

# Distributional Learning of Syntactic Categories

Malathi Thothathiri & Jesse Snedeker

Dept. of Psychology, Harvard University, Cambridge, MA 02138

## 1. Introduction

Distributional information (patterns of co-occurrence between words) is one cue to the syntactic category of a word. Several studies have shown that human adults can use this in conjunction with semantic or phonological cues to categorize words. But can human adults use distributional information *alone* for categorization? Two studies using artificial languages have found mixed results:

- Smith (1966) found failure using a set of 2-word sentences.
- Mintz (2002) found success using a set of 3-word sentences.

A current proposal for the disparity is that 3-word sentences used in the latter study provided frames (a\_\_b) that can be used for categorization (Mintz, 2003; hereafter referred to as the Frames Proposal).

## 2. Motivation

The above-mentioned studies differed in more ways than just the length of sentences. Specifically, unlike the *symmetric* design in Smith (1966), Mintz (2002) used a language where the two frames contrasted in test sentences differed in ways other than the critical categorial dimension (e.g., how frequent the middle words were during training and whether middle words only occurred in the middle).

- **In Expt. I, we ask whether adults can use a set of 3-word sentences with no distributional asymmetries to categorize words.**

An alternative to the Frames Proposal is that 3-word sentences are more advantageous because they provide *multiple* relevant adjacent dependencies (Monaghan & Christiansen, 2004).

- **In Expt. II, we ask whether frames are represented and used in this task.**

## 3. Experimental Procedure

Subjects listened to artificial language stimuli over headphones and answered questions on a computer.

- First they listened to words (in random order; once each).
- Then they listened to sentences (in random order; 4 times each).

During test, subjects were told that the sentences they heard were constructed according to certain rules and that they would now listen to new sentences and rate their grammaticality. They were instructed to use their gut feeling in answering the following questions:

- Does this sentence sound right? (Yes or No, recoded as +/-1)
- How sure are you? (Scale of 1 to 7; 7=most confident)

The composite rating for each sentence thus ranged from -7 to +7.

## 4. Sentences

Each of two categories (X and Y) was embedded in four frames. Each category consisted of two words  $[X = \{X_1, X_2\}; Y = \{Y_1, Y_2\}]$ .

**X sentences**  
aX<sub>1</sub>b, aX<sub>2</sub>b  
cX<sub>1</sub>d, cX<sub>2</sub>d  
eX<sub>1</sub>f  
gX<sub>2</sub>h

**Y sentences**  
iY<sub>1</sub>j, iY<sub>2</sub>j  
kY<sub>1</sub>l, kY<sub>2</sub>l  
mY<sub>1</sub>n  
oY<sub>2</sub>p

Language was counterbalanced such that half of the subjects heard aY<sub>1</sub>b, iX<sub>1</sub>j, etc.

Critical test sentences belonged to one of two types:

- **Grammatical:** eX<sub>2</sub>f, gX<sub>1</sub>h, mY<sub>2</sub>n, oY<sub>1</sub>p.
- **Ungrammatical:** eY<sub>2</sub>f, gY<sub>1</sub>h, mX<sub>2</sub>n, oX<sub>1</sub>p

In addition, in Expt. II, we tested frame-violations:

- **Frame-Violations:** aX<sub>1</sub>d, cX<sub>2</sub>b, iY<sub>1</sub>l, kY<sub>2</sub>j

## 5. Grammatical vs. Ungrammatical

The two types were **similar** in the following ways:

- They were novel (not heard during training).
- They contained words in their appropriate positions.
- Their constituent words occurred equally often during training.

The two types were **different** in one way only:

- Only in grammatical sentences did the middle word occur with the appropriate first and third words (as determined by where a sister category word had occurred in the training sentences).

