Overextensions that extend into adolescence: Insights from a threshold model of categorization

Steven Verheyen (steven.verheyen@psy.kuleuven.be)
Eef Ameel (eef.ameel@psy.kuleuven.be)
Gert Storms (gert.storms@psy.kuleuven.be)
Department of Psychology, University of Leuven
Tiensestraat 102, B-3000 Leuven, Belgium

Abstract

The development of the meaning of common nouns may continue well past the early years of language acquisition (Ameel, Malt, & Storms, 2008, 2011; Andersen, 1975). The nature of this development is investigated through the application of a formal model to categorization decisions by children aged 7 – 13 years and adults. Children up till the age of 13 overextend categories. It is established that these overextensions are not due to differences in the manner in which children and adults organize items with respect to the target categories. Even the youngest children appear to know about the internal category organization that informs categorization. Rather, the overextensions are the results of children imposing too tolerant a criterion for category membership on this organization. With age children learn of this discrepancy and gradually reduce it to converge upon the conventional adult meaning.

Keywords: categorization; word learning; overextension; later lexical development; typicality.

Introduction

Young children face the incredible task of having to master a vast number of categories. Unsurprisingly, they commit a number of errors in the process. Overextensions, for instance, occur when the child is excessively liberal in allowing items into a category (e.g., referring to all liquid containers as *bottles*). An excessively conservative use of a category label is called an underextension (e.g., restricting the use of *bottle* to milk bottles only). These departures from the conventional use of categories by adults provide a window into the manner in which children acquire category meanings. Most contemporary research on the matter would have us believe that this window has closed by the time children reach the age of six. Indeed, it seems to be (implicitly) assumed that the meanings of the circa 14,000 words that the child has acquired by this age are no longer subject to change. The developmental trajectory of these words is generally not investigated in children older than six.

This assumption appears particularly prevalent for nouns that are used to refer to the common household objects that children encounter at a very early age. We know of only three studies that looked into children's use of category labels to denote such objects *after* the early years of language acquisition. Surprisingly, all three studies suggest that the meaning of these category labels continues to change well after children reach the age of six. Andersen (1975) asked English-speaking children aged 3, 6, 9, and 12 years to name drinking vessels. She found that it wasn't until the age of 12 that children's naming behavior resembled that of adults. Similarly, Ameel et al. (2008) asked Dutch-speaking children aged 5, 8, 10, 12, and 14 years to name a set of household storage containers. They found that at least up to the age of 14 did the naming behavior of the children change. With age the naming behavior grew closer to that of adults. Recently, Ameel et al. (2011) demonstrated that this result not only holds in a production task, but also in a comprehension task. They asked Dutch-speaking children aged 7, 9, 11, and 13 years to decide whether the household storage containers used by Ameel et al. (2008) did or did not belong to each of a number of suggested categories. The study revealed that children up to the age of 13 overextended the category labels. With age these overextensions gradually gave way and children's categorization behavior converged upon that of adults.

The aim of this study is to investigate the nature of the development that takes place after the initial years of language acquisition. To this end we will reanalyze the categorization data that were collected by Ameel et al. (2011). The proposed method of analysis is fairly new to the study of word learning in that it involves the use of a formal model. It is our belief that a formal approach to this subject matter promises to yield new insights that are difficult to arrive at using traditional methods only. The next section recapitulates the procedures that Ameel et al. (2011) used to obtain the data that we will reanalyze. The subsequent sections provide details about the model that we will employ.

Data

Categorization

Ameel et al. (2011) had 20 children aged 7, 21 children aged 9, 20 children aged 11, 20 children aged 13, and 36 adults (mean age 23;1 years) complete three categorization tasks. All participants were native speakers of Dutch. In each of the tasks the same set of 73 pictured storage containers were presented in a random order for categorization. In American English the objects would presumably receive the name *bottle* or *jar*. Ameel et al. (2008) showed that they were likely to be called *fles*, *bus*, or *pot* in Dutch. Each of these labels served as a target category in one of the three categorization tasks. The order of the target categories was randomized across participants. The participants were invited to indicate for every item whether the suggested category label was appropriate (1) or not (0). This procedure resulted in three separate binary item-by-person matrices, one for each target category.

