# What infant-directed speech tells us about the development of compensation for assimilation

Helen Buckler [a]*, Huiwen Goy[b], Elizabeth K. Johnson[c]

*Corresponding author: helen.buckler@nottingham.ac.uk, Tel: +44 (0)115 74 86801
[a] The University of Nottingham, University Park, Nottingham, NG7 2RD, UK
[b] Ryerson University, 350 Victoria Street, Toronto, Ontario, M5B 2K3, Canada
[c] University of Toronto, 3359 Mississauga Road, Mississauga, Ontario, L5L 1C6, Canada

## Abstract

In speech addressed to adults, words are seldom realized in their canonical, or citation, form. For example, the word 'green' in the phrase 'green beans' can often be realized as 'greem' due to English place assimilation, where word-final coronals take on the place of articulation of neighboring velars. In such a situation, adult listeners readily 'undo' the assimilatory process and perceive the underlying intended lexical form of 'greem' (i.e., they access the lexical representation 'green'). An interesting developmental question is how children, with their limited lexical knowledge, come to cope with phonologically conditioned connected speech processes such as place assimilation. Here, we begin to address this issue by examining the occurrence of place assimilation in the input to English-learning 18-month-olds. Perceptual and acoustic analyses of elicited speech, as well as analysis of a corpus of spontaneous speech, all converge on the finding that caregivers do not spoon-feed their children canonical tokens of words. Rather, infant-directed speech contains just as many non-canonical realizations of words in place assimilation contexts as adult-directed speech. Implications for models of developmental speech perception are discussed.

**Keywords**
lexical development; connected speech processes; infant speech perception; place assimilation; speech register; child-directed speech

# What infant-directed speech tells us about the development of compensation for assimilation

## 1.0 Introduction

In conversations between adults, connected speech processes often lead to words being produced with variable realizations. For example, segments may be reduced, deleted, added, only appear in a given context, or adopt features of neighboring segments. These processes all lead to realizations that differ from a word's canonical,[1] or citation, form. In order to comprehend speech accurately, listeners must be able to compensate for this variation. Consider English place assimilation, where a word-final coronal segment may become more labial or velar depending on the place of articulation of the following sound (e.g. *green beans* sounds like *greem beans*). The listener has to be aware that this is a context-dependent change, and extract the meaning 'green' from the signal. Adults are rapidly able to accommodate for the connected speech process and access the intended meaning, even in cases where a connected speech process results in lexical ambiguity.

It has been argued that the ability to compensate for connected speech processes is phonological in nature, and therefore is part of the language-specific knowledge that the learner must acquire. Much experimental evidence supports this view. For example, listeners can accommodate the patterns of their native language, but not other languages (e.g. Darcy, 2002; Lahiri & Marslen-Wilson, 1991; Mitterer & Tuinman, 2012; Otake, Yoneyama, Cutler, & van der Lugt, 1996; Weber, 2001), or other varieties of their own language (Scott & Cutler, 1984;

[1] Note that throughout this paper we use the term 'canonical' to refer to a word that is uttered in its full, citation form. In the phonological literature this form would be the same as its underlying or phonologically unaltered form. We use the term 'non-canonical' to refer to any pronunciation variant that differs from the canonical form. Given the debates in the developmental literature surrounding the nature of phonological representations in the developing mental lexicon, we use the term 'canonical' here as we are primarily discussing the acoustic form in the child's input, and we do not wish to make theoretical claims about the nature of the child's phonological representations.

Tuinman, Mitterer, & Cutler, 2011). In addition, although L2 learners initially struggle with connected speech processes in the language being learned, this ability improves with increased proficiency in the L2 (Darcy, Peperkamp, & Dupoux, 2007). However, this view is not universally accepted. Other studies have suggested that the ability to compensate for connected speech processes depends primarily on language-general auditory skills (Mitterer, Csépe, Honbolygo, & Blomert, 2006).

A growing body of evidence from child language acquisition also suggests that at least some aspects of the ability to compensate for connected-speech processes are acquired. In studies with toddlers it has been found that the ability to compensate for native language assimilation patterns starts to appear at two to three years of age, coupled with an inability to compensate for non-native patterns (Skoruppa, Mani, & Peperkamp, 2013; Skoruppa, Mani, Plunkett, Cabrol, & Peperkamp, 2013). However, further studies suggest that the system is not fully mastered until much later in childhood. Two studies on liaison in French found that 6-year-old children make frequent errors in their production and comprehension of utterances involving liaison (Chevrot, Dugua, & Fayol, 2009; Dugua, Spinelli, Chevrot, & Fayol, 2009). Similarly, English-learning children only display adult-like comprehension of assimilation patterns at seven to eight years of age (Blomert, Mitterer, & Paffen, 2004; Marshall, Ramus, & van der Lely, 2011).

Although data from developmental studies indicate that children need to learn how to compensate for connected speech processes in their native language, very little is known about how they approach this learning problem, and how the ability is acquired. A primary source guiding children's language acquisition is the linguistic input they receive from their caregivers. In order to know how children may learn about connected speech processes, it is crucial to gain better understanding of how they are realized in speech addressed to children. Doing so will allow us to characterize the child's learning situation, and use this information to constrain theories of how the learning process develops.

Broadly speaking, there are two alternatives as to how often connected speech processes and non-canonical forms may occur in infant-directed speech (IDS). Either IDS contains fewer

connected speech processes and more canonical (i.e. citation) pronunciations than adult-directed speech (ADS), or, alternatively, IDS and ADS do not differ in the distribution of canonical pronunciations and connected speech processes used in each register. In the first case, parents may reduce the acoustic-phonetic variation the infant is exposed to in order to break the learning problem down for the child. First the child can learn the canonical form of words, that is, how they would be pronounced in isolation, and later they learn about connected speech processes in their language. This view is in line with the argument that IDS is simplified or hyperarticulated speech that caregivers use as a didactic device to teach their children about the specific features or contrasts in their language's phonology (Burnham, Kitamura, & Vollmer-Conna, 2002; Englund, 2005; Ferguson, 1964; Fish, García-Sierra, Ramírez-Esparza, & Kuhl, 2017; Kuhl et al., 1997, 2008; Liu, Kuhl, & Tsao, 2003; Uther, Knoll, & Burnham, 2007; Werker et al., 2007; Xu Rattanasone, Burnham, Kitamura, & Vollmer-Conna, 2013). The alternative hypothetical case is that IDS, like ADS, contains many non-canonical forms. In this case infants would be faced with a vast spectrum of acoustic-phonetic variation from which to extract both canonical forms and the processes or contexts governing changes that occur in connected speech. This alternative would fit with the argument that IDS is not exclusively designed to support linguistic development. Many of the reported "enhancements" in IDS are not reliable, or not to be beneficial to learning (Cristia & Seidl, 2014; Englund & Behne, 2005, 2006). It is argued that while IDS may have some features that are beneficial for the infant's linguistic acquisition, the parent's primary goal is to build social and emotional bonds with their child, and increased clarity is merely a side-effect of this (Benders, 2013; Cristia & Seidl, 2014).

