Syntactic categorization in early language acquisition: formalizing the role of distributional analysis

Cartwright and Brent (1997)

Presented by Alison Smizer

The role of distributional analysis in grammatical category acquisition

→ Children must learn the grammatical categories of individual words in order to acquire language. This is sometimes difficult because the same phonological word can be assigned to two different categories (see and sea).

→ Children use different information sources to categorize words:
  - Words occurring in similar contexts tend to belong to the same grammatical category. This is called distributional regularity.
  - Correlation between meanings of words and their grammatical categories, e.g. words referring to concrete objects are almost always nouns.
  - Phonological forms of words also correlate with syntactic category, e.g. bisyllabic words with final stress are more likely to be verbs than those with initial stress (“récord”, noun, and “recórd”, verb).

→ Analysis of distributional regularities might be more effective if the environments for those analyses were syntactic phrases smaller than complete sentences. Therefore, children might benefit from using whatever syntactic knowledge they have to try to find phrase boundaries. Furthermore, using sentential prosody such as pitch contours may provide some information about the locations of major syntactic boundaries.

Theories of category acquisition

Cognitive Theories

Distributional cues
→ Children might form grammatical categories by grouping together those words that occur in the same environments

Semantic cues
→ Truly grammatical categories are learned, but the process is begun with help from semantics. For example, the fact that a word refers to a concrete object can be used to label it as a noun. After that, words can be categorized either by distributional means or by further application of semantic correlations, although the distributional analysis always has the final say.
Phonological Cues

There are language-specific correlations between certain phonological properties of words and syntactic categories; children may be able to exploit these correlations at some stage of category acquisition. Further, language-independent correlations between phonological forms of words and their syntactic categories seem to provide limited information to category acquisition.

Computational models

Some computer simulations were more helpful than others in the history of the study of language acquisition.

**Hierarchical cluster analysis** – A computational model that groups together words whose distributional patterns are similar.

Started by Kiss (1973)

To use it for learning categories, define 3 parameters: list of words to be analyzed (target words), process by which vectors are derived from the context of larger words, and a distance measure.

Two differences between hierarchical cluster analysis and Cartwright and Brent’s proposed learning strategy:

1. C&B strategy results in a set of discrete categories of words, whereas HCA results in a large number of nested categories. HCA is not as efficient.
2. C&B strategy is incremental: operates on one sentence at a time, forgetting previous sentences. HCA proposes that all language input is processed in one batch. HCA is much more time consuming.

Reconstructing Distributional Categorization

Intuitive Distributional Analysis

Key term: minimal pair

I saw the cat
I saw the dog

Based on these, it is natural to assume that cat and dog belong to same grammatical category.

But minimal pairs rarely exist together in natural speech.

More often, we put words in categories and form templates, which are minimal pairs. Ex:

1. My cat meowed
2. Your dog slept

**Determinates**: My, your

**Nouns**: cat, dog

The new template:

1. Determinate + Noun + X
2. Determinate + Noun + Y

These templates can be viewed as generalized minimal pairs.
From Institutions to Theory

Preferences children use in deciding which groups to merge:
1. Minimize the number of templates. So, if there are two templates, ABCD and ABCE, create a new category for D and E and call it Z. New template is ABCZ.
2. Minimize the sum of the lengths of the templates. So, if there are four templates, AB, AC, PQRSTUV, and PQRSTUW, B and C are merged, V and W are merged. Children prefer the V-W merge because the longer minimal pair is less likely to have occurred by chance.
3. Create templates with the highest possible frequency. Commonly heard = better sources of distributional evidence because less likely to reflect ungrammatical noise.

Preferences of arranging words into groups:
4. Minimize the total number of groups. Every merge reduces the total number of groups by 1.
5. Put all instances of a word type together (unless there is evidence that they don’t belong together)
6. Minimize the number of types whose instances are divided among different groups. Better to have one type that belongs to several groups than several types each belonging to two groups
7. Minimize the number of words in each group. This preference works against merges. This is so children don’t lump every word into one meaningless group. Children merge conservatively.
8. Minimize the number of groups consisting of more than one type.
9. Maximize the frequency of word types within their groups. Go with the more strongly established grouping of the type. If only a small number of tokens occur in another group, it might be due to noise or error, and this group should not be used unless there is strong evidence for it.
10. Use large groups (in terms of numbers of types) in the templates. Use firmly established groups, not small ones that may have resulted from error.

Categorization as Optimization

→ List of preferences is based on the Minimum Description Length (MDL) Paradigm – the optimal description of some data is the shortest one. Formulate a hypothesis and a derivation.

Example of MDL in action:

This is a kitty → ABCD
This is a doggie → ABCD
What a cute kitty → ECFD
What a cute doggie → ECFD

A:{this} B:{is} C:{a} D:{kitty} (d1) {doggie}(d2) E:{what} F:{nice}(f1) {cute}(f2)
→ Now, the complete template list to describe all four sentences is ABCD, ECFD.
→ List templates corresponding to input sentences: 2, 2, 1, 1
Now each sentence can be completely reconstructed.

**Learning Algorithm**

➔ Children continue to merge until no merges can improve the arrangement of groups and templates. Also, an input sentence may never result in a merge if its grouping cannot be improved upon.

Four examples of algorithm in action
1. **No action**: no merge is warranted because the negative votes outweigh the positives.
2. **Minimal pair merge**: Reduces number of groups and reduces number and total length of templates.
3. **Chain reaction**: Preceding input establishes a group, thus templates for the sentences form a generalized minimal pair.
4. **Ambiguity**:
   Fred saw the cat
   Fred saw the dog
   Where is the cat?
   Where is the dog?

   *Where is the saw?*

Children use a specific formula to make sense of this confusing piece of information.

**EXPERIMENTS**
Conducted 5 experiments using the distributional model computer simulations.

**EXPERIMENT 1**
Computer simulation
Is the proposed strategy useful in learning the categories of an artificial language defined in truly distributional terms?

Results suggest that the learning strategy can be effective at categorization and can generalize quite rapidly from a small sample.

**EXPERIMENT 2**
Computer simulation
Can words be grouped into more than one category even when ambiguous words are added?
Results show that it takes fairly extreme conditions to make the algorithm perform poorly. Distribution based learning procedure performed well on input generated from simple, distributionally-defined artificial languages, even when presented with ambiguity.

EXPERIMENT 3
Computer simulation
But this time, used spontaneous, child-directed English
Is the proposed method as successful with linguistic data from real children as it was with the computer-generated input? Distribution-based simulation program did a better job categorizing the child-directed inputs than either baseline program.

EXPERIMENT 4
Computer simulation
Same as Experiment 3, only used substantially longer input files
Distributional program categorized words much more accurately than baseline, slightly more conservative than in Experiment 3.

EXPERIMENT 5
Does semantic information improve the performance of the distributional program? Yes. Even a small amount improved performance by leading to more correct group merges.

These results suggest that semantic and distributional analyses seem complementary. However, the authors admit that these experiments would need to be expanded to obtain more evidence. The results of Experiment 5 show that semantic information is useful, but the details of how it is best exploited remain to be worked out.