	instances of -telya: 265 / instances of -tel'-: 2509

This results in the following percentages:

Percentage of instances of -em that are also instances of -telem:	2%
Percentage of instances of -ya that are also instances of -telya:	.5%
Percentage of instances of -tel that are also instances of -telem:	5%
Percentage of instances of -tel that are also instances of -telya:	10%

<u>Discussion</u>

In the first search, it is clear that -k- is a very frequent feminine marker because 20% of all instances of -oj and 16.7% of all instances of -u are preceded by 'k.' It is also clear from the first search that 'tel' is not a very frequent marker since only 2% of the instances of -em and .5% instances of -ya are preceded by 'tel.'

In the second search, it becomes clear that -k is not a very reliable marker, given that only 9% of instances of -k precede instances of -oj and only 4% instances of -k precede instances of -u. On the other hand, 'tel' is a fairly reliable marker given that 5% of the instances of -tel precede instances of -em and 10% of the instances of -tel precede instances of ya. Although the percentages for -tel look low, it is important to remember that the overall frequency is very low so that, for such a low-frequency item, 'tel appears to be quite reliable. In turn, while the feminine marker is quite frequent, it is also not very reliable. The masculine marker, on the other hand, while not very frequent, is fairly reliable.

It is an open question as to whether or not these searches constitute adequate evidence for the richness in morphophonological markers required to induce learning of grammatical categories. The set size for an actual learner is substantially larger than the sets used in the experiments reported here and thus it is difficult to say whether or not the markings are rich enough. However, it should be kept in mind that -k and -tel are only two markings out of many used to mark gender in Russian; I estimate the number of markings to be somewhere around 30. Thus, considering that two markers out of 30 have the frequencies and reliabilities reported above, it seems likely that Russian is indeed rich enough in these markings to provide the kind of evidence required by a learner to form the syntactic categories described here. Recall that Kempe and Brooks (2001) found that between 35% and 40% of all nouns in child-directed speech contain diminutive suffixes. These suffixes, being derivational, are the same kinds of suffixes used in the experiments presented here. Thus it is probably the case that, at the very least, language directed at children contains an adequate number of morphophonological markings.

CHAPTER VI. CONCLUSIONS

This conclusion section consists of two parts: 1) a discussion of big-picture issues and 2) an outline of unresolved questions. Ideally, this dissertation will have contributed something to the framework in which we consider the larger theoretical issues such that they can be followed up by future researchers. Likewise, the long list of unresolved questions merits attention so that the theoretical issues can be continually revised and updated.

Big-Picture Issues

The experiments discussed here point to one general conclusion: non-semantic category learning is possible on the basis of morphophonological information alone. This has been something that several researchers have hypothesized for quite some time (beginning with Braine, 1963, and Smith, 1969) but have so far been unable to demonstrate experimentally. The work presented here, taken in conjunction with previous work done in this area, can at last provide convincing evidence that this kind of learning can take place.

The implications of this finding concern language learning in both infants and adults. First of all, recall that Pinker (1987) hypothesized that major syntactic categories (such as noun and verb) are learned *only* after the meanings of words are learned. For Pinker, it is the meaning which provides the crucial information necessary to map the learned lexical item onto the innate categories of noun and verb. For sub-categories, such as gender, Pinker had to posit a powerful learning mechanism capable of learning arbitrary categories.

The evidence presented here, however, calls Pinker's hypothesis into question. The subjects in the experiments for this dissertation did not know the meanings of the words they were learning. However, they were still able to learn grammatical categories, provided the input was sufficiently rich in morphophonological markings. Thus the need to learn meanings in order for category acquisition to take place seems dubious. Also, the existence of the rich morphophonological markings found in Russian calls into question Pinker's assumption that syntactic sub-categories are completely arbitrary. In Russian, at least, they appear to be anything but arbitrary. Thus, the results of these studies confront not only Pinker's account of sub-category learning, but also show that a closer look at major category learning (noun and verb) is warranted as well.