Existing literature addressing the question of how connected speech processes are realized in IDS provides some support for both of the situations described above, and does not allow us to clearly differentiate between the possibilities. An early study reported greater use of connected speech processes in IDS than ADS (Shockey & Bond, 1980). However, another study reports mixed results, with different patterns observed for different types of connected speech process (Bernstein Ratner, 1984). More recently, Lahey and Ernestus (2013) examined a corpus of natural speech and found, using acoustic and perceptual measures, that IDS contains as much

reduced speech as ADS, suggesting that parents are not increasing the clarity in the signal when talking to their child. However, this study only examined pronunciation variation in two highly frequent lexical items, leading to some doubt regarding the generalizability of the results. In another recent study, Dilley, Millett, McAuley, and Bergeson (2014) looked at regressive place assimilation in English. Tokens were classified by pronunciation type (Canonical, Assimilated, Glottalized, or Deleted), and more canonical pronunciations were found in IDS than ADS, suggesting that caregivers may be speaking more carefully to their children. However, this last study used read speech, which is more conservative than spontaneous speech (Nakamura, Iwano, & Furui, 2008; Warner & Tucker, 2011), and only looked at the acoustic classification of four word-pairs with no acoustic or perceptual measures. Again, this leads to questions regarding the generalizability of their results, as well as questions regarding how variation in the acoustic signal may be perceived.

In summary, there is mixed evidence as to whether the child's input with regard to connected speech processes is simplified or not. Theories of how children learn to cope with connected speech processes in their native language depend crucially on gaining a better understanding of how connected speech is realized in IDS. To date, no single study has investigated connected speech processes in IDS in a range of token types in both read and spontaneous speech. Nor have they used convergent approaches, combining acoustic, perceptual and classification analyses of tokens, as we do in the current study. Furthermore, past studies have not investigated the prevalence of connected speech processes in contexts which may give rise to lexical ambiguity, that is, where a lexical contrast would be neutralised if the speaker used a connected speech process.

The two hypothetical situations described above, namely whether IDS contains more canonical pronunciations than ADS or not, each create a different learning situation for the child, and present the child with different challenges. Spoon-feeding the child canonical forms may help them initially, but learning the citation form of a word does not teach them about how word forms may change in different contexts in connected speech. This is something that they must learn to cope with in order to function as a competent language user. Alternatively, IDS

contains as many non-canonical forms and connected speech processes as ADS. Although this presents the child with a more accurate picture of the variation that may occur in their language, they may struggle to know what the canonical, or citation, form of a word is, and hypothetically, may draw incorrect conclusions. If they hear a phrase like *greem beans*, for example, and building representations from the acoustic signal, they may speculate that *green* and *greem* each deserve their own lexical entry, much like *bean* and *beam*. Investigating the acoustic realization and perception of words in connected speech allows us to identify which learning situation children typically face, paving the way for future work into how they may overcome the specific learning challenges. For example, if the input provides the learner with clear, categorical variation, they may be able to use context-specific distributional statistics to learn the underlying, canonical form and its legitimate alternation (cf. Peperkamp & Dupoux, 2002). Alternatively, if the input is less categorical, the abundance of variation they are presented with may serve to inform the learner of all possible legitimate variation in the language and support the formation of generalisations across words and contexts.

### *1.1 The Current Study*

The current study investigates the prevalence of pronunciation variants in connected speech in English IDS, with particular focus on contexts that may give rise to lexical ambiguity. Place assimilation causes the final coronal segment of the first word to adopt place of articulation of the following word. In a phrase such as *cat box*, the final [t] of *cat* is influenced by the subsequent [b] of *box*, and can adopt a more labial pronunciation. Thus, place assimilation can result in lexical ambiguity, where the difference between *cat box* and *cap box* is minimised. We test whether IDS contains more unambiguous, canonical forms than ADS in connected speech generally, and in particular, whether there are fewer instances of place assimilation in contexts where assimilation is licensed.

We recorded a corpus of IDS and ADS. Mothers of 18-month-olds were recorded because of rapid vocabulary growth at this age. Large-scale studies of vocabulary development have demonstrated that productive vocabulary increases tenfold between 16 and 30 months

(Fenson et al., 1994). If parents were trying to support their child's vocabulary development by reducing ambiguity in their IDS we would expect them to do so during this period. Using this corpus we report data from a number of analyses. Firstly, we report data from two perceptual experiments in which adults were presented with a subset of tokens from the corpus and required to identify the intended target (cf. Lahey & Ernestus, 2013). Tokens used in these identification tasks were also classified by pronunciation type by trained phoneticians (cf. Dilley et al., 2014). Finally, all tokens in the corpus were analysed acoustically to gauge the degree of variation present in a large number of tokens (cf. Dilley & Pitt, 2007; Gow, 2001, 2002). We verify our findings by comparing the data from our elicited corpus with data from a corpus of spontaneous mother-child interactions. Thus, we show how acoustic properties are translated into perceptual judgments in a large corpus recorded in a laboratory setting, and how the acoustic properties of elicited speech compares to spontaneous speech. Taken together the different analyses provide a strong body of convergent evidence for how connected speech processes are realized in IDS, paving the way for future research into how connected speech processes are perceived, interpreted and acquired by the infant learner.

