Cross-situational language learning: The effects of grammatical categories as constraints on referential labeling

Padraic Monaghan (p.monaghan@lancaster.ac.uk)
Karen Mattock (k.mattock@lancaster.ac.uk)
Department of Psychology, Lancaster University
Lancaster LA1 4YF, UK

Abstract

Infants learn to map words onto situations, even though there is a bewildering array of potential referents for each word in their environment. Previous studies of cross-situational learning have shown that learning correspondences between words and referents is possible, when all words refer to objects. However, in child-directed speech, the infants' primary input is more complex as it comprises multi-word utterances from different grammatical categories, some of which do not form word-object pairings. In study 1, we confirmed in corpus analyses of child-directed speech that utterances typically contain words from several different grammatical categories. In study 2 we confirmed that participants could still learn from cross-situational statistics when (1) the language also incorporated words that did not refer to objects, and (2) when the language additionally contained function words that marked the referring and non-referring words. Cross-situational learning is robust to grammatical categories in acquiring word-object pairings.

Learning to map words onto objects

How do infants acquiring their first language learn to map words onto referents in their environment? In 1960, Quine illustrated the complexity of this task by describing the situation of a field linguist trying to determine the meaning of the word "gavagai" spoken by a native speaker as he points to a rabbit running past. The referent for the word could be the rabbit, but it could also be the action of running, a patch of fur, the general beauty of the scene, or infinite other interpretations. Additional constraints must be available to the learner in order to correctly ascertain the intended referent. The developmental literature reports several candidate constraints that reduce the space of possibilities, such as the whole-object constraint (Markman, 1990), whereby infants seem to assume that the referent for a word is a whole object and not a part of it. However, such perceptual constraints are still not sufficient to confine the interpretation to a single referent as there are often situations where the child's environment contains several separate but whole objects. In the same paper, Markman (1990) reports other possible constraints that arise from the child's computation of statistical co-occurrences between words and objects in the environment that can assist in solving the "gavagai" problem.

Smith and Yu (2008) illustrated one such statistical constraint in action by indicating that 12-month old infants could learn the relationship between particular names and objects across multiple learning situations, known as "cross-situational learning" (Siskind, 1996). In their study, six nonsense words and six unfamiliar shapes were paired. For each learning trial, participants heard two words and viewed the two objects to which they referred, and had to learn which of the words referred to which of the objects. The probability of hearing a word and seeing the target object was therefore 1, but the probability of hearing a word and seeing another of the objects was .2. After training, infants were presented with two objects and heard the word paired with one of those objects, and were found to preferentially look toward the target object. The infants were shown to rapidly determine across multiple learning trials the co-occurrence between names and objects.

Yu and Smith (2007) conducted a similar study for adults, but examined the effect of the number of words and objects presented at any one time. When the number of words and objects presented at any one time was either two, three, or four, the participants learned better than by chance the relationship between the target name and the target object. As with the infant study, adults could learn the link between particular words and objects from co-occurrences across several learning situations even when there were several – up to four – possible referents for each word.

However, in both these previous studies the association between a particular word and object in each learning situation was perfectly represented, in that for every word spoken there was an object to which it referred. This has the consequence that means that in the learning situation multiple factors may have been contributing to learning. First, the cross-situational association between each word and each object across multiple instances was a contributor, as highlighted by Smith and Yu (2008). Second, learning of *one* of the word-object pairs could assist in learning the relationships between the other word and object (as in Akhtar, 2002). Hence, another influence on learning in these studies could be the mutual exclusivity constraint of names for objects (Houston-Price, Plunkett, & Harris, 2005; Markman, Wasow, & Hansen, 2003), whereby learning of a name for one object precludes the child from using the same name for another object. So, learning the cross-situational statistics could be boosted in that knowing the connection for one of the word-object pairs provides information about the referent for the other word(s) that the participant hears.