What this work suggests is that infants might be able to learn a great deal about grammatical categories before they are aware of many word meanings. On the surface, this would seem counter-intuitive. How can a baby know that 'table' is masculine even before knowing what a table is? The evidence presented here points to the fact that humans can categorize words based on their morphophonological properties alone. So the grammatical and phonological markings on the word 'table,' taken in conjunction with knowledge of other grammatical and phonological markings, provide the infant with enough evidence to decide that there are X number of categories and that 'table' belongs to

Category A. The studies outlined above show that adults do just that when presented with enough of the right kind of information. Experiments in progress point to the fact that infants are also capable of the same kind of learning (Gerken, Wilson, & Lewis, in prep.).

There is another aspect of Pinker's hypothesis that should be revisited. On his view, learners are born with innate syntactic categories with mapping mechanisms to link up those already-existing categories with learned meanings. Learners later rely on distributiuonal cues to learn the categories of words that are not given semantically. If, as studies reported here suggest, humans are able to form syntactic categories on the basis of morphophonological information alone, then one reason for assuming innate knowledge of specific categories is called into question. The problem remains, however, about how to map syntactic categories onto the propositional structure of sentences. The mapping problem was largely solved by Pinker's approach, and another solution will be required for purely distributional accounts of category learning to be viable. That issue is beyond the scope of this dissertation.

In so far as the work presented here raises questions about innate knowledge, it highlights the role of innate learning mechanisms. In other words, humans may be hard-wired to learn categories under very specific conditions. The very fact that the observed effects here were previously so difficult to obtain points to the fact that very specific kinds of information must be present in order for learners to posit categories. In particular, it appears that the learning mechanism is well-suited to discern categories that are, in part, doubly-marked. Without this 'double-marking detector,' subjects would have either been 1) unable to learn the categories under any conditions or 2) able to learn the categories under any condition. Instead, learners formed the categories under very specific conditions. This points to an underlying learning mechanism, which, given the lack of experience in infants, is probably innate. One criticism of relying on distributional information in the input for learning is that doing so would lead to overgeneralization (Gleitman & Wanner, 1997). For example, you might group 'looking' and 'sing' into the same category because they both end in 'ing,' but 'look' and 's' do not belong in the same category. However, the research presented in this dissertation suggests that overgeneralization is unlikely. Recall that double-marking is required in order for category learning to take place; it would be highly unlikely for two words to share *two* category-markers and not belong in the same category. Again, this points to the fact that this learning mechanism is probably innate, since it saves learners from overgeneralization.

Whenever the issue of innateness is raised in language studies, other questions are raised concurrently: are some innate learning mechanisms specific to language learning? Or is it the case that learning mechanisms are general and can be employed for many learning purposes? Some recent work on statistical learning suggests that learning mechanisms may be general and not specific (Aslin, Saffran, & Newport, 1999), but the studies covered in this dissertation do not offer any evidence on the specificity of learning mechanisms. Theoretically, either scenario is compatible with the results discussed here.

The issue of innateness raises another issue as well. It is commonly assumed that there is a critical period in language learning (Lenneberg, 1967). One piece of the evidence for this is that after the age of about 12, humans rarely master either a first or a second language. However the work here (taken in conjunction with the infant studies) shows that adults and infants can learn syntactic categories under similar conditions. Therefore it might be possible that the observed critical period for language learning in humans is actually the artifact of some other process. For example, late language learners tend to focus on the meanings of words; this is natural, since they have been deprived of a means of communication, their primary desire is to communicate, not learn grammar. However, an infant's focus may be quite different. During the period of time that an infant is pre-verbal, it is possible that she is listening to the distributional regularities of the input and not concentrating on the meaning (Gerken, in press; Naigles, in press). This would mean that early and late learners have radically different learning styles. Perhaps this is the reason for the observed critical period. Future research could shed light on this question.