## 2.0 Creation of the Laboratory Corpus

We created a corpus of ADS and IDS by recording mothers of toddlers reading stories to their child (IDS) and the experimenter (ADS). The stories were appropriate for a young child and contained a number of two-word phrases that are potentially ambiguous due to place assimilation (e.g. *cat box* / *cap box*). Mothers were also recorded retelling the story to both listeners, thus eliciting both scripted and unscripted speech in both IDS and ADS. These different speech styles were chosen to reflect different situations that mothers use in everyday interactions with their children, and increase the generalizability of our results.

### *2.1 Participants*

Twelve mothers of 18-month-old children ($M_{age}$ = 1;6.14, range = 1;5.16-1;7.17) were recorded addressing their infant (IDS condition) or the experimenter (ADS condition). All

mothers had lived in Canada since childhood and were English dominant. A further eight mothers were recorded but these recordings did not contain sufficient tokens of high enough quality for further analysis.

*2.2 Materials*

Eight pairs of two-word phrases were created which did or did not license regressive place assimilation (e.g., *cat box* and *cap box).* The phrases were all potentially ambiguous if the speaker produced tokens with place assimilation.

In each of the eight pairs of phrases, the second word of each pair was the same, and the first words were minimal pairs differing in the place of articulation of the coda. The coda segment was a labial or coronal nasal or plosive. The onset segment of the second word was always a labial plosive or nasal.

A story was created that incorporated all pairs of phrases. Two versions of the story were written, with one member of the pair of all eight phrases included in each version. For example, *cat box* appeared in Version 1 of the story and *cap box* in Version 2. Phrases were balanced across versions of the story ensuring an equal number of assimilation-licensing and non-assimilating tokens in each story, and an equal number of nasal and plosive contexts in each story. Version 1 contained the tokens: *bean painter, cat box, cat burglar, comb maker, ape babies, Jem Pickles, teen bears, grape pie.* Version 2 contained the tokens: *beam painter, cap box, cap burglar, cone maker, eight babies, Jen Pickles, team bears, great pie.*

The phrases appeared in identical sentences in each version of the story. Sentences were semantically neutral and did not predict one member of the pair more than the other, for example, *Isn't the cat box pretty?* or *Isn't the cap box pretty?*

The two stories were illustrated and printed in colour onto fabric approximately 25 cm square. Each version of the story was sewn into a cloth book to reduce noise. An additional copy of each storybook was made using the same illustrations and only key-words from the story. The key-words included the target phrases. The complete text versions of the books were

used to elicit scripted speech, and the version with key-words was used to elicit unscripted speech.

*2.3 Procedure*

Mothers were required to tell one of the two versions of the story four times. They read the full-text version of the book to both their infant and the experimenter, and retold the story to each person using the key-word version of the book. Half of the mothers recorded the story first in IDS and then ADS, and half recorded ADS followed by IDS. Scripted speech was elicited immediately before unscripted speech.

Recordings were made in a quiet laboratory room. When recording IDS, the mother sat on a chair at a table with their infant on their lap. In the scripted speech condition, they were given the full-text version of the book and instructed to read the story to their child in a natural manner. They were subsequently given the key-word version of the book and instructed to retell the story to their child in their own words. When recording ADS, the experimenter sat on a chair across the table from the mother. The mother was instructed to read full-text version of the story as if reading aloud from the newspaper. For unscripted ADS they were instructed to retell the story using their own words.

Audio recordings were made on a Zoom Handy H4n digital audio recorder placed on the table in front of the mother. Recordings were made in .wav format with a sampling frequency of 44100 Hz. Tokens of the target phrases in the recordings were labelled in *Praat* (Boersma & Weenink, 2011). Tokens were excluded from any further analysis if the target phrase was unclear due to disfluencies in the speech, poor audio quality or noise.

**3.0 Perception Experiments**

We first report data from two perceptual experiments that address whether IDS is more intelligible than ADS, as would be expected if IDS contains more canonical pronunciations of words (for similar use of adult ratings see e.g. Gow, 2001, 2002, 2003; Mitterer & Blomert, 2003; Mitterer, Csépe, & Blomert, 2006; Zimmerer, Reetz, & Lahiri, 2009). Adult listeners

were required to identify the intended target in a subset of tokens taken from the corpus of elicited IDS and ADS (e.g. *cat box* vs. *cap box*). If IDS is clearer than ADS, we predicted that it should contain fewer ambiguous pronunciations as parents make the distinction between competing lexical items clear (i.e. cat box ≠ cap box). We predicted that IDS would contain more citation forms than ADS both in contexts where assimilation is licensed, and in contexts where it is not, and this would be reflected in adult listeners' identification accuracy. Specifically, we expected adults to identify the target more accurately in IDS than ADS. The corpus was designed to focus on place assimilation as a pronunciation variant; however, even in contexts that do not license place assimilation there are a number of variants that speakers may use (e.g. deletion or glottalization of the final segment) that may contribute to ADS being less clear than IDS. If IDS contains fewer assimilated tokens (in an assimilation licencing context), we predicted that the effect of assimilation context on listener accuracy would be greater for IDS than ADS.

We additionally expected listeners to identify the intended target more accurately in read speech, as this register is typically slower and produced with more articulatory effort than spontaneous speech (Nakamura et al., 2008; Warner & Tucker, 2011).

### *3.1 Experiment 1*

Adult listeners participated in a two-alternative forced choice task. Listeners heard a subset of the two-word phrases spliced from the laboratory corpus and were required to identify the phrase heard, for example, "cat box" or "cap box."

#### *3.1.1 Methods*

##### *3.1.1.1. Participant*s

Fifty-two adults were recruited from the undergraduate population at the University of Toronto ($M_{age}$ = 19 years, 38 females). All spoke English as their dominant language and had acquired English by the age of 5.

*3.1.1.2. Materials*

Stimuli were 224 tokens of mothers' speech selected from the corpus presented in Section 2.0. Target phrases were selected as stimuli if there was a good-quality recording of a given two-word phrase spoken in both IDS and ADS by the same mother in either Scripted or Unscripted speech. For each pair of phrases we selected an equal number of tokens that licensed assimilation and that did not, i.e. an equal number of *cat box* and *cap box* tokens. Scriptedness was also taken into consideration, and there were an equal number of scripted and unscripted tokens. If there were multiple repetitions of a token in a recording that were suitable for inclusion, we always selected the first repetition. A total of 224 tokens were selected as stimuli in the identification task. Of these, 56 were IDS scripted speech, 56 IDS unscripted speech, 56 ADS scripted speech and 56 ADS unscripted speech. Stimuli are broken down by phrase in Table 1.