Figure 1: Prototypical examples of the *fles* (left), *bus* (middle), and *pot* (right) categories.

Typicality

From Ameel et al. (2011) typicality judgments by 21 children aged 7, 21 children aged 9, 23 children aged 11, and 28 adults (mean age 18;9 years) were also available. None of them had participated in the categorization task. All of them were native speakers of Dutch. All participants were presented with the 73 pictured storage containers three times (each time in a random order) and asked to judge their typicality towards *fles*, *bus*, and *pot*. The presentation order of the target categories was randomized across participants. The participants were invited to indicate for every item how good an example it was of the suggested category label. They had a seven point rating scale at their disposal to provide their answers. A score of '1' was indicated to mean that the item was a *very poor* example of the category. A score of '4' was indicated to mean that the item was an *okay* example of the category. A score of '7' was indicated to mean that the item was a *very good* example of the category. To assist the children in providing their answer, these points were also accompanied by a schematic frowning face, a straight face, and a smiling face, respectively. The resulting typicality judgments for each category were averaged across each age group's members and z-transformed. The pictured storage containers in Figure 1 are considered prototypical exemplars of *fles*, *bus*, and *pot*, respectively, in all age groups.

A threshold model of categorization

Studies into developmental changes in word meaning are often framed in the context of prototype theory (e.g., Ameel et al., 2008; Kuczaj, 1982). The formal model that we will employ is therefore inspired by a version of prototype theory known as the Threshold Theory (Hampton, 2007). The Threshold Theory of categorization posits that prior to making a categorization decision the similarity between the item's representation and the category's representation is compared against an internal threshold. If the assessed similarity surpasses the internal threshold, it favors a positive categorization decision; otherwise it does not.

Prototype theory is often associated with abstracted summary representations of categories, but we will not make claims pertaining to this issue in our modeling endeavor. We will merely contend that the items become organized along a latent scale. This organization may result from the assessment of the similarity between the items' representations and an abstracted summary representation of the category, but doesn't necessarily have to. It may also result from similarity assessments involving other kinds of representations (e.g., an instantiated set of representative exemplars). Other sources of information may also inform the positioning of the items along the scale. If the target items are words, for instance, lexical familiarity might be of importance. The organization of the items along the scale will be derived from the categorization data. The resulting organization is thus to be understood with respect to a specific category. The model positions items that are seldom endorsed as members of the target category low on the scale. It positions items that are often endorsed as members of the target category higher on the scale.

At the heart of the model lies the assumption that the categorization scale is common to participants of all ages. Regardless of the age of the participants we thus assume that they organize the items with respect to the category in a similar manner. We expect participants to differ with respect to the categorization criterion they use, however. We assume that every participant employs a personal threshold in categorization, which is positioned along the same scale as the items. The distance between a participant's threshold and an item's position along the scale is assumed to inform the participant's categorization decision with respect to the item. If the item surpasses the threshold, the odds are that the participant will endorse it. The greater the distance between item and threshold, the greater the odds of a positive categorization decision. If the item does not surpass the threshold, the odds are that the participant will not endorse it. Under this circumstance, the odds of a negative categorization decision increase with the distance between item and threshold. The threshold position of every participant will be derived from the categorization data. The model positions thresholds of participants who endorse many of the items as members of the target category low on the scale. Many of the target items will surpass these thresholds. The model positions thresholds of participants who endorse fewer of the items higher on the scale. These thresholds will be surpassed by only a few of the items.

