Did I say dog or cat? A study of semantic error detection and correction in children

Abstract

While naturalistic studies of spontaneous speech suggest that young children can monitor their speech, the mechanisms for detection and correction of speech errors in children are not well understood. In particular, there is little research on monitoring semantic errors in this population. This study provides a systematic investigation of detection and correction of semantic errors in children between the ages of 5 and 8, as they produced sentences to describe simple visual events involving nine highly familiar animals (the *moving animals* task). Results showed that older children made fewer errors and corrected a larger proportion of the errors that they made than younger children. We then tested the prediction of a production-based account of error monitoring that the strength of the language production system, and specifically its semantic-lexical component, should be correlated with the ability to detect and repair semantic errors. Strength of semantic-lexical mapping, as well as lexical-phonological mapping, was estimated individually for children by fitting their error patterns, obtained from an independent picture naming task, to a computational model of language production (Foygel & Dell, 2000). Children's picture naming performance was predictive of

their ability to monitor their semantic errors, above and beyond age. This relationship was specific to the strength of the semantic-lexical part of the system, as predicted by the production-based monitor.

Keywords: speech errors; error repair; speech production

Introduction

Although it was once believed that pre-school children had little reflective awareness of their mental states (e.g. Piaget, 1976), evidence from observational and diary studies suggests that children are able to self-correct errors in word production almost as soon as they are able to speak (Clark, 1978; Jaeger 1992, 2004; Stemberger, 1989). Consistent with these claims, Levy (1999) showed that 2-3 year-old children could often respond appropriately to requests for clarification of what they had just said. Sometimes, though not always, they were able to repair their speech errors in response. Levy suggested that, even at this age, children have access to a speech-monitor capable of detecting and repairing errors in spoken output. A few studies have reported that self-repair abilities gradually develop and grow in pre-school children. Rispoli (2003) showed that the ability to respond to and replace grammatical errors in spoken language improved between the ages of 2 and 4 years. Importantly, he claimed that monitoring ability improved in line with a child's grammatical development. Similarly, Jaeger (2004) showed that the proportion of self-corrected errors in phonological, lexical, and syntactic categories increased in children between the ages of 1 and 5. She suggested a monitoring

process that develops over a span of time without reaching the level of the adult monitor by age 5 (Jaeger, 2004, p. 82).

However, all of these studies used a naturalistic approach in which evidence of monitoring ability was derived from children's spontaneous speech at home or in the classroom (e.g., Evans, 1985; Peets, 2009). Very few studies have used a structured task to investigate children's ability to monitor their speech. An exception is the work of Sasisekaran and Weber-Fox (2012) who showed that children's ability to monitor spoken recordings for the presence of particular phonemes increased steadily between the ages of 7 and 13 years. Nevertheless, Sasisekaran and Weber-Fox did not examine monitoring of self-produced speech errors. While observational studies have the advantage of capturing children's behavior in their natural environment, they have certain limitations. (1) Differences in the amount and content of speech that is produced by each child make group comparisons difficult. (2) The target utterance is not always clear to the investigator. Knowing the identity of the target is generally not a problem with syntactic and phonological errors because, for example, "I goed" (target: went) and "fiss" (target: fish) are not acceptable or meaningful utterances in English. However, unless the referent is known to the addressee (e.g., it is in sight), semantic errors can easily go undetected; if a child says "I saw a doggy", it is hard to verify whether the

child has indeed seen a dog, or whether he or she meant to name a different animal such as a “cat”. This may be the reason why the literature contains many more reports of how children detect and correct their phonological and syntactic as opposed to their lexical-semantic errors. (3) In unstructured conversations, unless the data collector knows a child's current productions intimately, it is easy to confuse knowledge errors (i.e., errors where the child does not know that a cat is not a dog) with speech errors (i.e., slips where the target word is known to the speaker, but fails to be produced on a given instance; Reason, 1990). For example, "goed" and "fiss" should only count as genuine speech errors if, most of the time, the child uses the words "went" and "fish" correctly. Our goal in this paper is thus to provide a systematic study of detection and correction of semantic errors in children between 5 and 8 years of age. Age 5 was chosen as the lower limit for two reasons: (1) to minimize knowledge errors for the materials used in our experimental task; (2) because most studies of self-correction of speech errors in children have focused on children before age 5 (e.g., Jaeger, 2004; Levy, 1999; Rispoli, 2003), with little information about how the monitor continues to develop past this age.

We used a child-friendly version of a task used by Nozari, Arnold and Thompson-Schill (2014) that was successful in eliciting a large number of lexical-

semantic errors in adult speakers. Children were asked to watch simple events involving cartoon animals as they changed positions on a computer screen, and to describe what they saw (e.g., "The dog goes above the cat. The lion and the cat go below the monkey."). There were nine different cartoon animals, whose names were repeated in various sentences throughout the experiment, thus giving rise to competition (e.g., Schnur, Schwartz, Brecher, & Hodgson, 2006) and semantic errors (e.g., "dog" for the target "cat"). At the beginning of the experiment, the children were told to correct any error that they noticed, but were not prompted to do so on individual trials. This task, which we refer to as the *moving animals* task, made it possible to capture children's spontaneous detection and correction abilities during production of sentences that describe meaningful events. Critically, the task has more ecological validity than other paradigms used to assess monitoring abilities such as phoneme monitoring, and it reflects the challenge that children face in everyday life of planning and sequencing words during sentence production better than single-word picture naming tasks. At the same time, it is structured enough to allow for systematic exploration of similarities and differences between individuals, and for clearly specifying target utterances against which the child's performance can be evaluated.