*Table 1. Number of tokens of each two-word phrase used as stimuli in Experiments 1 and 2. For each item half of the tokens were scripted speech, taken from the recordings of the mothers reading the story book, and half were unscripted, taken from recordings of the mothers retelling the story in their own words.*

| | ADS | IDS |
|---|---|---|
| Ape babies | 6 | 6 |
| Eight babies | 6 | 6 |
| Beam painter | 8 | 8 |
| Bean painter | 8 | 8 |
| Cap box | 6 | 6 |
| Cat box | 6 | 6 |
| Cap burglar | 10 | 10 |
| Cat burglar | 10 | 10 |

| | | |
|---|---|---|
| Comb maker | 8 | 8 |
| Cone maker | 8 | 8 |
| Grape pie | 2 | 2 |
| Great pie | 2 | 2 |
| Jem Pickles | 10 | 10 |
| Jen Pickles | 10 | 10 |
| Team bears | 6 | 6 |
| Teen bears | 6 | 6 |

*3.1.1.3. Procedure*

Participants were tested individually in a quiet room. The experiment was presented in *Praat* (Boersma & Weenink, 2011). In each trial, participants saw the orthographic form of the two competing phrases displayed in two boxes side-by-side on a laptop screen (e.g. *cat box* and *cap box*) and heard the two-word phrase to be identified over closed headphones (Sennheiser HD 280 pro). They were instructed to identify which of the phrases they heard by clicking in the corresponding box. Participants were also required to indicate, on a scale from 1 to 4, how confident they were in their judgment, where 1 indicated 'not at all sure' and 4 indicated 'very sure.' Participants were able to replay each token up to three times and received no feedback on the accuracy of their response. There were 224 trials presented in a randomized order. Progress through the study was self-paced and participants were able to take short breaks as desired. Participants typically took approximately 20 minutes to complete the task.

*3.1.2 Results & Discussion*

Fourteen trials were removed because the participants' response time was greater than 10s or they responded before they had heard the complete target phrase.

Accuracy data were analyzed using generalized logistic mixed effects model using the function *glmer* in the package lme4 (Bates, Maechler, & Bolker, 2015) in R (R Core Team, 2012). Fixed effects of Addressee (IDS or ADS), Scriptedness (scripted or unscripted) and

Assimilation Licensing Context (yes or no) were included, including the interactions of all of these fixed effects. Random intercept terms were included for Participant, Speaker and Item. This was the maximal random effects structure that achieved convergence (Barr, Levy, Scheepers, & Tily, 2013). All predictor variables were binary, and coded using an effects coding scheme (-.5, .5) so that parameter estimates would reflect the 'main effect' or the mean difference in log odds between the two conditions (Barr, 2008). Statistical significance was evaluated using likelihood ratio tests to calculate the change in model fit (log likelihood) between the full model and a reduced model without each fixed effect of interest (Barr et al., 2013). The difference between the two models was evaluated against the chi-square distribution.

Figure 1 shows accuracy scores by Addressee and Assimilation Licensing Context. We found a main effect of Addressee ($\beta_{\text{Addressee}} = .11$, $SE = .04$, $\chi^2(1) = 7.76$, $p = .005$); surprisingly, participants were more accurate in identifying the target when it was spoken in ADS ($M_{ADS} = .66$, $SD = .06$) than IDS ($M_{IDS} = .64$, $SD = .06$). We also found a main effect of Assimilation Licensing Context ($\beta_{\text{Assimilation}} = .58$, $SE = .27$, $\chi^2(1) = 4.\ 1$, $p = .04$). Participants were more accurate in identifying the target when the phrase contained an assimilation-licensing context ($M_{LicensingContext} = .72$, $SD = .09$), than when it did not ($M_{NoLicensingContext} = .58$, $SD = .1$). The interaction of Addressee and Assimilation Licensing Context was marginally significant ($\beta_{\text{Addressee:Assimilation}} = .13$, $SE = .08$, $\chi^2(1) = 2.7$, $p = .1$). Although listeners were more accurate overall in identifying tokens in ADS than IDS, this difference tended to be larger for tokens that contained an assimilation licensing context ($M_{ADS,\ LicensingContext} = .74$, $SD = .09$, $M_{IDS,\ LicensingContext} = .7$, $SD = .1$) than a non-assimilating context ($M_{ADS,NoLicensingContext} = .59$, $SD = .1$, $M_{IDS,NoLicensingContext} = .58$, $SD = .1$). Participants' accuracy in identifying tokens with no assimilating context (e.g. *cap box*) was similar in IDS and ADS, but if the phrase contained an assimilating context (e.g. *cat box*) they were slightly more accurate in identifying the intended target when it was uttered in ADS than IDS. Crucially, counter to our predictions, in no context were listeners more accurate at identifying the intended target in IDS than ADS. Furthermore, the effect of Assimilation Licensing Context was not stronger in IDS than ADS.

Interestingly, there was no significant effect of Scriptedness ($\beta_{\text{Scriptedness}}$ = .002, *SE* = .04, $\chi^2(1)$ = .002, *p* = .96), indicating that participants were equally accurate at identifying the intended target in Scripted as Unscripted speech ($M_{Scripted}$ = .65, *SD* = .06, $M_{Unscripted}$ = .65, *SD* = .05). That is, Scripted speech was not, as predicted, clearer than Unscripted speech. There was no interaction of Addressee and Scriptedness, ($\beta_{\text{Addressee:Scriptedness}}$ = .1, *SE* = .08, $\chi^2(1)$ = 1.36, *p* = .24). Finally, there was no interaction of Scriptedness and Assimilation Licensing Context ($\beta_{\text{Scriptedness:Assimilation}}$ = .04, *SE* = .08, $\chi^2(1)$ = 0.27, *p* = .6), and no three-way interaction of Addressee, Scriptedness and Assimilation Licensing Context ($\beta_{\text{Addressee:Scriptedness:Assimilation}}$ = -.12, *SE* = .08, $\chi^2(1)$ = 0.52, *p* = .47).