The cross-situational learning studies are useful abstractions from the real-life learning situations present when infants acquire knowledge of referents for words, yet they do not represent the natural situation in two potentially important respects. First, all the words in these situations have referents, whereas this is not the case in child-directed speech where only some of the words spoken to the child in each utterance are nouns referring to objects (see Yu & Ballard, 2007). Second, and relatedly, all the words in these situations refer to one of the objects in view, and so there are no cases where a word is spoken and a referent for the word is absent. If we incorporate these realistic features of child-directed speech into the cross-situational learning task then the mutual exclusivity constraint, and the assumption that every word has a referent, is not available to the language learner. In these cases, the role of cross-situational statistics as the sole driver of learning can be investigated. This, too, enables a stronger test of cross-situational learning under conditions that more closely resemble the natural-language situation.

In the first study, we investigated a large corpus of child-directed speech to determine the extent to which utterances consisted of words from more than one grammatical category. In the following corpus analysis, we were particularly interested in the co-occurrence of nouns and verbs in speech – in cross-situational learning only the nouns should be taken to identify with the object, though there are possibilities that a particular verb could also reliably co-occur with the object. In addition, we were also interested in the use of other content words alongside a noun for similar reasons – an adjective may reliably co-occur with a particular noun, which indirectly then may effectively co-occur with an object. We were also interested in the use of nouns with function words, such as “the” or “a” which occur frequently with nouns, and consequently could occur frequently with the object target for the noun. In these cases the probability of the object given the function word is high, but presumably the child would have to learn the non-specificity of the function word and disregard it as a potential label for an object.

Study 1: Corpus Analyses of Child-Directed Speech

Method

Corpus preparation

The corpus was taken from the English corpora submitted to the CHILDES database (MacWhinney, 2000). We selected all the adult speech spoken in the presence of infants, which comprised 5.7 million words in 1.3 million utterances. The corpus was automatically tagged by a parser with 95% accuracy (Sagae, MacWhinney, & Lavie, 2004).

Corpus analysis

We grouped the words in each utterance into different sets of grammatical categories. As we were particularly interested in the co-occurrence of words with referring nouns, we selected only those utterances that contained at least one noun. There were 608,008 such utterances. We then analysed these utterances in terms of the number of distinct nouns they contained – this addresses the question of the relative proportions of utterances that children are exposed to with either one or several nouns. Pronouns were not considered in the analyses, as once the child had acquired the pronouns then they could not be interpreted as referring to objects in the child’s environment, and they have a different distribution in that they tend not to be marked by a function word as with common nouns, which becomes relevant for the following analyses on the role of function words.

For each of these utterances containing at least one noun, we also counted the number of verbs they contained. Verbs are a frequent word category and we hypothesized that nearly all utterances would contain at least one verb. Certain verb tokens may be used in specific situations, and so could provide misleading variable information to the child about the identity of the referring word in speech, such as the verb “watch” that occurs in the same utterance as more than 20% of occurrences of “television” in the CHILDES corpus.

We also measured the other words that would not function as referents in the speech in each utterance, divided into two general categories. First, content words such as adjectives and adverbs which, as with the verbs, are varied in their usage and so could be misleading in terms of their link to particular objects in the environment. Second, we measured occurrence of function words, comprising articles, numerals, conjunctions, and prepositions, that are likely to occur with a referring noun, but, unlike specific verbs, adjectives, and adverbs, they occur frequently in speech and so the child has to learn that the lack of variation in these words’ usage indicates they are poor candidates for mapping words to objects.

Not all nouns in child-directed speech are used to refer to an object in the child’s environment (for instance, “parliament” and “senate” occur once each in the CHILDES corpus, yet it is unlikely this is used to refer to an object in the child’s immediate environment). However, the results do provide an indication of the potential co-occurrence of nouns with non-referring words in child-directed speech. Aslin, Woodward, LaMendola, and Bever (1996) instructed parents to teach 12-month old infants a novel word, and they tended to use the word in multi-word utterances including verbs and function words, suggesting that the general pattern of utterances for learning word-object pairings is not qualitatively distinct from general patterns of parent-child discourse in terms of the range of grammatical items used.