This design allowed us: (1) to investigate whether children's ability to detect and correct their semantic errors for familiar words improves during the early years of elementary school; (2) to investigate whether increased ability to detect and correct errors is accompanied by a decrease in the number of errors that children make as they get older; (3) most importantly, it allowed us to investigate the nature of the system that children use in error monitoring. We elaborate on this last point below.

Relationship between the maturity of production and monitoring systems

Karmiloff-Smith (1986) provided evidence that the ability of 4-12 year-olds to detect and correct their speech errors far exceeded their explicit metalinguistic awareness of why there was an error. This finding was among the pieces of evidence taken to argue for a monitoring system that operates independent of conscious comprehension (for a review see Nozari, Dell & Schwartz, 2011; Postma, 2000). Nozari et al. (2011) proposed a new theory, called the conflict-detection theory of monitoring, in which error detection depends on the internal dynamics of the production system. When activating a word, other words (competitors) that share semantic and phonological features with that word also become activated (e.g., Dell, 1986; Rapp & Goldrick, 2000; See Dell, Nozari, & Oppenheim, 2014 for a review). When these competitors are highly activated, there will be more conflict with the target word for selection, and the chance of a slip increases. In the model, this conflict generates a signal that is translated by an executive center (most likely the

anterior cingulate cortex) into an error signal that causes the speakers to stop and revise their utterances (Nozari et al., 2011). As conflict and the subsequent error signal are generated automatically, part of the monitoring process can be completed before conscious awareness of the exact nature of the problem.

Critically, the stronger the production system, the greater will be the difference between the amount of conflict on error and correct trials. A neurotypical native adult speaker of a language often experiences low conflict during the production of a word such as "cat". So, on occasions when there is competition with other words, this conflict is highly predictive of an upcoming error. On the other hand, when the production system is weak, either due to immaturity or brain damage, conflict is much higher on all trials, thus making it a weak signal for error detection. This is similar to how one responds to a smoke detector alarm. With the right level of sensitivity, it is a perfectly reliable indicator of a serious problem. However, if it goes off every time one fries vegetables, it has much less reliability in signaling a genuine problem. Thus, Nozari et al.'s model predicts that as the production system matures, conflict will discriminate better between error and correct trials, and the quality of error detection will improve.

Critically, the model makes a very specific prediction: improvement in detection of each error type depends directly on the maturation of the specific part of the speech production system from which that error type arises. Thus, detection of a semantic error depends on how well the semantic-lexical part of the production system works. To illustrate this point, we provide a brief overview of the language production system using the model of Foygel and Dell (2000) that serves as the

underlying framework for Nozari et al.'s (2011) error detection model, and show how the strength of the different parts of the production system can be quantified using this model.

Insert Figure 1 about here

According to Foygel and Dell (2000), word production has two stages: lexical selection and phonological encoding. Lexical selection starts when the units representing a target word's semantic features are given a jolt of activation that then spreads throughout the network. The semantic features activate the lexical representation of the target word (at the "word" level in the bottom panel of Figure 1) and, potentially, the lexical representation of other words that share some of those semantic features. The most highly activated lexical node is chosen after a fixed time period at which point lexical selection is complete. The second part of word production (phonological encoding) starts with the selected lexical node at the word level receiving a new jolt of activation that sends activation to its associated phonemes. The phoneme layer consists of slots for onsets, vowels, and codas, and produces single-syllable consonant-vowel-consonant (CVC) words as output (see

Figure 1). Because of feedback, these phonemes activate lexical representations of other words that share some of the activated phonemes. Those lexical representations, in turn, activate new phonemes that belong to them. After a further fixed time period, the most highly activated phoneme in each of the onset, vowel, and coda clusters are selected and combined to form a whole word.

It is possible to simulate individual differences in naming by reducing either the connection weights between the semantic and lexical layers (the semantic or S parameter) or the connection weights between the lexical and phonological layers (the phonological or P parameter) or both (e.g., Dell et al., 1997; Nozari et al., 2010). These two parameters can be varied independently to simulate differences in the number and type of errors that individuals make. A reduction in the strength of the semantic parameter would lead to the production of a relatively large number of semantic errors such as dog (target: cat). A reduction in the strength of the phonological parameter would lead to a relatively large number of phonologically related nonword errors such as dat (target: cat), and some increase in the number of phonologically-related word errors, also known as formal errors, such as cap (target: cat). Why does the strength of the parameters matter? Because the process of activating nodes is noisy at both the word and the phoneme layers. small amounts of noise are added to the input that each node receives from other nodes that send

activation to it. When S and P are strong, noise has little effect. On the other hand, when these parameters are weak, activation of nodes is dominated by noise, making production much more error prone, similar to what happens following brain damage.

Nozari et al. (2011) used the basic framework of Foygel and Dell's model, and the ratio of connection weight to noise to deduce mechanisms for how speakers may monitor their own production. If the connections between semantic features and a target word are strong (as determined by the model's S parameter), selection is clean and easy. If not, competition and conflict between target and competitors will be consistently high, making for a poor error detection signal. Thus, a stronger semantic-lexical part of the system, as indexed by a larger S parameter in the model, predicts better detection of semantic errors (Figure 1). The strength of the connections between words and phonemes is indexed by the model's P parameter, and, according to the conflict-detection theory, its strength should be indicative of the detection of phonological, and especially nonword errors, but not semantic errors. Nozari et al. (2011) confirmed this prediction in a population of individuals with post-stroke aphasia. The current study tests this prediction in children.