Note to Publisher: Insert Figure 1 about here

We additionally analyzed confidence data using the same model parameters. We found a significant effect of Addressee ($\beta_{\text{Addressee}}$ = .06, *SE* = .01, $\chi^2(1)$ = 17.94, *p* < .001). Similar to the accuracy data, participants were more confident in their ability to identify the target in ADS ($M_{ADS}$ = 3, *SD* = .34) than IDS ($M_{IDS}$ = 2.94, *SD* = .37). Even though listeners were more accurate in identifying targets with an assimilation licensing context, this was not reflected in their confidence, and we found no effect of Assimilation Context ($\beta_{\text{Assimilation}}$ = .01, *SE* = .06, $\chi^2(1)$ = .04, *p* = .85). The interaction of Addressee and Assimilation Context was not significant ($\beta_{\text{Addressee:Assimilation}}$ = -.03, *SE* = .03, $\chi^2(1)$ = 1.21, *p* = .27). The effect of Scriptedness approached significance ($\beta_{\text{Scriptedness}}$ = -.02, *SE* = .01, $\chi^2(1)$ = 2.68, *p* = .1), and participants were marginally more confident in their judgments of unscripted tokens ($M_{Unscripted}$ = 2.98, *SD* = .34) than scripted tokens ($M_{Scripted}$ = 2.96, *SD* = .36). However, this pattern of results was the same in both IDS and ADS, and we find no significant interaction of Addressee and Scriptedness ($\beta_{\text{Addressee:Scriptedness}}$ = .03, *SE* = .03, $\chi^2(1)$ = 1.1, *p* = .29). As with accuracy, there was no interaction of Scriptedness and Assimilation Licensing Context ($\beta_{\text{Scriptedness:Assimilation}}$ = -.04, *SE* = .03, $\chi^2(1)$ = 2.49, *p* = .11), and no three-way interaction ($\beta_{\text{Addressee:Scriptedness:Assimilation}}$ = .06, *SE* = .05, $\chi^2(1)$ = 1.45, *p* = .23).

Based on earlier work (Mitterer & Blomert, 2003; Zimmerer et al., 2009), we had expected that listeners' identification accuracy would reflect the clarity of the speech heard. Therefore, if IDS were clearer than ADS, adults would be more accurate in identifying the target in IDS than ADS. However, listeners' judgments of IDS were not more accurate than of ADS, either when consider all tokens together, or only the tokens in an assimilation-licensing context. This could imply that parents are not reducing ambiguity in IDS by using more canonical pronunciations than in ADS; however, there are a number of other possible explanations that need to be explored.

Perhaps listeners found the task difficult, and the design of the study may have made the identification of IDS tokens particularly challenging for adults (who are presumably more familiar with ADS). It is noteworthy that overall accuracy in the task was 65%, which is reasonably low for a word-recognition task. Two elements in the design may have contributed to how challenging listeners found the task. Firstly, listeners had very little speech material to base their decision on, and secondly, they had no expectation of which register the coming trial would be spoken in. Regarding the first point, previous research has found that adult listeners do not find words spliced out of continuous speech more intelligible in IDS than ADS (Bard & Anderson, 1983, 1994). The authors argue that adults use more predictable sentence structures, repetition, and use of extra-linguistic cues in IDS, which compensates for less clear articulation of individual words. Similar factors may be contributing to our data, as participants heard only two-word phrases without any contextual cues. Furthermore, constantly changing register may have created a greater challenge than if the same register was maintained throughout, as listeners could not predict whether the following token would be IDS or ADS, and had to adapt to the register upon hearing the token. Given that adult listeners are less familiar with IDS, this may have affected their ability to reliably judge IDS tokens more than ADS tokens.

In Experiment 2, we address the above concern by presenting participants with more contextual information and blocking stimuli presentation by register. Instead of hearing just the two-word phrase, listeners heard the phrases in its sentence context. Stimuli presentation was

blocked, and participants heard a succession of IDS trials followed by a succession of ADS trials.

Note to Publisher: Insert Figure 2 about here

*3.2 Experiment 2*

Experiment 2 was designed to assess whether listener accuracy in identifying the intended target in Experiment 1 was affected by lack of contextual information and unfamiliarity with the different speech registers. Participants heard tokens from the same recordings of mothers' speech as in Experiment 1, but now, instead of just hearing a two-word phrase, they were presented with the whole sentence that the phrase was uttered in (e.g. *"Isn't the cat box pretty?"* or *"Isn't the cap box pretty?"*). Trial presentation was also blocked by register.

*3.2.1 Methods*

*3.2.1.1. Participants*

Thirty-two adults participated in Experiment 2 ($M_{age}$ = 18.3 years, range = 17-23 years, 27 female). Participants were recruited from the same population as in Experiment 1, and met the same eligibility criteria.

*3.2.1.2. Materials*

Stimuli were taken from the same corpus of mothers' speech described in Section 2. Target phrases were the same as presented in Experiment 1, however, in Experiment 2, participants were presented with the whole sentence containing the target phrase.

*3.2.1.3. Procedure*

The experimental procedure was identical to that reported in Experiment 1, with the exception that stimuli were now presented in blocks of IDS or ADS speech. There were 8 blocks of 28 trials, and they alternated between blocks of IDS and ADS tokens. Block

presentation was counterbalanced such that half of the participants heard IDS in the first block, and half heard ADS.