Yu and Bannard (2007) provided some highly-detailed analyses of two small corpora of child-directed speech (281 and 321 utterances, respectively) when parents were speaking to their children in the presence of various toys in the child’s immediate environment. They found that the utterances were generally grammatically complex, and that children would have to learn to disregard certain misleading associations between words and objects, and that a computational model maximizing the likelihood of descriptions to match environmental objects could effectively learn the associations between referring words and objects from these corpora. Our corpus analyses extend these results by providing a perspective from a substantially larger corpus as to the extent of the complexity of words of different grammatical categories present in child-directed speech.

Table 1. Proportion of utterances containing at least one noun in child-directed speech corpus.

Number of Nouns	Number of utterances	Proportion of utterances
1	419469	0.69
2	135044	0.22
3	35456	0.06
4	10611	0.02
5	3796	0.01
6+	3632	0.01

Table 2. Proportion of the utterances containing one noun and zero or one or more other words of each grammatical grouping.

Verbs	Adj/Adv	Function	Proportion
0	0	0	0.10
0	0	1+	0.21
0	1+	0	0.01
0	1+	1+	0.06
1+	0	0	0.02
1+	0	1+	0.13
1+	1+	0	0.07
1+	1+	1+	0.40

Results and Discussion

Table 1 shows the number of utterances containing different numbers of nouns. The results indicate that the majority of utterances contain only one potential referring noun, however 31% of utterances contained more than one noun, and 10% contained three or more. Whereas the learning situation of several nouns occurring in an utterance is frequent (the learning situation of Smith & Yu, 2008, for instance is reflected in 22% of utterances that contain two nouns, and 2% of utterances reflect the learning situation of Yu & Smith, 2007, where four nouns are present), the typical exposure the child experiences is of utterances containing just one word as a referent.

Table 2 indicates the utterances that contain just one noun and the number of other words of each of the other groupings of grammatical categories in the corpus – verbs, adjective and adverbs, and function words. 10% of the utterances containing one noun contained no other words – these were potentially referring nouns that were spoken in isolation and may have provided ideal information about the pairing of the word with an object in the child's environment. However, the most common occurrence was when a single noun is also accompanied by at least one other word that could not operate as a potential referring word. 62% of utterances containing one noun also contained at least one verb, 54% contained at least one adjective or adverb, and 80% contained at least one function word. 40% of the utterances containing just one noun also comprised at least one verb, at least one adjective or adverb, as well as at least one function word, indicating that most child-directed speech utterances were grammatically complex.

These proportions were maintained when all the utterances containing one or more nouns (and not only those containing a single noun) were considered. 8% of these utterances contained only nouns, so 92% consisted of at least one noun and at least one other grammatical category word. 45% of utterances containing at least one noun contained at least one of all three of the other categories – one or more verbs, adjectives/adverbs, and function words.

The corpus analyses confirmed that children are often exposed to situations where several nouns, potentially referring to objects in the environment, occur in the same utterance. However, far more frequently, children hear an utterance that contains several words other than the noun. So, the task of cross-situational learning requires learning which of the words in a multi-word utterance may relate to an object in the environment and which do not. The words to be rejected are either those that co-occur with a particular object, such as verbs, adverbs, or adjectives, but are not paired with a particular object in the environment, or those that co-occur reliably with an object but also co-occur with many other objects, such as the function words).

Our results therefore confirmed the small corpus analyses of Yu and Bannard (2007) and showed that the complexity of child-directed speech extended to a larger corpus more representative of the variety of input to which the child is exposed. Our second study tested whether cross-situational learning is possible when learners hear object labels alongside a range of other words that did not relate to objects. The study was performed on an adult population to determine whether the referring and non-referring words could be isolated in a language learning task. Additional studies on infants would enable the results to be generalized to the language acquisition process, but previous studies of artificial language learning have indicated similar qualitative patterns of results in adult and infant studies (e.g., Smith & Yu, 2008; Yu & Smith, 2007).

Study 2: Cross-situational learning task

Method

Participants

The participants were 48 undergraduate students from Lancaster University. There were 13 male and 35 female participants, with mean age 19.5 years (range 18-24 years), and 24 participants were randomly assigned to each of the two conditions.