In this study, we used an independent picture-naming test to estimate the strength of S and P parameters for each child. The picture-naming test had a diverse

vocabulary, on which children made a variety of errors, including semantic and phonological errors. The S and P parameter strengths were estimated by fitting Foygel and Dell's model to the pattern of errors that each child made when naming the pictures (See Methods for details of model fitting). The success of this technique in estimating the strength of semantic and phonological processing in children's production system, and in detecting increases in the strength of the S and P parameters as children get older, has been previously established (Budd, Hanley & Griffiths, 2011; Budd, Hanley & Nozari, 2012). We then tested the model's prediction that the strength of S, but not P, should be predictive of the detection of semantic errors in the *moving animals* task. Figure 1 outlines the steps involved in testing this prediction.

In summary, we investigated children's ability to spontaneously detect and correct the semantic errors that they made on familiar words as they produced sentences from meaningful visual events. We examined whether these errors decreased as children got older and whether their ability to monitor them improved. We then tested the prediction of the conflict-detection theory of monitoring that the maturity of the production system in general, and the lexical-semantic part of it in particular (estimated by their accuracy at naming single pictures and their pattern of errors), should be predictive of children's ability to detect and correct a semantic

error during sentence production.

Methods

Participants

Participants were 65 typically-developing monolingual English speaking children who were pupils from a primary school in Colchester, UK. There were 24 children from the reception class (mean age = 5 years and 2 months, SD= 3.98 months), 20 children from the year 1 class (mean age = 6 years and 3 months, SD = 3.96 months) and 21 children from the year 2 class (mean age = 7 years and 4 months, SD = 2.94 months). The sample comprised 29 males and 36 females. Prior to the beginning of the study, informed consent was obtained from the head-teacher of the participating school, the children and their parents.

Materials and Procedure

A sentence production and error monitoring task (the *moving animals* task), a picture-naming task, and a digit span task were used. The digit span task was administered to control for differences in children's working memory capacity, in case it had an effect on performance. Each child was tested individually in two sessions. The digit span and picture-naming tasks were conducted in the first session and the *moving animals* task was administered in the second session. The

sessions for each child always took place on separate days and every session was recorded for offline transcription.

Moving Animals task. This task was presented by a Macintosh Macbook Pro computer using a *Powerpoint* presentation display. On each trial, 9 colored cartoon pictures of familiar animals were presented on the computer screen. At the beginning of each trial, these animals were simultaneously presented in different positions on the screen (see Figure 1 for an example screen shot). The initial positions were randomized. After a brief interval, one of the on-screen animals moved either above or below two other animals. This was followed by two of the other animals moving on top of, or underneath, another animal. When the movement was finished, the child was instructed to make two statements, one for each action. They were asked to state which animals had moved and whether they had moved above or under the other animals. The following would be a fully correct response on a trial in the *moving animals* task: "The dog moved below the rabbit and the monkey. The elephant and the pig moved above the sheep". The children were told to correct any error that they noticed in their speech. The task contained thirty experimental trials that followed a practice phase. The trials were self-paced. The experimenter (the second author) started the practice by performing a trial herself. The children were then exposed to at least three other practice trials. If necessary,

the children were assisted during practice by the experimenter. Practice was continued until the experimenter was satisfied that the instructions were understood and that the child could complete the two sentences required on each trial. All children in this study were able to do that.

Each trial required production of the name of 6 animals and 2 prepositions ("above" or "below"). We limited the locations to "above" and "below", as even very young infants show evidence of early organization of spatial memory for these two categories (e.g., Quinn, 1994), and excellent comprehension of the linguistic preposition applied to them (Meints, Plunkett, Harris, & Dimmock, 2002), thus decreasing the chance of knowledge errors. Since the focus is on semantic errors, syntactic errors such as agreement errors (e.g., "cat move") were not recorded. Thus there were 240 error opportunities for each child (8 items per trial by 30 trials). Presentation was self-paced and children were allowed to take breaks between trials.

Picture-naming. The task was taken from Budd et al. (2011) and comprised 56 black and white line-drawings whose names comprised short monosyllabic words containing 3 to 4 phonemes (e.g. "vase"), and long words of either 3 or 4 syllables containing 6 to 10 phonemes each (e.g. "elephant"). Half of the pictures had a rating of 2.50-4.35 (high-frequency) on a scale of 1 to 5, while the other half was

rated between 1.45 and 2.05 (low-frequency) (Morrison, Chappell, & Ellis, 1997). The material for the picture-naming task was presented via *SuperLab*. The children were asked to state the name of the object as clearly and quickly as possible. At the start of each trial, a fixation cross was followed by the target picture and a simultaneous beep. Trials were self-paced, and a child responded "don't know" if they were unable to name the picture. Following each response, the experimenter pressed the spacebar on the keyboard to move on to the next trial. Before the experiment took place, the children were given four practice trials to ensure that they understood the instructions. Children were allowed to take breaks between trials as needed. Immediately afterwards, a word-picture matching test with the same items was used to assess which errors on the naming task were knowledge errors. Those errors were excluded when scoring the picture naming task.

The digit span test. This test was taken from the Wechsler Intelligence Scale for Children (WISC-IV, Wechsler, 2003) and was administered both forwards (DSF) and backwards (DSB). Maximum score was 17 (9 forward and 8 backward).

Results and Discussion

In total, the *moving animals* task yielded 902 errors, 366 of which were self-corrected. All children were able to complete the 30 trials without great difficulty.