Formally, the application of the model to an item-by-person categorization matrix yields an item position estimate β_i for every item i and a threshold position estimate θ_p for

every person p. The categorization decision of every participant p towards every item i is treated as a Bernoulli trial. The participant's estimated threshold position θ_p and the item's estimated position β_i combine according to Equation (1) to yield the corresponding probability of a positive decision:

$$\Pr(x_{pi} = yes) = \frac{e^{(\beta_i - \theta_p)}}{1 + e^{(\beta_i - \theta_p)}} \quad (1)$$

Explanatory versions of the threshold model

In what follows we will try to capture the categorization decisions of different age groups with the model in Equation (1). Since this involves the assumption that the organization of the items with respect to the target categories is the same for all participants (regardless of age) this constitutes a test of the hypothesis that even the youngest children among the participants know about the internal structure of the categories, but do not (yet) agree with adults on the conventional extension. For this hypothesis to have any merit at all, the model will have to pass a goodness of fit test after it has been applied to the categorization data. It needs to be shown that the model, with its assumption of a single latent dimension for all, can account for the variability in the observed responses.

To allow for a second test of this hypothesis we propose to extend the model through linear components that introduce external information. While the model that is expressed in Equation (1) is merely a descriptive one that only yields estimates of items' and persons' positions along a latent scale, the introduction of external information into the model renders it explanatory. The external information allows one to account for the observed variability in item or person position estimates. On the item side we will subsequently express the β_i's as a linear function of each age group's typicality judgments:

$$\beta_i = \delta_0 + \delta_1 TYP_i + \varepsilon_i \text{ with } \varepsilon_i \sim N(0, \sigma_\varepsilon^2) \quad (2)$$

with δ_1 expressing the effect of the z-transformed typicality scores of a particular age group, and δ_0 taking the role of intercept. Random item variation is allowed to influence the prediction of the β_i's as well. This allows for the assessment of the magnitude of the explanatory power of each age group's typicality judgments. Typicality is perhaps the most direct measure of internal category structure that one can obtain. If we are correct in assuming that even the young children know about the categories' internal structure and use it for their categorization decisions, their typicality judgments should explain about as much of the variance in the item positions as the adult typicality judgments do.

It is further hypothesized that, with the knowledge about the internal category structure already in store, children learn about the appropriate region among the organized items to place their threshold criteria as they become older. Since Ameel et al. (2011) reported overextension for the categories under investigation we expect to see a more conservative placement of the threshold criteria with age. As age increases we thus expect a shift from low threshold values (positioned low on the latent scale and surpassed by many of the target items) to high threshold values (positioned higher on the scale and surpassed by fewer of the items). To evaluate this hypothesis, we will on the person side of the model express the θ_p's as a linear function of four dummy variables X_1, X_2, X_3, and X_4 (with the respective weights termed γ_1 to γ_4) that recode age group membership following the Helmert scheme in Table 1. The Helmert coding scheme allows for the detection of monotonic trends. It contrasts each level with the mean of the preceding levels. It is therefore ideally suited to test the hypothesis that with age participants impose a higher threshold on category membership. If γ_n is found to differ reliably from 0 it constitutes evidence that the n youngest age groups overextend the category with respect to the $(n+1)th$ group.

Table 1: Helmert coding of age levels.

Age	X_1	X_2	X_3	X_4
7-year olds	−1	−1	−1	−1
9-year olds	1	−1	−1	−1
11-year olds	0	2	−1	−1
13-year olds	0	0	3	−1
adults	0	0	0	4

Model estimation

All model analyses were performed in WinBUGS. For every combination of model and item-by-person categorization matrix five chains were run with 10,000 samples each, for a total of 50,000 samples from the posteriors of the models' parameters. The resulting distributions allow us to answer a number of substantive questions. We start, however, with a discussion of the fit of the models to the categorization data. Space restrictions only allow us to show the fit results for *fles*, but they are indicative of the results for the other two categories.