Materials

From the corpus analyses, it was clear that the majority of utterances to which children are exposed contain both a noun and a verb. The "noun-verb" condition incorporated this fact into the language learning task. For this condition, we selected six geometric shapes printed in black on a grey background, taken from Fiser and Aslin's (2002) study.

There were 12 nonsense two-syllable words spoken by a female voice in a neutral tone: *jeelow, pakrid, rakken, makkot, fooglow, shellbye, vinnoy, bimdah, zawyer, trepier, haagle, and wiertat*. For each participant, six of the words – the referring words – were randomly paired with one of the shapes, and the other words formed the non-referring word set. The randomization was performed to avoid any possible effects of preference for certain words describing certain shapes (see, e.g., Westbury, 2004).

In the "function word" condition, two additional nonsense words were used, *tha* and *noo*. The function words were initially randomly paired with either the category of referring words or the non-referring words. Figure 1 shows an example of a learning situation from the "function word" condition. The participant hears four words and sees two pictures, and has to learn that one of the words ("makkot") refers to the picture on the left. Another word not heard in this learning situation refers to the picture on the right. The word "tha" is the function word indicating the referring word. "Pakrid" is the non-referring word, and "noo" is the function word indicating the non-referring word.

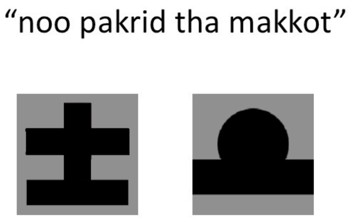

Figure 1. An example of a learning situation for one trial in the "function word" condition.

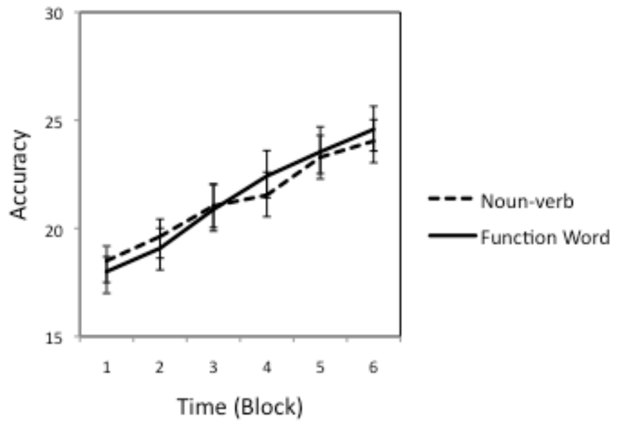

Figure 2. Accuracy for "noun-verb" and "function word" conditions of cross-situational learning across the six training blocks. 15 is chance level.

Procedure

In each trial, participants heard a sentence and viewed two pictures. In the "noun-verb" condition, one of the words was selected from the referring word set and the other was taken from the non-referring word set. The picture that was paired with the referring word appeared on the screen along with one of the other five pictures. The "function word" condition was identical except that the sentence comprised four words, the referring function word, the referring word, the non-referring function word, and the non-referring word. The function words always occurred immediately before the referring or non-referring word, but the order of the referring and non-referring word was counterbalanced.

Each trial began with the two pictures appearing on a computer screen, 500ms later, the sentence began. The referring/non-referring words lasted 500ms each, and the function words lasted 250ms. The participant was instructed to press the "1" key if they thought the sentence described the left picture, and the "2" key if the sentence described the right picture. 1000ms after the participant's response the next trial began. The order of the pictures (left/right) was counterbalanced. No feedback was given as to the participant's accuracy.

Accuracy of judgments was measured for every 30 blocks, in which each referring word appeared with its target picture 5 times. We also recorded reaction times of the responses, timed from the offset of the final word in the sentence.