Returning to the main issues raised by this research, if meaning is not necessary for category learning to take place, it is appropriate to question what role meaning does play in the acquisition of grammar. As shown in Braine (1987), meaning can contribute to category learning if that meaning is useful (in other words, if the meaning of the words form obvious semantic categories). If, however, meaning does not serve as a cue to categories (as in Brooks, Braine, et al. (1993) and Kempe and Brooks (2001)), it is possible that meaning might actually interfere with learning. This is because both the Brooks and the Kempe study contained the "double-marking" necessary for category learning, but both studies provided only weak evidence that the subjects had learned the categories. This fact is especially relevant considering the differences in training times among the studies: in the Brooks study subjects were trained for nearly four hours; the Kempe study included two hours of training; the subjects in most of the current studies were trained for only seven minutes. Given that the Brooks, et al. and the Kempe & Brooks studies contained the necessary markings on the items and that subjects were trained on those items for a considerable amount of time, the lack of success must be due to the main difference between those experiments and the experiments reported here: the Brooks and the Kempe studies required the subjects to learn the meanings of the words; the present experiments did not.

Although this seems to be a valid hypothesis of the role of meaning in the acquisition of grammar, it is unclear as to how this is actually operationalized. One possibility is that there is simply too much demand on the short-term memory system both to remember meaning and to detect regularities of form; another possibility is that the category learning system attempts to integrate all available information and the presence of semantic information throws the system off since that information is not reliable in determining categories; still a third possibility is that, for adult learners, meaning is potentially more useful than form regularities and subjects simply devote more energy to learning meanings, ignoring form for the most part. One way to probe this question would be to run the experiments described here and add a semantic component, thus verifying that meaning is indeed the crucially different factor between the Brooks and Kempe studies and the present dissertation.

One criticism of relying on distributional information in the input for learning is that doing so would lead to overgeneralization (Gleitman & Wanner, 1997). For example, you might group 'looking' and 'sing' into the same category because they both end in 'ing,' but 'look' and 's' do not belong in the same category. However, the research presented in this dissertation suggests that overgeneralization is unlikely. Recall that double-marking is required in order for category learning to take place; it would be highly unlikely for two words to share *two* category-markers and not belong in the same category. Again, this points to the fact that this learning mechanism is probably innate, since it saves learners from overgeneralization.

Returning to the issue of adult learners (which is usually an instance of second language learning), it is important to point out that this dissertation is one of the few studies which documents implicit learning in a second language. The objection could be made that there were no formal measures to determine whether or not subjects were conscious of the learning they underwent. This is true. However, the reaction from the majority of subjects was so strong and they expressed such disbelief that they had learned anything that I feel confident in asserting that the learning was implicit. The fact that implicit learning is now shown to be possible is very important since it had been assumed for so long but had not been documented. Indeed, many researchers were beginning to question whether or not is was possible (see literature review).

However, the issue of meaning, raised above, is crucial to the issue of second language learning. Is it possible, or even desirable, to learn a second language without meaning for a (possibly long) period of time? Can an entire second language be learned implicitly, or will there always be a need for some explicit learning to take place?

The questions raised here are numerous and complex and far outside the domain of this dissertation. However, one idea springs to mind as a way to easily integrate this kind of category learning into the language classroom. During introductory lessons, when students are typically engaged in learning the sound system of a language, it would be easy to construct input so that it could have the additional benefit of inducing some category learning. For example, it would be possible for a teacher of Russian to use the experimental sets presented in this dissertation to have students practice writing Russian words and practice the Russian sound system. Then students would be equipped not only with Russian letters and sounds, but possibly grammatical categories as well. It remains to be seen whether this kind of implicit learning can immediately transfer to explicit, ready-for-production grammatical structures, but there is a chance that such an exercise would give those students a "heads up" on the grammatical system.