*3.2.2 Results & Discussion*

Data were analyzed in the same manner as Experiment 1. There was no effect of counterbalancing order and therefore this effect was not included in the main analysis ($\beta_{\text{Counterbalance}}$ = .01, *SE* = .13, $\chi^2(1)$ = 0.006, *p* = .94). Our primary variable of interest was participants' accuracy in identifying the intended target and whether this was mediated by Addressee and/or Assimilation Licensing Context. As in Experiment 1, we find a significant main effect of Addressee ($\beta_{\text{Addressee}}$ = .22, *SE* = .06, $\chi^2(1)$ = 14.83, *p* < .001). Again, participants were more accurate in identifying the intended target when spoken in ADS ($M_{ADS}$ = .7, *SD* = .08) than IDS ($M_{IDS}$ = .66, *SD* = .05). We also again find a significant effect of Assimilation Licensing Context ($\beta_{\text{Assimilation}}$ = 1.31, *SE* = .22, $\chi^2(1)$ = 18.52, *p* < .001), with increased accuracy for tokens that licensed assimilation, such as *cat box* ($M_{LicensingContext}$ = .81, *SD* = .1), than those that do not, such as *cap box* ($M_{NoLicensingContext}$ = .56, *SD* = .12). There was no interaction of Addressee and Assimilation Context ($\beta_{\text{Addressee:Assimilation}}$ = -1.1, *SE* = .11, $\chi^2(1)$ = 0.96, *p* = .33). The advantage that listeners have for interpreting ADS over IDS does not vary depending on whether the phrase licenses assimilation or not ($M_{IDS,\ NoLicensingContext}$ = .53, *SD* = .13, $M_{IDS,\ LicensingContext}$ = .79, *SD* = .12; $M_{ADS,\ NoLicensingContext}$ = .59, *SD* = .13, $M_{ADS,\ LicensingContext}$ = .82, *SD* = .1). As in Experiment 1, in Experiment 2, we find no significant effect of Scriptedness ($\beta_{\text{Scriptedness}}$ = -0.07, *SE* = .06, $\chi^2(1)$ = 1.64, *p* = . 2). Participants were not more accurate in identifying tokens in scripted, read speech ($M_{Scripted}$ = .67, *SD* = .06) than unscripted, spontaneous speech ($M_{Unscripted}$ = .69, *SD* = .07). The interaction of Addressee and Scriptedness was not significant ($\beta_{\text{Addressee:Scriptedness}}$ = .15, *SE* = .11, $\chi^2(1)$ = 1.68, *p* = .19), however, there was a significant interaction of Scriptedness and Assimilation Licensing Context ($\beta_{\text{Scriptedness:Assimilation}}$ = .53, *SE* = .11, $\chi^2(1)$ = 22.74, *p* < .001). The effect of Assimilation Licensing Context on accuracy was greater for scripted than unscripted tokens ($M_{Scripted,\ NoLicensingContext}$ = .52, *SD* = .13, $M_{Scripted,\ LicensingContext}$ = .82, *SD* = .11; $M_{Unscripted,\ NoLicensingContext}$ = .6, *SD* = .13, $M_{Unscripted,\ LicensingContext}$ = .79,

*SD* = .1). The three-way interaction was not significant $\beta_{\text{Addressee:Scriptedness:Assimilation}}$ = .33, *SE* = .22, $\chi^2(1) = 2.19$, $p = .14$).

Confidence data were analysed using the same model parameters. Again, listener confidence reflects accuracy, with a marginally significant effect of Addressee ($\beta_{\text{Addressee}}$ = .03, *SE* = .016, $\chi^2(1) = 3.54$, $p = .06$). Participants tended to be more confident in their ability to identify the target when it was spoken in ADS ($M_{ADS}$ = 3.11, *SD* = .38) than IDS ($M_{IDS}$ = 3.08, *SD* = .37). Similarly, Assimilation Licensing Context also affected confidence ($\beta_{\text{Assimilation}}$ = .12, *SE* = .05, $\chi^2(1) = 5.11$, $p = .02$), and listeners were more confident in their judgment of targets that licensed assimilation ($M_{LicensingContext}$ = 3.15, *SD* = .38) than those that did not ($M_{NoLicensingContext}$ = 3.05, *SD* = .37). There was no significant interaction of Addressee and Assimilation Context ($\beta_{\text{Addressee:Assimilation}}$ = -.01, *SE* = .03, $\chi^2$ (1) = .05, $p = .83$).

Scriptedness did not affect listeners' confidence ($\beta_{\text{Scriptedness}}$ = .02, *SE* = .2, $\chi^2(1) = 1.01$, $p = .32$). However, the interaction of Addressee and Scriptedness approached significance ($\beta_{\text{Addressee:Scriptedness}}$ = -.05, *SE* = .03, $\chi^2(1) = 2.65$, $p = .1$). The difference in listeners' confidence rating of scripted and unscripted tokens tended to be greater in IDS ($M_{IDS,Scripted}$ = 3.1, *SD* = .35, $M_{IDS,Unscripted}$ = 3.06, *SD* = .39) than ADS ($M_{ADS,Scripted}$ = 3.11, *SD* = .39, $M_{ADS,Unscripted}$ = 3.12, *SD* = .38). Different from the accuracy data, the interaction of Scriptedness and Assimilation Licensing Context was not significant ($\beta_{\text{Scriptedness:Assimilation}}$ = .004, *SE* = .03, $\chi^2(1) = 0.02$, $p$ = .89), however, the the three-way interaction of Addressee, Scriptedness and Assimilation Licensing Context approached significance ($\beta_{\text{Addressee:Scriptedness:Assimilation}}$ = .12, *SE* = .07, $\chi^2(1)$ = 3.23, $p = .07$), indicating that Assimilation Context had an effect on the size of the difference of listeners' confidence rating of scripted and unscripted tokens in IDS and ADS.

In both Experiment 1 and 2 we find the somewhat surprising result that adult listeners are more accurate at identifying the intended target in ADS than IDS, and that they are not more accurate in identifying the target in an assimilation licensing context in IDS than ADS. This does not rule out the possibility that IDS does contain more canonical forms than ADS, but

indicates that perceptually, to an adult listener, IDS is not clearer or less ambiguous than ADS. To investigate whether these results reflect differences in the acoustic stimuli or adults' perceptual judgments, we supplemented the data from the two identification experiments with a classification analysis of pronunciation variation (cf. Dilley et al., 2014; Dilley & Pitt, 2007).

### *3.3 Phonetic Classification*

The tokens from the stimuli set that contained an assimilation-licensing context (e.g. *cat box*) were classified by pronunciation variant. This provides categorical data of pronunciation variants used by speakers, and allows us to quantify whether there are more canonical pronunciations in IDS than ADS. This analysis investigates the distribution of pronunciation variants in IDS and ADS in our stimuli, and how this affected listeners' accuracy in the identification task

#### *3.3.1 Procedure*

The tokens used in the perception study that contained an assimilation-licensing context, for example *cat box*, were classified by pronunciation type. Three phonetically trained coders used spectrographic information to classify the pronunciation of each token into one of three categories: Canonical, Assimilated, or Other. These categories were based on the categories established by Dilley and Pitt (2007) and Dilley et al. (2014). We collapsed their categories of Glottalized and Deleted into the single category of Other. A token was classified as Assimilated if there was evidence in the preceding vowel of a downward movement in the F2 that would be associated with a transition into a labial, and therefore assimilated, place of articulation of the following segment. A token was classified as Canonical if the word-final obstruent (e.g. the [t] of *cat* in the phrase *cat box*) was perceived as being present, and without voicing irregularity, and if the formant transition in second formant of the previous vowel was consistent with a coronal place of articulation. A token was classified as Other if the word-final obstruent was not present, was glottalized, or if the pronunciation did not meet the criteria of the Canonical or Assimilated categories.