However, three children (one 5-, one 6-, and one 7-year old) committed a high rate of errors on “above” and “below” (39%, 42% and 53% of the 60 opportunities afforded by these terms), which put them 3 SD above the average rate of errors on these two terms in this study population (M= 8%; SD = 10%). Moreover, these three children’s error rates on animal names were comparable to others (7%, 2%, and 6% for the 180 opportunities), and within the normal limits (M = 5%; SD = 4%). Thus, these children may have had special difficulty with spatial processing, or processing of spatial language, and were excluded from further analyses[1]. For the remaining children, we followed a predetermined criterion that if a child consistently made an error on a word (the 9 animal names, or "above" and "below") it would be coded as a knowledge error and not included in the analyses. However, this was not the case with any of the words in the *moving animals* paradigm, as the set was small and chosen to be familiar to children. There were only two formal errors (cat > kit, sheep> sleep) and one unrelated error (dog> above) in the *moving animals* task that were removed from the analysis of semantic errors.

Alternative labels, as long as they were semantically acceptable replacements, were not coded as errors. For example, "kitten" for "cat", "underneath" or "under" for ‘below", and "on" or "on top of" for "above" were

[1] Their inclusion, however, did not change the direction or significance of any of results.

accepted as correct variations. Semantic errors were thus defined as lexical substitutions of animal names (e.g., “dog” for “cat”). Lexical blends (e.g., “dat”) and fragments (e.g., “/kæ/”) were rare and were not included in the coding of semantic errors. Self-corrections were classified as words that the child uttered after having made an initial response. Sometimes the child explicitly implied that they wished to change their response using utterances such as ‘"no, I mean", and sometimes they simply replaced the word with a different one. Only when the child showed no indication that he or she wished to replace an error with a new word, was that error coded as "uncorrected".

Exclusion of the unsuitable participants and items left 792 semantic errors for analysis, of which 344 were detected. Mean error rate per child was 13.75 (SD = 8.35), and the average proportion of corrected errors was 0.52 (SD = 0.26). Figure 2 shows the semantic error counts (left panel) and the proportion of corrected errors (right panel) in sentence production as a function of age. Linear regression was used to determine the effect of age (in months) on the number of errors and the proportion of those errors detected and corrected. Children made marginally fewer semantic errors as they got older ($t = -1.78$, $p = 0.08$), and corrected a significantly larger proportion of those errors ($t = 3.66$, $p = 0.001$).

Insert Figure 2 about here

As can be seen in Figure 2, while detection and correction of speech errors increase with age, there was still much variability among children within the same age range. In fact, the model with age as the only dependent variable explained just 17% of variation in error detection (Adjusted $R^2 = .17$). Nozari et al.'s (2011) model predicts that the ability to detect and correct errors should depend directly on the strength of the production system, which can be captured by pure picture naming ability outside of sentence production. Not surprisingly, children's performance on the picture-naming test also improved with age ($t = 3.03$, $p = 0.004$; Adjusted $R^2 = .12$). So the critical question is whether variations in naming ability, beyond that explained by age, are predictive of children's monitoring and repair ability. When naming was added to the regression model with age as a predictor, the new model explained 26% of variance in error detection and correction (Adjusted $R^2 = 0.26$), with both age and naming scores having significant effects ($t = 2.56$, $p = 0.013$ for age, and $t = 2.87$, $p = 0.006$ for naming scores). This finding shows that picture-naming ability, as an index of the maturity of the language production system, predicts children's ability to detect and correct their errors in sentence production

above and beyond age.

Below, we explore in more detail the claim that the state of the production system predicts monitoring ability. Before that, however, we examine the contribution of a potential confound: children had to hold the event in working memory and verbalize it afterwards. Given the presence of multiple animals in each event, it is possible that uncorrected errors were those in which the child misremembered the event. If so, the strength of children's working memory should be predictive of the proportion of corrected errors. To assess this possibility, digit span scores were entered into the model as a third predictor. The mean digit span score was 10.13 (SD = 4.29). Addition of this variable did not change the model fit (Adjusted $R^2 = 0.26$), and the effect of digit span was not reliable ($t = 0.75$, $p = 0.45$). The effect of both age and naming, however, remained significant in this model ($t = 2.23$, $p = 0.03$ for age, and $t = 2.24$, $p = 0.029$ for naming scores). Together, these results show that children's monitoring behavior is predicted by their age, as well as by the quality of their production system, even when the influence of working memory is excluded.

Next, we focus on testing more specific hypotheses about the relationship between word production and error monitoring during sentence production. Specifically the strength of the semantic-lexical part of the production system

(parameter S in Nozari et al.'s model) should predict detection and correction of semantic errors, but the strength of a different part of the production system, namely the lexical-phonological component (parameter P) should NOT show such a correlation. The next section details how S and P parameters were estimated for each child based on his or her performance on the picture naming task, and relates those to how well each child detected and corrected his or her errors on the *moving animals* task.

Model fitting

Children's errors on the picture-naming task were used to estimate their S and P parameters. Mean accuracy on this task (out of 56 items) was 34.85 (SD = 6.74). Errors were classified into the following groups: semantic, formal, mixed, unrelated, nonword, and other. Table 1 shows all the possible response categories, in an example when the target is "cat". A formal error was scored if the response was a real word that had a phoneme in the correct position in common with the target word (e.g. cat> "cap"). If a response was a real word that was both semantically and phonologically related to the target word, it was considered a mixed error (e.g. cat > "rat"). Conversely, if a response was a real word that did not meet any of the abovementioned criteria it was considered to be an unrelated error

(e.g. cat > "pen"). Any neologisms were classified as nonwords, even if they had several phonemes in common with the target (e.g. cat > "dat"). Consistent with Budd et al. (2011), responses were not considered as incorrect if they were clearly a result of articulatory impairments or accents present throughout the child's entire performance in all tasks. Responses not fitting into any of the previously discussed categories were scored as '"other". "Other" responses included fragments, descriptions (cat > "it meows"), visual errors (microwave > "TV") as well as failing to give a response.