Unresolved Questions

Recall that in Experiments 1-3, the mean scores for the UR, GC, and UC items form a curious pattern. Those means are repeated here:

Experiment 1:

Word Type	Mean Score	Error
GR (grammatical/regular)	.8766	.0321
UR (ungrammatical/regular)	.4610	.0612
GC (grammatical critical)	.4545	.0880
UC (ungrammatical critical)	.4773	.0920

Experiment 2:

Word Type	Mean Score	Error
GR (grammatical/regular)	.8265	.0377
UR (ungrammatical/regular)	.6327	.0550
GC (grammatical critical)	.3214	.0899
UC (ungrammatical critical)	.3929	.0922

Experiment 3:

Word Type	Mean Score	Error
GR (grammatical/regular)	.8429	.0581
UR (ungrammatical/regular)	.4286	.0522
GC (grammatical critical)	.2000	.0816
UC (ungrammatical critical)	.3750	.1070

In Experiment 1, the means for the Grammatical Regular items are high (as expected, since the subjects are familiar with them) and the other three types of items are about at chance, which is expected since subjects are equally unfamiliar with all of them. However, in Experiments 2 and 3 we see a different pattern. In those experiments, there is a difference between the Ungrammatical Regular items and the Critical items (both grammatical and ungrammatical). Why are the Ungrammatical Regular items somehow more acceptable to the subjects than the Critical items? This is especially curious in Experiment 3, where the counterpart to the critical item (that is, the "cell above the empty cell") was repeated in the training so that subjects would be as familiar with the stem of that lexical item as they would be with the stems of all the other lexical items. So, theoretically, there should be no reason for the critical items to seem even less familiar than the ungrammatical items. This issue warrants future research because it is clear that some process is occurring here and it is possible that that process may be an important component of actual learning. At the very least, since we have assumed an innate learning mechanism, it would give us a clearer picture of what the innate learning mechanism is like; one has to study both instances of learning and failures to learn in order to fully understand the innate learning system.

Another unanswered question is the role of palatalized consonants in the learning of masculine and feminine categories in Russian. As discussed in a previous section, it is clear that the distribution of palatalized consonants is not completely free when it comes to gender (Wade, 1992). That is, nouns ending in -m' (palatalized m) are uniformly feminine, for example, while nouns ending in -r' (palatalized r) are uniformly masculine. Other palatalized consonants do not have such a clean distribution: b', v', and d' usually indicate feminine nouns, but there are some exceptions. A clear inventory of this distribution needs to be made.

Future research could investigate the role that palatalization might play as a cue to gender learning. At the very least, the existence of at least semi-gendered palatalization points to the fact that gender classes are not arbitrary, but in fact come with a whole host of markings, both morphological and phonological (in Russian, at least). Given the results of Gerken, Gómez and Nurmsoo, it seems highly likely that the distribution of palatalized consonants could play a role in the learning of gender categories.

A third unresolved question is the issue of Experiments 8 and 9. Recall that Experiment 8 was another version of Experiment 7 which fixed the problem of the ungrammatical masculine items (these items contained an additional cue that could have been responsible for the results of Experiment 7). The results of Experiment 8 did not pattern with the idealized outcome based on learning. The results of Experiment 9, however, which only differed from Experiment 8 in the "placement of the empty cell" in the set of training items, did pattern with the idealized outcome based on learning. This is very strange. One possible explanation is that two different tape machines were used to run Experiments 8 and 9, the one which was used to run Experiment 9 was of substantially higher quality. A thorough test of the equipment could provide a viable explanation for the difference between the two experiments. If, however, equipment is not the reason for the difference in outcomes, several more experiments need to be performed in order to determine the precise cause of the difference. The reasons for this are the same reasons stated above for the necessity of future research into Experiments 1-3: to completely understand the nature of the learning mechanism, it is necessary to thoroughly understand the experiments where learning did not take place.