Three trained coders classified all 112 tokens. Coding data was compared across coders, and for tokens where there was disagreement between coders they were permitted to reconsider their classification. If disagreement remained, the classification that two out of the three coders agreed upon was taken as the coding value. Percent agreement between coders was 81.2%, resulting in a Fleiss Kappa of $\kappa = 0.8$, indicating substantial agreement between coders (Landis & Koch, 1977).

*3.3.2 Phonetic Classification and Identification Accuracy Results & Discussion*

Tokens were classified as having a realization that was Canonical, Assimilated, or Other. In IDS, 46% of tokens were classified as having a Canonical pronunciation. Only 11% of tokens were pronounced with definite assimilation, and 43 % were classified as Other. In ADS 43% of tokens had a Canonical pronunciation, with 14% Assimilated and 43% Other pronunciation variants. The distribution of Canonical and Non-Canonical (Assimilated and Other) pronunciations in IDS and ADS did not differ, $\chi^2(1) = .04$, $p = .85$.

A logistic mixed model was used to test whether listeners' identification accuracy was predicted by a token's pronunciation (Figure 2). We included fixed effects of Pronunciation Classification (Canonical or Non-Canonical), Addressee (IDS vs. ADS), and the interaction of Addressee and Pronunciation Classification. We also included random intercept terms of Participant, Speaker, and Experiment (1 or 2). There was a significant effect of Pronunciation Classification, $\beta_{\text{Pron.Classification}} = 0.64$, $SE = .06$, $\chi^2(1) = 136.3$, $p < .001$. Participants were more accurate in identifying targets that were classified as having Canonical pronunciations ($M_{Canonical} = .81$, $SD = .12$) than non-canonical pronunciations ($M_{NonCanonical} = .70$, $SD = .12$). Participants were more accurate in identifying targets in ADS ($M_{ADS} = .77$, $SD = .1$) than IDS ($M_{IDS} = .74$, $SD = .12$), $\beta_{\text{Addressee}} = 0.19$, $SE = .05$, $\chi^2(1) = 14.02$, $p < .001$. However, Pronunciation Classification was not more predictive of accuracy in IDS or ADS and there was no interaction of these effects, $\beta_{\text{Pron.Classification,Addressee}} = -0.11$, $SE = .11$, $\chi^2(1) = 1.12$, $p = .29$; $M_{ADSCanonical} = .82$, $SD = .11$; $M_{ADSNonCanonical} = .72$, $SD = .13$; $M_{IDSCanonical} = .8$, $SD = .15$; $M_{IDSNonCanonical} = .68$, $SD = .14$...

*3.4 Discussion*

Results of both Experiment 1 and 2 were remarkably similar. In both experiments listeners were consistently more accurate in identifying the intended target when it was uttered in ADS. In both experiments participants were more accurate in identifying the target when it appeared in context that licensed assimilation (e.g. *cat box*) than when it did not (e.g. *cap box*), however, this did not differ by IDS or ADS. Together these results indicate that perceptually IDS is not clearer or less ambiguous than ADS, and does not contain more canonical forms (either in contexts where assimilation is licensed or where it is not). This result is further supported by the classification analysis of the tokens with an assimilation-licensing context, with the same distribution of pronunciation variants attested in IDS and ADS. Given that the distribution of pronunciation variants was so similar in IDS and ADS, it is interesting that adult listeners were more accurate when listening to ADS than IDS. There are a number of possible explanations for this discrepancy.

One explanation may be found in prosody. Despite IDS and ADS tokens being elicited from identical texts, it is probable that the prosody was not identical and could have favored the ADS register (cf. Shattuck-Hufnagel & Turk, 1996). Furthermore, there was variety in the sentences used when retelling the story that likely had an effect on the prosodic structure of the target phrases. Given that prosodic structure is known to affect the phonetic realisation of segments (Cho, 2004; Cho, Kim, & Kim, 2017), and that listeners use prosodic structure in speech perception (Durvasula & Kahng, 2016; Mitterer, Cho, & Kim, 2016) we cannot rule out the possibility that prosody influenced our data. In order to understand the structure of IDS more completely future research should consider prosodic structure in more detail.

Familiarity with the two speech registers may have played a role in our data, as adults typically have more experience with ADS than IDS.[2] However, previous research has shown

[2] Pitch deviations in IDS may be greater than in ADS, and this may be particularly difficult or distracting for adult listeners. We manipulated the pitch of the tokens used in Experiment 2 such that it was flattened to the mean F0, and a group of adults participated in the same identification task ($N$ = 32). Even with this pitch manipulation we find the same pattern of results of Experiments 1 and 2, namely increased accuracy

that adults are adept at processing IDS (e.g., Golinkoff & Alioto, 1995; Jesse & Johnson, 2012). Bard and Anderson (1983, 1994) argue that adults find words spliced out of ADS more intelligible than IDS because IDS contains more extralinguistic cues and speakers use sentence structure to help the infant. However, this cannot be the only factor at play in our data because even in Experiment 2, where listeners heard the whole sentence containing the target phrase, ADS was still more intelligible.

