Word categorization from distributional information: Frames confer more than the sum of their (Bigram) parts

Toben H. Mintz [a,b,*], Felix Hao Wang [a], Jia Li [a]

[a] Department of Psychology, 3620 McClintock Ave., University of Southern California, Los Angeles, CA 90089-1061, United States
[b] Department of Linguistics, University of Southern California, United States

ARTICLE INFO

ABSTRACT

Grammatical categories, such as noun and verb, are the building blocks of syntactic structure and the components that govern the grammatical patterns of language. However, in many languages words are not explicitly marked with their category information, hence a critical part of acquiring a language is categorizing the words. Computational analyses of child-directed speech have shown that distributional information—information about how words pattern with one another in sentences—could be a useful source of initial category information. Yet questions remain as to whether learners use this kind of information, and if so, what kinds of distributional patterns facilitate categorization. In this paper we investigated how adults exposed to an artificial language use distributional information to categorize words. We compared training situations in which target words occurred in frames (i.e., surrounded by two words that frequently co-occur) against situations in which target words occurred in simpler bigram contexts (where an immediately adjacent word provides the context for categorization). We found that learners categorized words together when they occurred in similar frame contexts, but not when they occurred in similar bigram contexts. These findings are particularly relevant because they accord with computational investigations showing that frame contexts provide accurate category information cross-linguistically. We discuss these findings in the context of prior research on distribution-based categorization and the

  * Corresponding author at: Department of Psychology, 3620 McClintock Ave., University of Southern California, Los Angeles, CA 90089-1061, United States.
    E-mail address: tmintz@usc.edu (T.H. Mintz).

broader implications for the role of distributional categorization in language acquisition.

1. Introduction

Grammatical categories—e.g., noun, verb, adjective, etc.—are the building blocks that structure human languages and the units over which syntactic and morphological processes operate. Across typologically diverse languages, categories govern the ordering of words in sentences and the combinations of affixes and word stems. For example, in the sentence *Anna is kicking the ball*, the stem *kick-* occurs with the inflection *–ing* and follows the auxiliary verb *is* by virtue of being a verb, and in particular, a verb marked with present progressive tense and aspect. Because syntactic processes apply to categories, the present progressive morphosyntax does not need to be learned or represented, item by item, for each verb of English; rather, a word's status as a verb is sufficient for licensing its occurrence in this construction. Moreover, when hearing an unfamiliar word in this construction—e.g., *is lorping*— an English speaker can identify *lorp* as a verb stem, and then, by virtue of the category, knows a host of other operations and constructions available to the stem. For example, the speaker knows that the stem can be inflected with a past-tense morpheme to form *lorped*, and that it can be sequenced with a different auxiliary verb to form *will lorp*, and so on. In sum, categories provide language users with an efficient and powerful means of representing regularities in linguistic structures and processes.

However, words do not appear in utterances explicitly marked with category information, so there is a puzzle as to how speakers first acquire the knowledge that allows them to leverage the power of categories. Consider the previous *lorp* example: An English speaker can use her implicit knowledge of English syntax to identify the structural position of *lorp* in the sentence as the head of a verb phrase (VP), and thus categorize *lorp* as a verb. But novice English learners cannot call on this knowledge, as it is precisely the mapping of surface strings—e.g., *is lorping*—to syntactic structures—e.g., VP—that they are in process of learning. Learners must therefore be able to assign at least some words in an utterance to categories in order to be able to learn about the abstract grammatical patterns in their language in the first place. How do language learners take the first steps in assigning words to categories? What information could they use to initially categorize words?

In this paper, we provide new evidence that learners perform *distributional analyses* of the sequences of words in their input and form categories of words that appear in similar distributional contexts. For example, using distributional information, a learner could analyze the utterance *Can you lorp it?* and categorize *lorp* with other verbs, not by identifying the position of *lorp* as the head of a VP, but by categorizing it with other words that occur in similar contexts. This is because words in English that are immediately surrounded by *you* and *it* are almost exclusively verbs. While other studies have presented similar evidence (Mintz, 2002; Reeder, Newport, & Aslin, 2013), our findings are significant because they demonstrate that learners are especially responsive to a particular type of distributional pattern called a *frequent frame* (like the English *you_it* frame just mentioned), which has been shown computationally to be an especially accurate source of grammatical category information cross-linguistically (Chemla, Mintz, Bernal, & Christophe, 2009; Erkelens, 2009; Stumper, Bannard, Lieven, & Tomasello, 2011; Wang, Hohle, Ketrez, Kuntay, & Mintz, 2011; Weisleder & Waxman, 2010). Our study thus sheds light on the particular kinds of distributional patterns to which human learners attend, and their potential relevance in human language acquisition. Furthermore, some studies have only found evidence of distributionally-based categorization in situations where grammatical categories were also marked by converging sources of information, such as semantic or phonological information (Braine, 1987; Frigo & McDonald, 1998; Gomez & Lakusta, 2004; Smith, 1966). Counter to those findings, we show that learners can acquire categories solely from distributional information (Mintz, 2002, 2011; Reeder et al., 2013), and we propose an account that unifies the apparent discrepancies in prior research on distributional categorization.