A fourth unresolved issue is the different results obtained in Experiments 10 and 11. These were follow-up experiments designed to test the limits of learning which comes about as a result of the double-marking of some of the lexical items. In Experiment 10, only the masculine items were marked; in Experiment 11, only the feminines were marked. Experiment 10, however, failed to yield results consistent with the idealized outcome based on learning. Experiment 11 did yield results consistent with learning, and the results were very solid. This is quite an odd results considering that the masculine marking 'tel' is likely much more salient than the feminine marking 'k.' As noted above, it is possible that there was a difference based on the equipment used; it is worth testing the equipment with both sets of stimuli to see if this is the case. If, however, the difference is not a result of differences in equipment, the results here are very difficult to explain. In the larger context of Russian, this would be an expected results since 'k' is a much more frequent suffix than 'tel,' but in terms of these experiments, the two endings were of equal frequency. A thorough examination of these two studies is necessary, with particular attention paid to the phonetic/phonological properties of the individual items used in both experiments. It is possible that there was some subtle phonological cue present in the items used in Experiment 11 which was overlooked in the creation of the stimuli. If this is indeed the case, the presence of such cues leading to learning would do nothing but strengthen the hypothesis presented here: that is, category learning is possible on the basis of morphophonological information alone.

One last unresolved issue that warrants mentioning is the make-up of the subjects. In the experiments described here, no subjects were excluded (or included) because of their native language background. However, due to the composition of the university as a whole (and the students of Psychology 101 in particular) most of the subjects were native English speakers. This fact might have played a role in the outcome of the experiments. The subjects who were not native English speakers came from a variety of linguistic backgrounds: subjects were native speakers of Spanish, Cantonese, Japanese, Persian, Arabic and other languages. To the best of my knowledge, no subjects were native speakers of a Slavic language. In the future, it might be interesting to analyze the results of the native English speakers separately from the results of native speakers of other languages. In theory, these experiments should work with speakers of any language. If they do not, then serious consideration needs to be paid to the design and procedure of the experiments.

Final Conclusions

The research here points to a view of language learning that is heavily dependent upon the distributional regularities found naturally in human language to support the learning of syntactic categories. It shows that humans are extremely sensitive to subtle morphological and phonological cues present in linguistic input. Future research should explore the nature of this sensitivity and focus on creating a model of learning which accounts for the results presented here.

REFERENCES

Alanen, R. A. (1992). Input enhancement and rule presentation in second language acquisition. , University of Hawaii at Manoa.

Aslin, R., Saffran, J., & Newport, E. (1999). Statistical Learning in Linguistic and Nonlinguistic Domains. In B. MacWhinney (Ed.), The emergence of language . Mahwah, NJ, USA: Lawrence Erlbaum Associates, Inc.

Braine, M. (1987). What is learned in acquiring word classes-- A step toward an acquisition theory. In B. MacWhinney (Ed.), Mechanisms of Language Acquisition . Hillsdale, N. J.: Erlbaum.

Braine, M. D. S. (1963). On learning the grammatical order of words. Psychological Review, 70, 323-348.

Braine, M. D. S. (1965). The insufficiency of a finite state model for verbal reconstructive memory. Psychonomic Science, 2, 291-292.

Braine, M. D. S. (1966). Learning the position of words relative to a marker element. Journal of Experimental Psychology, 72, 532-540.

Braine, M. D. S. (1971). On two types of models of the internalization of grammars. In D. Slobin (Ed.), The ontogenesis of grammar . New York: Academic Press.

Brooks, P. J., Braine, M. D. S., Catalano, L., Brody, R., & Sudhalter, V. (1993). Acquisition of Gender-like Noun Subclasses in an Artificial Language: The Contribution of Phonological Markers to Learning. Journal of Memory and Language, 32, 76-95
.

Corbett, G. C. (1991). Gender. Cambridge: Cambridge University Press.

DeKeyser, R. (1995). Learning Second Language Grammar Rules. Studies in Second Language Acquisition, 17, 379-410.