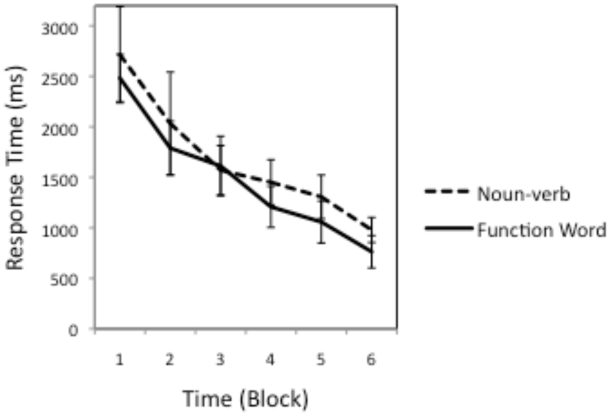

Figure 3. Response times for the "noun-verb" and the "function word" conditions of the cross-situational learning task.

Results and discussion

For the accuracy of responses, Figure 2 shows the results for the "noun-verb" and the "function word" conditions. A repeated measures ANOVA with block (1 to 6) as within subjects factor and condition ("noun-verb" or "function word") as between subjects factor was performed. There was a main effect of time, $F(5, 230) = 28.74$, $p < .001$, indicating that responses became more accurate with time.

There was no significant main effect of condition and no significant interaction between time and condition, both *F* < 1, indicating that accuracy was at a similar level in both conditions and improved at a similar rate.

When compared to chance level of 15 from 30 in each block, both conditions resulted in performance significantly better than chance from the first block, all $t(23) > 4.37$, all $p < .001$. In both conditions, participants were able to learn the mapping between the referring word and the target picture quickly, despite the presence of words that did not refer to any pictures, as well as the presence of a foil picture that was not referred to in the learning trial.

Reaction times are shown in Figure 3. A repeated measures ANOVA with time as within subjects factor and condition as between subjects factor resulted in a similar pattern of effects as for the accuracy measures. There was a main effect of time, $F(5, 215^1) = 27.76$, $p < .001$, indicating that responses became quicker with time, and there was no significant main effect of condition and no interaction between time and condition, both $F < 1$. As with the accuracy measures, the additional complexity of the speech containing two additional words for the "function word" condition did not impede responses compared to the "noun-verb" condition.

General Discussion

There is substantial complexity in the situation facing the child in learning the link between words and their referents in the environment. This complexity is certainly present in the world, in terms of the variety of possible objects, object parts, actions, and emotions that surround the child. Yet, the complexity is also present in the language itself, in that only some of the words that the child hears in each utterance have potential referents in that environment. Our corpus analyses provide a snapshot of the proportion of utterances containing words that the child must learn are not candidates for word-object pairings. Yet, as Yu and Bannard (2007) point out, the low likelihood of certain words occurring only with certain objects can cause many of these words to be disregarded, such as the frequent and diverse usage of function words.

Yet, other categories of words, such as certain pairs of nouns, or certain noun-verb or adjective-noun pairings, may be strongly co-occurrent in the corpus causing difficulties in forming the object-word pair. By chance, for instance, in Yu and Bannard's (2007) corpora, "eye" and "bird" were both highly associated with the appearance of a bird in the child's environment. Previous studies of learning from cross-situational statistics have indicated that, when all the words present are paired with one object each then learning can occur (Smith & Yu, 2008; Yu & Smith, 2007), yet learning that certain words may not have referents is a more realistic reflection of the situation that the child faces in language acquisition.

As in the experiments conducted by Smith and Yu (2008) and Yu and Smith (2007), learning in our study could only take place as a consequence of determining the associations between particular words and pictures. However, we have additionally indicated that the learning of these associations is robust against the presence of additional information that may have obscured the linking between the referring word and the picture in the form of words that did not have a referent in the "noun-verb" condition, and in the presence of additional words that co-occurred with all pictures in the "function word" task. We have shown that cross-situational learning is therefore sensitive to the mutual dependence of one word with a picture and does not occur only under circumstances where every word has a referent. The "function word" condition illustrated that the function words which were always present with each picture but did not provide information about the referent did not interfere with learning – there was no detriment in learning compared to the noun-verb condition where only two words were present.

We have also demonstrated that cross-situational learning is not dependent upon learning based on mutual exclusivity. In previous studies of cross-situational language learning, determining the mapping between one of the words and its referent provides additional information about the referent for the other words in the learning situation. In our design, participants had to learn that (at least) one of the words and one of the pictures provided no useful information for forming the mapping between the referring word and its referent.