Insert Table 1 about here

Each child's error profile was used to obtain the best-fitting S parameter (the strength of the associative connections between the semantic and lexical layers) and P parameter (the strength of the associative connections between the lexical and phonological layers) for that child. The fitting process entailed inputting to the model the proportions of each of the different response categories described above. Using a maximum likelihood technique, the model then estimated the values of the S and P parameters that provided the closest simulation each child's pattern of responses in order to produce the highest likelihood of simulating his or her error

profile. For example, imagine a child had the following error profile: correct responses = 70%, semantic errors = 10%, formal errors = 10%, mixed errors = 1%, unrelated errors = 5% and nonword errors = 1%. Through a search of space parameters, the algorithm determines that the strength of S should be set at 0.017 and the strength of P should be set at 0.028. This is because the model's estimated pattern of errors, given these two parameter values, would be correct responses = 73%, semantic errors = 9%, formal errors = 10%, mixed errors = <1%, unrelated errors = 7% and nonword errors = 1%, which is quite close to the child's actual performance (for more details on the fitting process see Budd et al., 2011). The normal range of S and P parameters is between 0 and 0.04, with a neurologically-intact adult speaker's parameters hovering close to 0.04. Immature and damaged systems have lower values for one weight or both. Damage to one weight can be independent of damage to another weight, which means that in large samples, S and P parameters are independent of each other (Dell et al., 2013).

Average values of the strength of the S and P parameters were 0.015 and 0.035, and the two were not reliably correlated (Pearson's $r = 0.16$, $p = 0.21$), allowing us to assess their influence independently. Figure 3 shows the relationship between the S (left panel) and P (right panel) parameters and the detection and correction of semantic errors in children. As predicted by Nozari et al.'s (2011)

conflict-detection theory, the S weights were positively and reliably correlated with the ability to correct errors that arose during semantic-lexical mapping (Pearson's r = 0.34, p = 0.008), but there was no reliable correlation with P weights that indexed the strength of a different part of the production system, namely lexical-phonological mapping (Pearson's r = 0.06, p = 0.62). When both variables were entered into a regression model together, the effect was significant for the S (t = 2.65, p = 0.01), but not for the P parameter (t = 0.85, p = 0.93). In summary, the results showed that the state of children's production system predicted how many errors they detected and corrected, and that this was stage-specific: detection and correction of semantic errors was only predicted by the strength of semantic-lexical mapping in production.

Insert Figure 3 about here

General Discussion

The exact mechanism by which children detect and correct their speech errors is not well understood. While several researchers have used observational or

small-corpus data to look into children's error detection, systematic studies of self-correction in children before their teenage years are scarce. This study used a paradigm to elicit semantic errors in which children produced sentences describing meaningful visual events, with the purpose of investigating how they detected and corrected those errors without external prompt. The set of target words was intentionally small, and included nine animals that were all highly familiar to children of the ages tested (5-8 years). As such, demands on knowledge were minimal, allowing for a clear analysis of how 'slips' were detected and corrected. This, however, does not mean that the task was trivial. Repeated production of a small set of semantically-related items produces a high level of interference and makes the speech prone to semantic errors. As such, the paradigm was ideal for studying semantic error detection under a high load of lexical competition.

We found that older children detected and corrected more semantic errors than younger children. Furthermore, above and beyond chronological age, the maturity of children's lexical retrieval system, as determined by accuracy on an independent picture-naming task, was a reliable predictor of how well they detected and corrected their errors during sentence production. It is important to keep in mind that: (1) the S and P parameters are determined using a picture-naming task with a large set of items different from those used in the *moving animals* task; (2) it

is *error commission* that is measured by the picture-naming task, and *error detection and correction* that are measured by the *moving animals* task; (3) pictures are named one at a time during the picture-naming task with no demand on planning a sentential structure or holding items in working memory. Thus, errors made on a picture-naming task are a purer index of the internal dynamics of lexical retrieval than those made during a sentence production task. As such they provide a viable measure for testing the predictions of the conflict-based monitor.

The fact that detection and correction of semantic errors paralleled the maturation of the lexical-retrieval system is aligned with, and complementary to, previous findings that children's ability to revise a sentence grows with the development of their grammatical skills (Rispoli, 2003). The results are also compatible with predictions of the conflict-detection theory of monitoring (Nozari et al., 2011), which proposes that conflict between two or more representations provides a strong signal for error, and that the strength of this signal grows as the underlying production system matures. Importantly, the claim that error detection depends on the internal dynamics of the production system is consistent with Karmiloff-Smith's (1986) finding that 4-12 year-olds detect and correct their speech errors without necessarily having explicit metalinguistic awareness of what the error was. Electrophysiological studies have shown that the conflict signal is

generated as part of the production process, and can act independently of conscious awareness (e.g., Nieuwenhuis et al., 2001).

Nozari et al.'s (2011) model makes another testable prediction: monitoring and repair of semantic errors should only depend on the maturity of the semantic-lexical part of the production system, as this is where conflict leads to the generation of semantic errors. To test this prediction, we simulated the strength of semantic-lexical (S) and lexical-phonological (P) mapping for each child by fitting his or her picture naming data to a computational model. We then showed that children's S, but not their P, parameters reliably predicted detection and correction of semantic errors in the *moving animals* sentence-production task.