Results

Posterior predictive

The black dots in the right bottom panel of Figure 2 summarize the *fles* categorization data the models were fitted to. For each of the 73 items there is a dot representing the proportion of participants providing a positive categorization response. The items are ordered along the horizontal axis according to their mean β_i estimate across all samples. The other panels show the categorization data separated per age group. These panels maintain the same ordering of items that was adopted in the right bottom panel.

The plots demonstrate a developmental change in the extension of the category. With age participants become more conservative. An increasingly smaller number of items are endorsed by a large number of the participants. This aspect of the data corresponds to the pattern of overextension in children that was observed by Ameel et al. (2011). The plots add to this finding in that they suggest that the developmental change occurs in an organized manner. Indeed, all panels

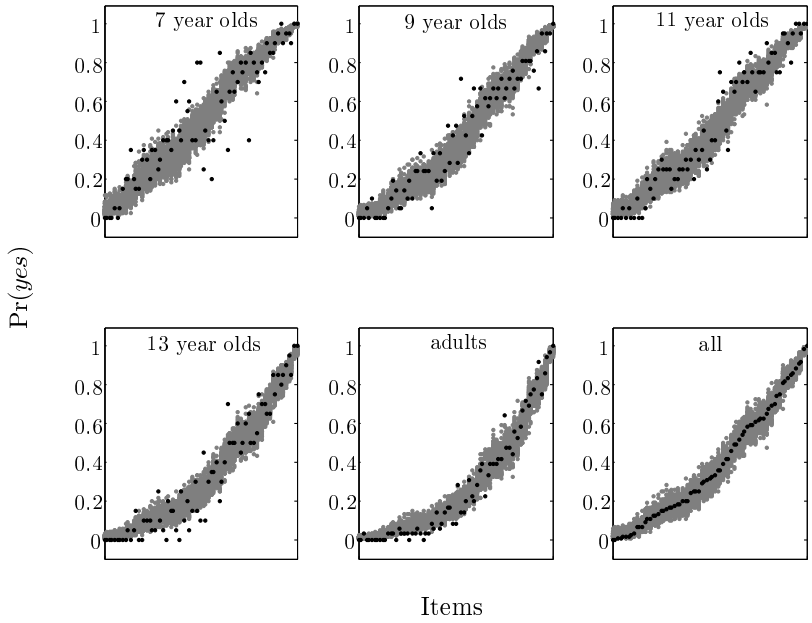

Figure 2: Posterior prediction of the threshold model for the *fles* categorization data, separated per age group and across all participants (bottom right panel). The black dots show the data (per item the proportion of participants providing a *yes* response). The gray areas show 100 posterior predictive samples of Pr(*yes*) for every item. The items are ordered along the horizontal axes according to the mean β_i estimate across all samples.

show an increase in the proportion of *yes* responses with the estimated position along the latent scale. While this increase is situated along the first diagonal in the youngest children's panel, it gradually takes on a different shape as the panels present the data from older participants. Nevertheless, it appears that in all age groups the increase roughly follows the latent scale extracted by the model. This suggests it is appropriate to assume that even the youngest children in the sample have a rather good idea about the organization of the items with respect to the target category and make use of this organization for their categorization decisions.

The gray areas in Figure 2 show 100 posterior predictive samples of Pr(*yes*) for every item. An adequate model should be able to describe the patterns in the data it was fitted to. The posterior predictive is a distribution over data that allows one to assess whether this is the case. It represents the relative probability of different observable outcomes after the model has been fitted to the data. From Figure 2 it becomes apparent that the model is able to pick up the most striking aspects of the data. The gray areas quite closely mirror the observed data. Admittedly, there are a number of observations - especially in the youngest age group - that are not well accounted for by the model, but as a first formal approximation of this kind of data, the model fares quite well. In the *Discussion* section we will elaborate on how the model might be extended to relate to the data more closely, but for now we will continue with the simple model and its explanatory versions and see what more we can learn from them.

