Cognition 128 (2013) 13–25

Contents lists available at SciVerse ScienceDirect

Cognition

journal homepage: www.elsevier.com/locate/COGNIT

Word frequency, function words and the second gavagai problem

Jean-Rémy Hochmann *

SISSA, International School for Advances Studies, Cognitive Neuroscience Department, via Bonomea, 265, 34136 Trieste, Italy
Harvard University, Department of Psychology, William James Hall, 33 Kirkland Street, Cambridge, MA 02138, USA

ARTICLE INFO

Article history:
Received 20 December 2010
Revised 18 February 2013
Accepted 20 February 2013
Available online 2 April 2013

Keywords:
Function words
Word frequency
Word learning

ABSTRACT

The classic gavagai problem exemplifies the difficulty to identify the referent of a novel word uttered in a foreign language. Here, we consider the reverse problem: identifying the referential part of a label. Assuming "gavagai" indicates a rabbit in a foreign language, it may very well mean "*a* rabbit" or "*that* rabbit". How can a learner know whether rabbit is actually said "vagai", "gava" or "gavagai"? Here, we report evidence suggesting that infants can identify potential function words on the basis of their high frequency and avoid considering them when associating labels and referents. In three experiments, 17-month-old infants were first exposed to an artificial speech stream where frequent and infrequent syllables alternated (e.g., ... *gibuvokugifevodegita*...). Infants then saw a novel object and heard the repetition of a bisyllabic label consisting of one frequent and one infrequent syllable (e.g., *vomu*). The frequent syllable was the initial syllable of the label in Experiment 1 and the final in Experiments 2 and 3. We then presented infants with both the previous and now familiar object and a novel object. We asked whether infants would be more likely to orient first towards the familiar object when hearing a label with a *new frequent* and the *previous infrequent* syllables (e.g., *gimu*), or when hearing a label with a *new infrequent* and the *previous frequent* syllables (e.g., *vona*). Results suggest that the infrequent syllable was associated more strongly with the object, than the frequent one, only when the perceived position of the frequent syllable was constant all along the experiment.

1. Introduction

1.1. The second gavagai problem

Quine (1960) famously introduced the "gavagai" problem, whereby a child acquiring language is facing the problem of the indeterminacy of translation when trying to understand the meaning of a novel word. In Quine's metaphor, children are in the same situation as an adventurer witnessing an island native that would point at a running rabbit and say "gavagai". Does "gavagai" refer to the rabbit, its ears, its color, its running or something else? This problem has since received several solutions, involving the use of conceptual biases (Markman, 1990; Markman, Wasow, & Hansen, 2003), socio-pragmatic (Akhtar & Tomasello, 1996; Baldwin, 1991, 1993; Csibra, 2003; Hirsh-Pasek, Golinkoff, & Hollich, 2000; Sperber & Wilson, 2004) and syntactic cues (Bernal, Lidz, Millote, & Christophe, 2007; Brown, 1957; Gleitman, 1994; Naigles & Kako, 1993; Waxman & Booth, 2001, 2003). However, Quine's situation contains a second problem that has hardly been addressed. Assuming that the learner solved the classic gavagai problem and identified the referent, what part of speech actually refers to the referent? In other words, given that the island native intended to name the rabbit, is rabbit "gavagai" in his language? Or is it "gava", "vagai" or "gai"? This second gavagai problem is not anecdotal, as words, especially nouns, are rarely pronounced in isolation. Only 7–12% of child directed speech utterances consist of

* Address: Harvard University, Department of Psychology, William James Hall, 33 Kirkland Street, Cambridge, MA 02138, USA. Tel.: +1 6174962695.
E-mail address: jr.hochmann@gmail.com

0010-0277/$ - see front matter © 2013 Elsevier B.V. All rights reserved.
http://dx.doi.org/10.1016/j.cognition.2013.02.014

isolated words (Brent & Siskind, 2001; Christiansen, Allen, & Seidenberg, 1998; Fernald & Morikawa, 1993; Fernald & Simon, 1984). In particular, nouns in languages like English or Italian are usually associated with a determiner. Therefore, instead of just "rabbit", "gavagai" may well mean "*a* rabbit". Learners would thus need to strip the part that corresponds to the determiner, before pairing the noun and its referent.[1]

A solution to the second gavagai problem may involve early knowledge about a category of determiners, or more broadly of function words. If infant learners discriminate between function and content words and know that function words are not referential, they may correctly focus on content words when learning the label of a novel object or kind.