The specific relationship between the strength of a part of the production system and the success of detecting errors of a certain type is not solely of abstract interest to theories of speech monitoring, but has implications for learning and treatment of language disorders. Recently, Schwartz et al. (2014) showed that spontaneous error detection in individuals with aphasia marked the strength of the underlying production system. Critically, detection of semantic (but not phonological) errors had the added benefit of learning: simply having detected errors on a previous naming attempt, without external feedback, led to more successful future attempts at naming the same picture, presumably by

strengthening the connections between semantic features and lexical items. This finding suggests that monitoring and detection of lexical-semantic errors can be used as both a diagnostic tool for assessing the severity of lexical retrieval problems, as well as a treatment method, in conjunction with other methods, to improve lexical retrieval.

In sum, this study provides the first systematic report of detection and correction of semantic errors in a structured task by children before their teenage years, along with quantitative predictions from a falsifiable model. The reduction in the number of semantic errors that children made on the moving animals task as they got older reflects the maturation of the semantic-lexical component of the speech production system during this stage of linguistic development. More importantly, the increased efficiency of the semantic-lexical component of children's speech production system appears to be directly related to their ability to detect and repair the errors that it generates. Our study therefore shows that the meta-linguistic behavior of self-correction has its roots in the underlying system that gives rise to errors, and that the maturation of the two happens in close parallel.

The current study also lays the foundation for a number of other critical questions to be taken up in future research. It provides a compelling case for the feasibility of experimental studies to assess the detection of other error types, such

as phonological errors. It is likely that a more linguistically demanding task will be necessary to make possible phonological and grammatical errors. This requires the use of paradigms that elicit a large number of target errors such as tongue-twisters (e.g., Nozari & Dell, 2012; Nozari & Thompson-Schill, 2013). Given the dissociation previously demonstrated between the detection of semantic and phonological errors (e.g., Nozari et al., 2011 and references therein), we do not expect detection of semantic errors to depend critically on that of phonological errors. Thus, the absence of large quantities of phonologically-related errors in the current experiment does not pose a problem for the interpretation of the results on semantic error detection and correction. However, it is entirely possible that increasing task difficulty by requiring the production of sentences that are syntactically and phonologically more complex could decrease the efficacy of the monitor in detecting all types of error. According to the conflict-detection theory, all conflict signals are relayed to a central conflict detection center, the anterior cingulate cortex, which would translate the conflict into an error signal and in collaboration with the prefrontal cortex signal the need for correction. Such a system can have bottlenecks, and as such can be susceptible to general cognitive load of the task. Identifying these bottlenecks is a task for future studies.

Another future question is the role that the comprehension system plays in

detection and correction of various error types in children. While once thought to be the primary cognitive system involved in error detection (e.g., Levelt, 1983; 1989), it is now understood that the role of comprehension in error-detection and correction is only complementary, and most likely limited to detection of errors through the auditory perception of a misspoken word (e.g., Huettig & Hartsuiker, 2010; Nozari et al., 2011; Postma, 2000). Moreover, the contribution of comprehension may differ depending on the error type, as phonological errors seem to rely more strongly on comprehension for detection, at least in adult speakers (e.g., Hartsuiker & Kolk, 2005). A complete understanding of children's monitoring behavior requires exploration of these issues. We believe that the current study takes the first step on this path.

References

Budd, M-J., Hanley, J.R., & Griffiths, Y. (2011). Simulating children's retrieval errors in picture naming: A test of Foygel & Dell's (2000) semantic/phonological model of speech production. *Journal of Memory & Language, 64*, 74-87.

Budd, M-J., Hanley, J.R., & Nozari, N. (2012). Evidence for a non-lexical influence on children's auditory repetition of familiar words. *Journal of Psycholinguistic Research,* **41,** 253-266.

Clark, E. V. (1978). Awareness of language: Some evidence from what children say and do. In A. Sinclair, R. J. Jarvella, & W. J. M. Levelt (Eds.), The child's conception of language (pp. 17-43). Berlin, Germany:Springer-Verlag.

Dell, G. S. (1986). A spreading-activation theory of retrieval in sentence production. *Psychological Review, 93,* 283–321.

Dell, G. S., Nozari, N., & Oppenheim, G. M. (2014). Word production: Behavioral and computational considerations (pp. 88-104). *The Oxford Handbook of Language Production*, Oxford UK: Oxford University Press.

Dell, G. S., Schwartz, M. F., Martin, N., Saffran, E. M., & Gagnon, D. A. (1997). Lexical access in aphasic and nonaphasic speakers. *Psychological Review*, *104*, 801.

Dell, G. S., Schwartz, M. F., Nozari, N., Faseyitan, O., & Coslett, H. B. (2013). Voxel-based lesion-parameter mapping: Identifying the neural correlates of a computational model of word production. *Cognition*, *128*(3), 380-396.

Evans, M. A. (1985). Self-initiated speech repairs: A reflection of communicative monitoring in young children. *Developmental Psychology*, *21(2),* 365-371.

Foygel, D., & Dell, G. S. (2000). Models of impaired lexical access in speech production. *Journal of Memory and Language, 43*, 182-216.

Hartsuiker, R. J., & Kolk, H. H. J. (2001). Error monitoring in speech production: A computational test of the perceptual loop theory. *Cognitive Psychology, 42*(2), 113-157.

Huettig, F., & Hartsuiker, R. J. (2010) Listening to yourself is like listening to others: External, but not internal, verbal self-monitoring is based on speech perception. *Language and Cognitive Processes, 25,* 347-374.