Item variation

Each age group's typicality judgments proved reliable predictors of the item variation found in categorization. The Bayesian credibility interval (BCI) of parameter δ_1 from Equation (2) (i.e., the region around the mean δ_1 estimate that contains 95% of the mass of its posterior distribution) did not include 0 in any of the categories. For *fles* the BCI's were $[1.71, 2.40]$, $[1.51, 2.30]$, $[1.99, 2.41]$, and $[1.89, 2.36]$ for typicalities provided by children aged 7, 9, 11, and adults, respectively. For *bus* these BCI's were $[1.01, 1.57]$, $[1.12, 1.63]$, $[1.36, 1.72]$, and $[1.10, 1.58]$. For *pot* they were $[1.18, 1.99]$, $[1.60, 2.21]$, $[1.80, 2.32]$, and $[1.84, 2.31]$. More importantly, however, is the variance in the β_i estimates accounted for by the typicality judgments. Table 2 shows the percentage of variance accounted for (VAF) per category by each age group's typicality judgments. In each of the categories the VAF increases with age, but the magnitude of the increase is rather small: From 79% to 86% for *fles*, from 62% to 65% for *bus*, and from 80% to 85% for *pot*. These differences are most likely due to age related differences in the reliability of the typicality judgments. Indeed, Ameel et al. (2011) report an increase in these reliabilities with age. We therefore take the results to indicate that even the youngest children in our study know about the internal structure of the categories *fles*, *bus*, and *pot* and use it to provide categorization decisions. Typicality judgments arguably provide the most direct measure of a category's internal structure. Here, the typicality judgments by participants of different ages all accounted for the variation of items along the categorization

Table 2: Item variance accounted for by typicality.

Age	*fles*	*bus*	*pot*
7-year olds	.79	.62	.80
9-year olds	.81	.62	.81
11-year olds	.84	.64	.82
adults	.86	.65	.85

dimension about equally well. The item organization that the model extracts from the categorization data is reliably predicted by the internal category structure that judgments of typicality capture - even those provided by the 7-years olds.

Note that typicality has been taken to reflect item-category similarity (e.g., Hampton, 2007). To the extent that this assertion is true, the results for *bus* in Table 2 suggest that there are additional considerations that inform categorization: Part of the item variation is left unexplained by typicality. An investigation into other sources that govern categorization is beyond the scope of the current paper, but might easily be initiated by an inspection of the items with β_i's that are not well predicted by typicality.

Person variation

Figure 3 presents per category and age group boxplots of the mean θ_p's that were obtained with the exploratory model. The value of θ_p reflects the position of person p's threshold along the categorization scale, with larger values indicating positions further along the scale. For all three of the categories an age related shift in the employed threshold criteria is notable: With age the estimated θ_p values grow larger. Apparently, when children become older, they place their categorization thresholds further along the scales, thereby becoming more restrictive of the items they allow in the categories. In terms of the results from the previous section, it would appear that with age items need to be of increasingly high typicality (i.e., item-category similarity) to be endorsed as category members.

The results of the analyses with the model that incorporates age group membership through the Helmert coding in Table 1 support these informal observations following inspection of Figure 3. The majority of the BCI's of the parameters γ_1, γ_2, γ_3, and γ_4 were found not to include 0. This indicates that the contrasts that the dummy variables X_1, X_2, X_3, and X_4 constitute are reliable: The mean threshold criterion that is employed by the participants of a particular age group is indeed more conservative than the mean threshold criterion employed by the participants in the younger age groups. For *fles* the BCI's of γ_1, γ_2, γ_3, and γ_4 were $[.04, .74]$, $[-.18, .20]$, $[.13, .40]$, and $[.17, .34]$, respectively, indicating that all contrasts - except for the one comparing 11-year olds with 7- and 9-year olds - were reliable. For *bus* these BCI's were $[-.26, .34]$, $[-.12, .23]$, $[.06, .31]$, and $[.18, .34]$. For *pot* they were $[.02, .72]$, $[-.17, .23]$, $[.10, .39]$, and $[.32, .50]$. Note that none of the BCI's for γ_4 include 0, supporting the claim by Ameel et al. (2011) that ovextension errors for common nouns may persist into adolescence.