1.2. Early representations of syntactic categories

The question of whether young children and infants represent syntactic categories has been debated in the field of language acquisition. One view proposes that young children rely on item-based templates that are syntactically unanalyzed (Dabrowska, 2001; Tomasello, 2003). According to such item-based or usage-based theory of language acquisition, children initially store a certain amount of fixed sentence templates. This view predicts that children's initial production merely reflects their input and is not productive. A child that says "*eat pasta*" and "*want cookie*" would not necessarily say or understand "*eat cookie*" and "*want pasta*". Support to this theory mainly comes from corpus analysis of infants' spontaneous productions (Pine & Martindale, 1996; Tomasello, 2003).

This view is contradicted, however, by studies showing productive use of morphological and syntactic constructions, even leading to overgeneralization (Marcus et al., 1992). For example, in a classic study, Berko (1958) introduced 4-year-old preschool children with a novel animal that she called a *wug*: "this is a wug". She then showed two exemplars of the same animal, and asked children to complete her sentence: "these are two…". Children spontaneously answered "wugs!", generalizing the rule for plural formation in English to the novel word they had just learned in the singular form. Comprehension studies suggest that this ability initially develops between 20- and 24-months of age, well before production, and is further refined in the third year after birth (Kouider, Halberda, Wood & Carey, 2006).

Moreover, Lidz, Gleitman, and Gleitman (2003) showed that 3-year-olds already analyze the structure of simple sentences, and use that structure to constrain their interpretation of the sentence (see also Lidz & Gleitman, 2004). They studied learners of Kannada, a language where causative meaning can be marked in two ways: by a morphological affix on the verb or by a transitive construction, as it is classically done in English. In infants' and children's input, causative morphological marking is more frequently and more consistently associated with causative meaning than is the transitive structure. The usage-based theory would thus predict that young children should be better at interpreting the morphological marker construction than the transitive construction. Lidz et al. (2003) showed that this is not the case. Three-year-old Kannada learners based their interpretation of the meaning of a novel verb on the number of arguments (one or two) rather than on the presence or absence of the causative morphological marker. Thus, rather than using the meaning of template sentences to learn about the structure of their language, Kannada learners rely on a structural analysis to learn the meaning of novel verbs. In that respect, they behave as young learners of English (Arunachalam & Waxman, 2010; Naigles & Kako, 1993; Yuan & Fisher, 2009), Mandarin (Lee & Naigles, 2005, 2008), or Turkish (Göksun, Küntay, & Naigles, 2008).

Beyond lexical categories such as nouns and verbs, the debate has also revolved around the question of whether young children have functional categories such as determiners. In particular, researchers have asked whether young children show overlap in their use of determiners. For example, do infants who have learned "*a ball*" and "*the duck*" also say "*the ball*" and "*a duck*"? Or is their production item-based? Analyzing children's production corpora and using different criteria, Pine and colleagues (Pine & Lieven, 1997; Pine & Martindale, 1996) argued for the latter, while Valian and colleagues (Valian, 1986; Valian, Solt, & Stewart, 2009) argued for the former possibility. Yang (2010) recently proposed that the pattern of children's production could be explained by the general phenomenon that relatively few words are used very frequently, whereas most words are rarely used (a statistical distribution known as Zipf's law). Reanalyzing production data taking this distribution into account, Yang (2010) claimed that the data ultimately support the view that very young children are already equipped with productive syntactic rules.

Other evidence for the representation of syntactic functional categories consists in how these constrain word-to-world mapping. For example, Waxman and Booth (2003) showed that 14-month-old infants associate a novel word to an object kind if it is preceded by a determiner (*the blicket*) and to a property of the object if ending in *-ish* and followed by the pronoun *one* (*the blickish one*). Bernal et al. (2007) showed that 23-month-old infants map a novel word onto a novel action if this word appeared following a pronoun, but not if it followed a determiner (see also Brown, 1957).

In our opinion, experimental evidence thus favors the view that young children and infants as young as 14-months have syntactic categories, or at least precursors of these categories. How do they acquire or recognize these categories? Distributional analysis and phonological cues may play a role in this process.

1.3. Forming syntactic categories

A relatively successful strategy to classify words into nouns and verbs is to classify them according to the frames

[1] The second gavagai problem, we have to stress, cannot be reduced to the problem of segmentation, which has been the focus of much investigation in recent years (Aslin, Saffran, & Newport, 1998; Bonatti, Peña, Nespor, & Mehler, 2005; Buiatti, Peña, & Dehaene-Lambertz, 2009; Gómez, Bion, & Mehler, 2011; Mehler, Peña, Nespor, & Bonatti, 2006; Saffran, Aslin, & Newport, 1996; Shukla, Nespor, & Mehler, 2007 and many others). Even after one has segmented *gavagai* into distinct words such as *ga* and *vagai*, one still needs to identify which of these words refers to RABBIT.