Though the natural language situation is more complex than the small-scale tasks employed in these laboratory tests of cross-situational learning, this complexity may feasibly facilitate learning the word-object mappings. If the child can learn not only that an article such as "the" or "a" not only does not independently pair with a referent but also that it generally precedes a noun that can be paired with a referent, then the language internal structure may assist in constraining the possible mappings available between words and objects (see Yu, 2006, for preliminary work on grammatical category information constraining the mappings). In our "function word" experiment, the function words provided additional information about which of the other words was the referring word. This confluence of word-word and word-object associations may have boosted learning. The extra complexity of four words, only one of which was a referring word, in the "function word" condition did not produce a detrimental effect on learning compared to the "noun-verb" condition. This absence of impact may have been because the language internal information provided additional constraints on the locus of the word-object mapping.

Conclusion

Learning to pair words to objects in language acquisition is a difficult task due to the enormous number of possibilities for forming links between words and objects in

[1] Four participants' data was not available for the response time analysis due to problems in recording.

the environment. We have confirmed that the majority of utterances in child-directed speech contain words that have no referents. Incorporating these natural language characteristics into a study on cross-situational statistical learning indicated that participants could still form word-object associations even when there were several words in each utterance that related to no objects in the learner's environment. These natural language properties preclude the effective use of strategies such as mutual exclusivity to learn the associations. We contend that these language properties that introduce extra complexity also generate additional constraints on the language that may indeed promote the child's language learning.

Author Note

Thanks to Jenny Ball of Lancaster University for lending her voice for recording the stimuli.

References

Akhtar, N. (2002). Relevance and early word learning. *Journal of Child Language, 29*, 677-686.

Aslin, R.N., Woodward, J., LaMendola, N., & Bever, T.G. (1996). Models of word segmentation in fluent maternal speech to infants. In J.L. Morgan & K. Demuth (Eds.), *Signal to Syntax: Bootstrapping from speech to grammar in early acquisition (pp.117-134)*. Mahwah, NJ: Lawrence Erbaum Associates.

Fiser, J. & Aslin, R.N. (2002). Statistical learning of higher-order temporal structure from visual shape-sequences. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 28*, 458-467.

Houston-Price, C., Plunkett, K., & Harris, P. (2005). 'Word-learning wizardry' at 1;6. *Journal of Child Language, 32,* 175-189.

MacWhinney, B. (2000). *The CHILDES Project: Tools for analyzing talk. Volume 2: The database, 3rd edition*.

Markman, E.M. (1990). Constraints children place on word learning. *Cognitive Science, 14*, 57-77.

Markman, E.M., Wasow, J.L., & Hansen, M.B. (2003). Use of the mutual exclusivity assumption by young word learners. *Cognitive Psychology, 47*, 241-275.

Quine, W.V.O. (1960). *Word and object.* Cambridge, MA: MIT Press.

Sagae, K., MacWhinney, B., & Lavie, A. (2004). Automatic parsing of parental verbal input. *Behavior Research Methods, Instruments and computers, 36*, 113-126.

Siskind, J.M. (1996). A computational study of cross-situational techniques for learning word-to-meaning mappings. *Cognition, 61*, 39-61.

Smith, L. & Yu, C. (2008). Infants rapidly learn word-referent mappings via cross-situational statistics. *Cognition, 106*, 1558-1568.

Westbury, C. (2004). Implicit sound symbolism in lexical access: Evidence from an interference task. *Brain and Language, 93*, 10-19.

Yu, C. (2006). Learning syntax-semantics mappings to bootstrap word learning. In R. Sun (Eds.) *Proceedings of the 28th Annual Conference of the Cognitive Science Society* (pp. 924-929). Mahwah, NJ: Lawrence Erlbaum Associates.

Yu, C. & Ballard, D.H. (2007). A unified model of early word learning: Integrating statistical and social cues. *Neurocomputing, 70*, 2149-2165.

Yu, C. & Smith, L. (2007). Rapid word learning under uncertainty via cross-situational statistics. *Psychological Science, 18*, 414-420.