Age group membership accounts for 34% of the variance in θ_p estimates in *fles*, for 34% in *bus*, and for 48% in *pot*. Given that the change in threshold position with age is present in all three categories it is reasonable to expect a certain degree of correspondence between the threshold estimates obtained in the context of one category and the threshold estimates obtained in the context of another. However, if certain children within an age group could be found to display categorization behavior that resembles that of younger or older children *and* this behavior would generalize across categories, an even higher degree of correspondence is to be expected. The estimated threshold differences within an age group would then also carry valuable information. To investigate this possibility we correlated the θ_p estimates obtained in the context of one category with the θ_p estimates obtained in the context of another (r_{obs}) and compared the resulting correlation to a correlation that was obtained by treating all participants in an age group as if they were interchangeable (r_{age}). To achieve this the θ_p estimates within each age group were randomly permuted prior to computing the correlation between categories. This procedure was repeated for all combinations of categories and all 50,000 samples from the thresholds' posterior distributions. For the combination of *fles* and *bus* the mean of the resulting distribution for r_{obs} was .53 compared to .34 for r_{age}. For *fles* and *pot* the mean of r_{obs} was .55 compared to .41 for r_{age}. For *bus* and *pot* the mean of r_{obs} was .61 compared to .41 for r_{age}. The observed correspondence in threshold use is greater than what would be expected based upon age alone. Not only the θ_p differences between age groups are thus of importance. Within age groups the θ_p values reflect substantive differences as well. Clearly, the later years of language development know stable and pronounced inter-individual differences, much like the ones that are characteristic of the early years of language acquisition.

Discussion

The aim of the current study was to gain insight into the changes in common nouns' meanings after the early years of language acquisition through the use of a formal model. The proposed model offers a process account of the decreasing pattern of overextension observed in children aged 7-13 by Ameel et al. (2011). The results of the model analyses suggest that children and adults organize items in a similar manner with respect to the target categories. The items are organized along a dimension that closely resembles typicality. On this dimension the categorizers impose threshold criteria. When an item surpasses a categorization threshold the odds are that it will be endorsed. When the item does not surpass the categorization threshold the odds are that it will not be endorsed. Overextension is observed in children up till the age of 13 because they place their threshold criteria lower on the categorization dimension than adults do. It appears that the children have acquired sufficient knowledge about the categories' meanings to construct the dimension that underlies