Jaeger, J. J. (1992) 'Not by the chair of my himy hin hin': some general properties of slips of the tongue in young children. *Journal of Child Language, 19*, 335-366.

Jaeger, J. J. (2004). *Kids' slips: What young children's slips of the tongue reveal about language development*. Psychology Press, New York, NY.

Karmiloff-Smith, A. (1986). From meta-processes to conscious access: Evidence from children's metalinguistic and repair data. *Cognition, 23*, 95-147.

Levelt, W. J. M. (1983). Monitoring and self-repair in speech. *Cognition, 14,* 41–104.

Levelt, W. J. M. (1989). *Speaking: From intention to articulation*. Cambridge, MA: MIT Press.

Levy, Y. (1999) Early metalinguistic competence: Speech monitoring and repair behavior. *Developmental Psychology, 35,* 822-834.

Meints, K., Plunkett, K., Harris, P. L., & Dimmock, D. (2002). What is 'on' and 'under' for 15-, 18-and 24-month-olds? Typicality effects in early comprehension of spatial prepositions. *British Journal of Developmental Psychology*, *20*(1), 113-130.

Morrison, C. M., Ellis, A. W., & Quinlan, P. T. (1992). Age of acquisition, not word frequency, affects object naming, not object recognition. *Memory and Cognition, 20*, 705-714.

Nieuwenhuis, S., Ridderinkhof, K. R., Blow, J., Band, G. P. H., & Kok, A. (2001). Error-related brain potentials are differentially related to awareness of response errors: Evidence from an antisaccade task. *Psychophysiology*, *38(5)*, 752–760.

Nozari, N., Arnold, J.E., & Thompson-Schill, S.L. (2014). The effects of anodal stimulation of the left prefrontal cortex on sentence production. *Brain Stimulation,* 7, 784-792.

Nozari, N., & Dell, G. S. (2012). Feature migration in time: Reflection of selective attention on speech errors. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *38*(4), 1084.

Nozari, N., Dell, G.S., & Schwartz, M.F. (2011). Is comprehension necessary for error detection? A conflict-based account of monitoring in speech production. *Cognitive Psychology, 63,* 1-33.

Nozari, N., Kittredge, A. K., Dell, G. S., & Schwartz, M. F. (2010). Naming and repetition in aphasia: Steps, routes, and frequency effects. *Journal of Memory and Language*, *63*(4), 541-559.

Nozari, N., & Thompson-Schill, S. L. (2013). More attention when speaking: does it help or does it hurt? *Neuropsychologia*, *51*(13), 2770-2780.

Peets, K.F. (2009). Profiles of dysfluency and errors in classroom discourse among children with language impairment. *Journal Communication Disorders, 42,* 136-154.

Piaget, J. (1976) *The grasp of consciousness: Action and concept in the young child.* Cambridge, MA: Harvard.

Postma, A. (2000). Detection of errors during speech production: A review of speech monitoring models. *Cognition*, *77*, 97-131.

Quinn, P.C. (1994). The categorization of above and below spatial relations by young infants. *Child Development*, *65*, 58-69.

Rapp, B., & Goldrick, M. (2000). Discreteness and interactivity in spoken word production. *Psychological Review, 107*, 460-499.

Reason, J. (1990) *Human Error.* Cambridge: Cambridge University Press.

Rispoli, M. (2003). Changes in the nature of sentence production during the period of grammatical development. *Journal of Speech, Language, and Hearing Research, 46*, 818–830.

Sasisekaran, J., & Weber-Fox, C. (2012). Cross-sectional study of phoneme and rhyme monitoring abilities in children between 7 and 13 years. *Applied Linguistics, 33*, 253-279.

Schnur, T.T., Schwartz, M.F., Brecher, A., & Hodgson, C. (2006). Semantic interference during blocked-cyclic naming: Evidence from aphasia. *Journal of Memory and Language, 54,* 199-227.

Schwartz, M.F., Middleton, E., Nozari, N., Brecher, A., Gagliardi, M., & Garvey, K. (2014). Learning from your mistakes: The functional value of spontaneous error monitoring in aphasia. *Frontiers in Psychology.* doi:10.3389/conf.fpsyg.2014.64.00070

Stemberger, R.P. (1989). Speech errors in early child language production. *Journal of Memory and Language, 28,* 164-188.

Wechsler, D. (2003). *Wechsler Intelligence Scale for Children-4th Edition (WISC IV)* San Antonio, TX: Harcourt Assessment.

Figure 1

Steps involved in determining the strength of the semantic-lexical and lexical-phonological components of children's language production system (S and P parameter's in Foygel & Dell's (2000) model), and using those parameters to predict the efficacy of spontaneous semantic error detection and repair.