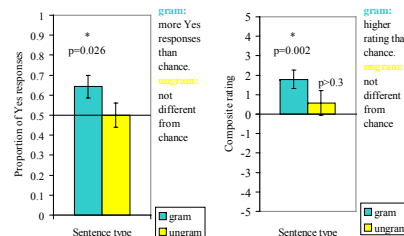
Thus, a difference in acceptability between the two types is evidence for categorization. We evaluated categorization success using two dependent variables:

- Proportion of Yes responses to each type (expect greater Yes responses to grammatical sentences).
- Composite ratings for each type (expect greater ratings for grammatical sentences).

## 6. Expt. I: Subjects accept Grammatical

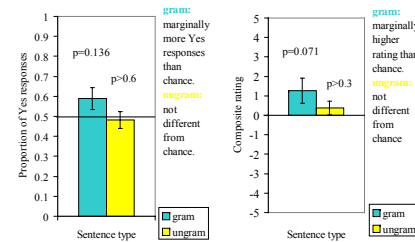
We investigated whether categorization would occur with a *symmetric* training language with 3-word sentences. Test sentences were either grammatical or ungrammatical. 14 college-aged adults participated

(7M, 7F)\*. \* 2 additional adults were excluded because they gave 75% or more responses of one type.



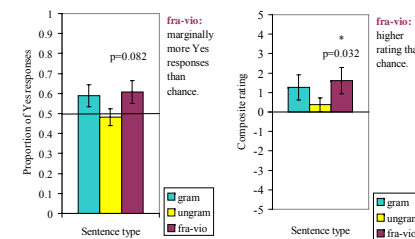
## 7. Expt. II: Subjects accept Grammatical

We tested frame-violation sentences in addition to grammatical and ungrammatical sentences. 14 college-aged adults participated (4M, 10F)\*. \* 6 additional adults were excluded because they gave 75% or more responses of one type.



## 8. Expt II: Subjects accept Frame-Violations

Yes responses and ratings for frame-violations were similar to those for grammatical sentences.



## 9. General Conclusions

- Human adults can use distributional information alone to categorize words.
- Across two experiments, subjects accepted grammatical sentences more than expected by chance (Yes responses:  $t(27)=3.87, p=0.001$ ; Rating:  $t(27)=2.931, p=0.007$ ). They did not accept ungrammatical sentences above chance (Yes responses:  $t(27)=1.336, p=0.19$ ; Rating:  $t(27)=-0.254, p>0.8$ ).
- A significant majority of subjects showed the expected pattern of preferring grammatical over ungrammatical sentences (21/28 across two experiments;  $X^2(1)=7, p=0.008$ ).
- Thus, we provide additional support for previous findings with 3-word sentences (Mintz, 2002) using a symmetric language.
- Distributional asymmetries do not appear to be *necessary* for categorization.
- It is likely that 3-word sentences are more advantageous for categorization than 2-word sentences.

## 10. Evaluating the Frames Proposal

Our findings do not support the Frames Proposal for categorization.

- In Expt. II, subjects accepted frame-violations just as much as they accepted grammatical sentences.
- We suggest instead that single or multiple *adjacent* dependencies drive categorization in our experiments (Monaghan & Christiansen, 2004).
- Languages with a higher ratio between number of category words and number of frames may lead to the utilization of *non-adjacent* dependencies (Gomez, 2002).
- An alternate explanation is that the high ratings given to frame-violations were due to perceived similarity to heard sentences. Frame-violations contained adjacent pairs that had occurred during training. This process may be qualitatively different from the one used to judge the acceptability of grammatical and ungrammatical sentences. Further research is required to rule out this possibility.

## 11. References

- 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.
- Mintz, T. H. (2002) Category induction from distributional cues in an artificial language. *Memory & Cognition*, 30 (5), 678-686.
- Mintz, T. H. (2003) Frequent frames as a cue to grammatical categories in child-directed speech. *Cognition*, 90 (1), 91-117.
- Monaghan, P. & Christiansen, M. H. (2004) What distributional information is useful and usable for language acquisition? Proceedings of the 26<sup>th</sup> annual meeting of the Cognitive Science Society.
- Gomez, R. L. (2002) Variability and detection of invariant structure. *Psychological Science*, 13(5), 431-436.