Doughty, C. (1991). Second Language Instruction Does Make a Difference: Evidence From an Empirical Study of SL Relativization. Studies in Second Language Acquisition, 13, 431-469.

Ellis, N. (1993). Rules and Instances of in Foreign Language Learning: Interactions of Explicit and Implicit Knowledge. European Journal of Cognitive Psychology, 5, 289-318.

Emmorey, K., Bellugi, U., Friederici, A., & Horn, P. (1995). Effects of age of acquisition on grammatical sensitivity: Evidence from on-line and off-line tasks. Applied Psycholinguistics, 16, 1-23.

Frigo, L., & McDonald, J. (1998). Properties of phonological markers that affect the acquisition of gender-like subclasses. Journal of Memory and Language, 39, 218-245.

Gerken, L. A. (in press). Early sensitivity to linguistic form. Annual Review of Language Development.

Gerken, L. A., Gómez, R., & Nurmsoo, E. (1999). The Role of Meaning and Form in the Formation of Syntactic Categories. . Albuquerque, NM: Society for Research in Child Development.

Gerken, L. A., Wilson, R., & Lewis, W. (in prep.). The induction of syntactic categories by 17-month-olds. .

Gleitman, L., & Wanner, E. (1997). Current issues in language learning. In M. H. Bornstein & M. E. Lamb (Eds.), Developmental psychology: An advanced textbook (2nd Ed). (pp. 297-356).

Green, K., & Zampini, M. (1998). Second language perception . Unpublished ms. University of Arizona.

Gregg, K. (1984). Krashen's Monitor and Occam's Razor. Applied Linguistics, 5, 79-100.

Gvozdev, A. N. (1969). Voprosy Izuchenia Detskoj Rechi. Moscow: Izd. Akadem. Nauk RSFSR.

Karmiloff-Smith, A. (1976). The Interplay Between Syntax, Semantics and Phonology in Language Acquisition Processes. In R. Campbell & P. Smith (Eds.), Recent Advances in the Psychology of Language . New York: Plenum.

Kempe, V., & Brooks, P. J. (2001). The Role of Diminutives in the Acquisition of Russian Gender: Can Elements of Child-Directed Speech Aid in Learning Morphology? Language Learning, 51(2), 221-256.

Krashen, S. (1981). Second Language Acquisition and Second Language Learning. Oxford: Oxford University Press.

Lenneberg, E. (1967). Biological Foundations of Language. New York: Wiley.

Lewis, W. (2002). The Web as Linguistic Resource: Methods for Harvesting and Analyzing Linguistically Relevant Data off the Web. Unpublished Doctoral Dissertation, University of Arizona, Tucson, AZ.

Naigles, L. (in press). Form is easy, meaning is hard: Resolving a paradox in early child language. Cognition.

Pinker, S. (1987). The bootstrapping problem in language acquisition. In B. MacWhinney (Ed.), Mechanisms of Language Acquisition . Hillsdale, N. J.: Erlbaum.

Popova, M. I. (1958). Grammaticheskie elementy jazyka v rechi detej preddoshkol'nogo vozrasta. Voprosi Psixologii, 4(3), 106-17.

Robinson, P. (1997). Generalizability and automaticity of second language learning under implicit, incidental, enhanced and instructed conditions. Studies in Second Language Acquisition, 19, 223-247.

Schmidt, R. (1990). The Role of Consciousness in Second Language Learning. Applied Linguistics, 11, 129-158.

Slobin, D. I. (1966). The Acquisition of Russian as a Native Language. In F. Smith & G. Miller (Eds.), The Genesis of Language . Cambridge, Mass: MIT Press.

Smith, K. H. (1966). Grammatical intrusions in the recall of structured letter pairs: Mediated transfer or position learning? Journal of Experimental Psychology, 72, 580-588.

Wade, T. (1992). A Comprehensive Russian Grammar. Cambridge, Massachusetts: Blackwell.