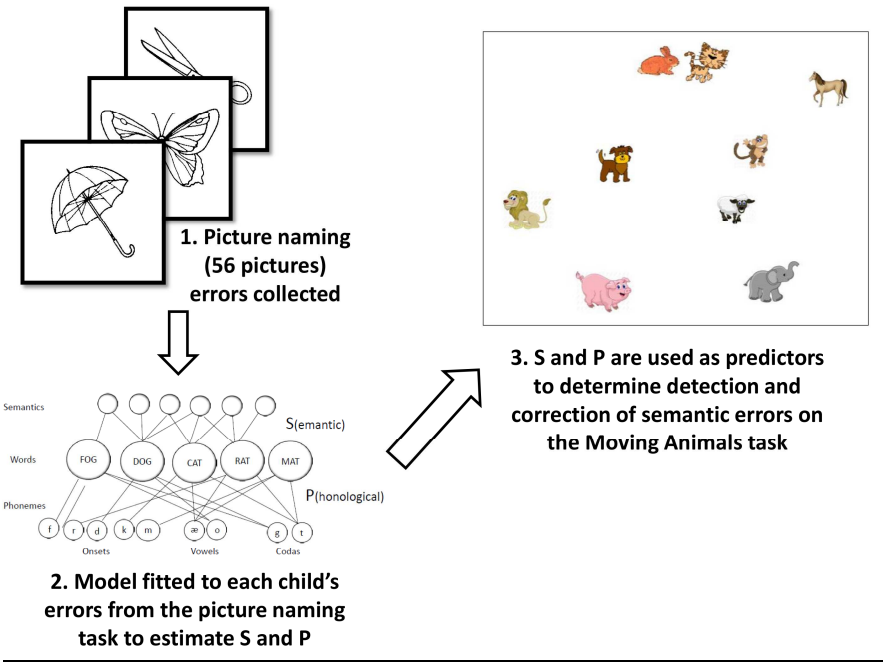

Figure 2

Number of errors (left panel) and the proportion of corrected errors (right panel) made by 5, 6 and 7-year old children on the sentence production task as a function of age in months

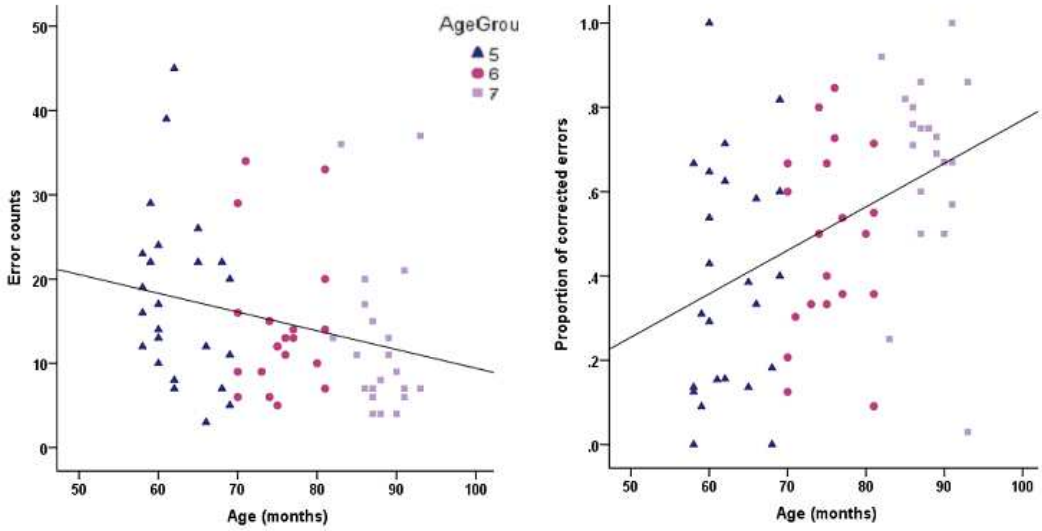

Figure 3

The relationship between detection and correction of semantic errors and the strength of semantic-lexical mapping (model's S parameter; left panel), and the strength of lexical-phonological mapping (model's P parameter; right panel).

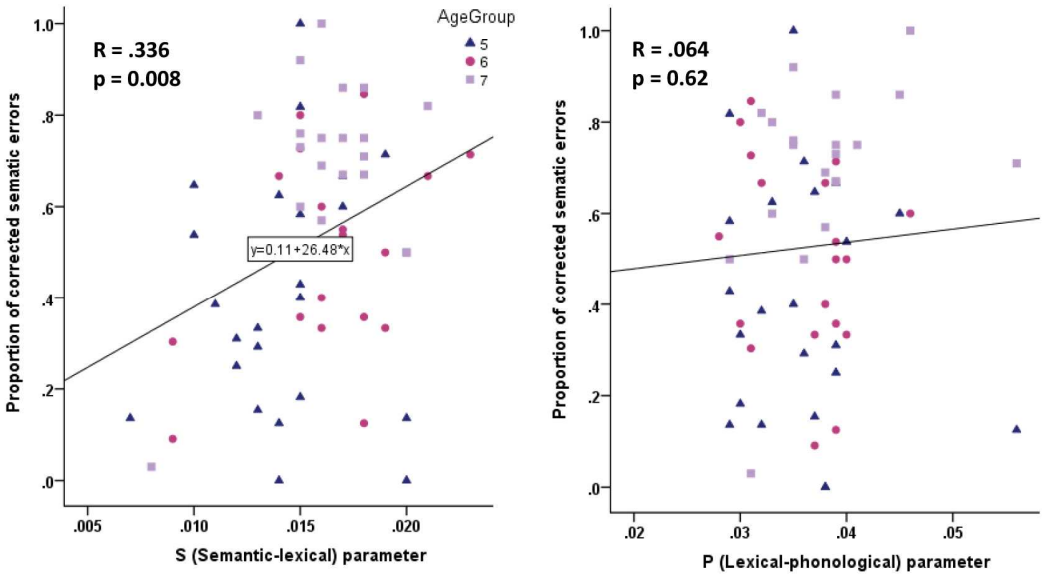

Table 1

The seven possible response categories for a target word on the picture naming task. The table shows an example when the target word was “cat”.

Correct	Semantic	formal (word)	mixed	unrelated (word)	nonword	other
cat	dog	cap	rat	pen	dat	<no response>/ fragments (e.g., k…)/ descriptions (e.g., it meows)/etc.