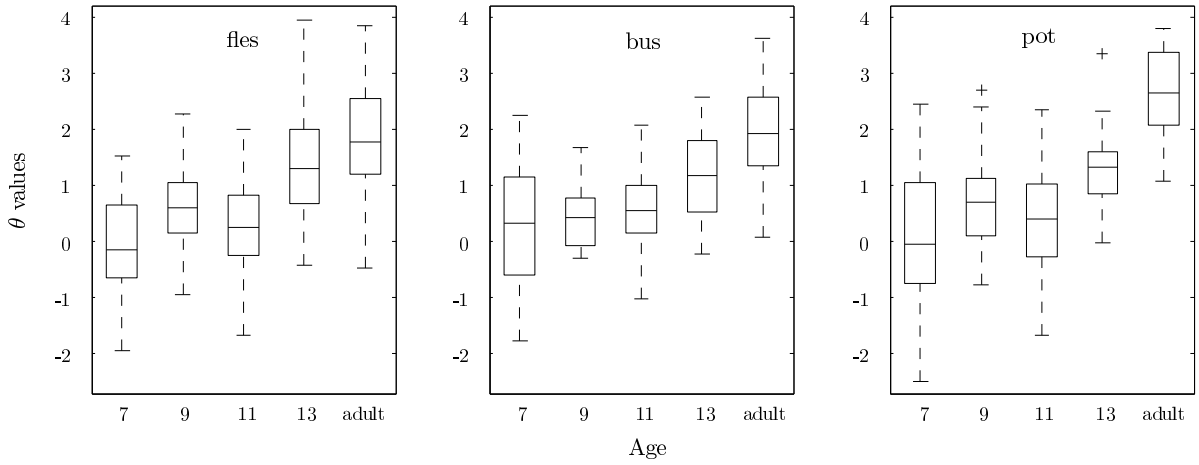

Figure 3: Boxplots of the mean θ_p estimates (across samples) per category and age group.

categorization, but do not yet agree with adults on the appropriate region to place the categorization threshold.

Part of the explanation for this observation probably lies in everyday language use, where a category label is primarily used to denote its prototypical exemplars. Items that reside at the category boundary are awarded the category label less often and might even be awarded a different label depending on the context (Clark, 1997). This makes it difficult for children to discover what the appropriate extension of the category is. Only with age, as children experience that their use of a category label does not always accord with that of adults, or when they are old enough to have this explicitly pointed out to them, can they start to shift their membership criteria in the direction of those used by adults. This process of gradually conforming to what is conventional in the language community is known as the principle of *conventionality* (Chouinard & Clark, 2003; Golinkoff, Mervis, & Hirsh-Pasek, 1994). According to the results of our model analyses it might very well be the main principle that governs the development of meaning after the early years of acquisition.

The model that we have forwarded is one of the first pertaining to word meaning development. As such it is subject to many improvements. In Figure 2, for instance, we saw that departures from the predicted pattern were most prevalent for the youngest children. These children might have only just acquired the meaning of the target categories, making their category representations still somewhat instable (Kuczaj, 1982; Thomson & Chapman, 1977). To accommodate this we are considering introducing noisy responding in the model. If initial instability of category representations is responsible for the departures in Figure 2, the noisy responding component should be more pronounced among the younger children. Incorporating manners to capture dependencies between items might also improve the fit of the model as it promises to identify chaining. Chaining is at work when children provide a different answer to an item than what is expected based on item-category similarity (i.e., typicality) because of the item's similarity to one or more other items. The categorization decision then not merely entails the comparison of item-category similarity against an internal threshold, but also becomes dependent upon the decision made for one or more other items. These dependencies might constitute an important aspect of the children's categorization behavior that is overlooked in the current version of the model.

References

Ameel, E., Malt, B. C., & Storms, G. (2008). Object naming and later lexical development: From baby bottle to beer bottle. *Journal of Memory and Language*, *58*, 262-285.

Ameel, E., Malt, B. C., & Storms, G. (2011). *Mowgli in the jungle of words: Comprehension and later lexical development.* Manuscript submitted for publication.

Andersen, E. S. (1975). Cups and glasses: Learning that boundaries are vague. *Child Language*, *2*, 79-103.

Chouinard, M. M., & Clark, E. V. (2003). Adult reformulations of child errors as negative evidence. *Journal of Child Language*, *30*, 637-669.

Clark, E. V. (1997). Conceptual perspective and lexical choice in acquisition. *Cognition*, *64*, 1-37.

Golinkoff, R. M., Mervis, C. B., & Hirsh-Pasek, K. (1994). Early object labels: The case for a developmental lexical principles framework. *Journal of Child Language*, *21*, 125-155.

Hampton, J. A. (2007). Typicality, graded membership, and vagueness. *Cognitive Science*, *31*, 355-384.

Kuczaj, S. A. (1982). Young children's overextensions of object words in comprehension and/or production: Support for a prototype theory of early object word meaning. *First Language*, *3*, 93-105.

Thomson, J. R., & Chapman, R. S. (1977). Who is "daddy" revisited: The status of two-year olds' overextended words in use and comprehension. *Journal of Child Language*, *4*, 359-376.