An additional finding that warrants discussion is that listeners were more accurate at identifying the target in a context that licenses assimilation than one that does not, although it should be noted that this did not differ by speech register and as such does not impact on the primary question of investigation. On first impressions this result seems surprising, as tokens with a non-assimilating context (e.g. *cap box*) are expected to show less variation than tokens where assimilation may occur. However, there are a few possible explanations for this finding. One possibility is that by presenting the two orthographic forms, participants' awareness of place assimilation was activated and influenced their decision-making. If they perceived *cap box*, there are two possible alternative interpretations, and both of these are presented visually; either it is a surface-match of the intended target *cap box*, or it is an assimilated pronunciation of the intended target *cat box*. In the non-assimilating context participants' accuracy was not much higher than chance (58% in Experiment 1, 56% in Experiment 2), suggesting that when only cues to a labial place of articulation were heard they entertained each possible interpretation as plausible. However, if any acoustic cue relating to a coronal place of articulation is perceived, then the choice is restricted as it is much more likely that the intended target is *cat box* and not *cap box*. The classification analysis of tokens with a context that licensed assimilation indicated that only 12.5% of tokens were completely assimilated, meaning that the majority of the tokens likely had some acoustic cue to a coronal place of articulation. Although previous studies have found that that when assimilation gives rise to lexical ambiguity listeners accurately extract the intended word-form and do not access the unintended lexical item (Gaskell & Marslen-Wilson,

---

for ADS tokens than IDS tokens, and increased accuracy in a context that licenses assimilation than one that does not.

2001; Gow, 2002), these studies used a priming paradigm, an online measure, rather than an offline task.

Listeners may also be making use of frequency information. Tokens that licensed assimilation tended to be of higher frequency than tokens that did not license assimilation, and listeners may have assumed that the token heard was the more frequently occurring of the pair displayed. Frequency was not a factor in our primary manipulation of interest, and this was not one of the key criteria used when creating stimuli for this study. All phrases were recorded in both ADS and IDS, ensuring that frequency information should affect both registers equally, and not affect our measure of interest.

Data from Experiments 1, 2 and the classification analysis all converge on the finding that IDS is not less ambiguous than ADS. However, all three analyses are based on the same 224 tokens. While this data set is comparable in size to previous studies (cf. Lahey & Ernestus, 2013), it is only a small subset of the data available in the corpus of laboratory speech collected. To extend the scope of our analysis, we now present the results of more detailed acoustic analyses of all analysable tokens in the corpus.

**4.0 Acoustic Analysis**

The results of Section 3 suggest that parents do not use less assimilation in IDS than ADS. Here we test whether those results generalise to a wider sample of tokens. The primary variable of interest is variation in the frequency of the second formant (F2) at the end of the vowel, as the F2 is affected by the place of articulation of the following segment. This is the most frequently reported measure of place assimilation in the literature (e.g. Dilley & Pitt, 2007; Gow, 2001, 2002, 2003; Zimmerer et al., 2009). We also included another measure that has been used in studies of place assimilation; variation in the amplitude of the second formant (A2) at the end of the vowel (Gow, 2001, 2002, 2003). In addition, we examined variation in two acoustic features that are known to vary between IDS and ADS, namely fundamental frequency (F0, the primary acoustic correlate of pitch) and vowel duration (e.g. Albin & Echols, 1996; Fernald et al., 1989).

*4.1 Materials and Analysis Methods*

The method for creating the Laboratory Corpus is detailed in Section 2. There were 1743 tokens in the corpus, however 301 tokens were excluded due to disfluencies in the speech, poor audio quality or noise. Data from the remaining 1442 tokens are presented.

Start and end points of the vowel in the first word of the two-word phrase (i.e. the [æ] of *cat* in the phrase *cat box*) were marked manually in *Praat* (Boersma & Weenink, 2011). The frequency and intensity of the second formant at the endpoint of the vowel, as well as vowel duration and mean pitch were extracted automatically using a custom *Praat* script. Formant information was verified by hand, and where the formant estimates generated by *Praat* deviated from the spectrogram the spectrogram reading was followed.

*4.2 Results & Discussion*

There were 1442 tokens from 12 speakers included in the analysis; each speaker contributed between 113 and 181 tokens. 812 tokens (56.3%) were uttered in ADS, and 630 (43.7%) in IDS. For the purposes of the acoustic analysis items were clustered into five Vowel Groups based on the vowel of the first word of the phrase. For example, *ape* and *grape* were grouped together, as were *beam* and *team*. Table 3 presents the number of tokens broken down by vowel group and IDS/ADS.

*Table 3. Number of tokens of each two-word phrase analysed in the corpus of laboratory speech, and whether uttered in adult-directed speech (ADS) or infant-directed speech (IDS). Phrases are clustered according to the vowel of the first word.*

| Vowel Group | Phrase | ADS | IDS |
|---|---|---|---|
| eɪ | Ape babies / Grape pie | 61 | 50 |
| | Eight babies / Great pie | 68 | 52 |
| i: | Beam painter / Team bears | 114 | 75 |
| | Bean painter / Teen bears | 98 | 77 |

| | | | |
|---|---|---|---|
| æ | Cap box / Cap burglar | 107 | 67 |
| | Cat box / Cat burglar | 81 | 68 |
| oʊ | Comb maker | 47 | 41 |
| | Cone maker | 66 | 41 |
| ɛ | Jem Pickles | 78 | 78 |
| | Jen Pickles | 92 | 81 |

Variation in each of the four features of interest (F2, A2, F0 and duration) was analysed using a linear mixed effects model with fixed effects of Addressee (IDS or ADS), Scriptedness (scripted or unscripted) and Assimilation Licensing Context (yes or no), and all two- and three-way interactions of these. Random intercept terms were included for Speaker and Vowel group, as well as random slopes for Assimilation Licensing Context by Speaker and Vowel Group. In the interest of space we report only a selection of the results that relate to differences between IDS and ADS, however the complete model outputs are presented in Appendix A for reference.

The frequency of the second formant reflects place of articulation of the following segment, with a lower F2 expected prior to a labial than a coronal segment. We find an effect of Assimilation Licensing Context ($\beta_{\text{Assimilation}} = 90$, $SE = 21.99$, $\chi^2(1) = 16.65$, $p < .001$), with lower F2 measured in a non-assimilating context ($M_{\text{NoLicensingContext}} = 1865.2$, $SD = 187$) than a context that licenses assimilation ($M_{\text{LicensingContext}} = 1952.77$, $SD = 158.35$). This is expected, given that there are more labial pronunciations as in a context that does not license assimilation (e.g. *cap box*) than in a context where a labial pronunciation is optional (e.g. *cat box*). We find a significant effect of Addressee, $\beta_{\text{Addressee}} = -72.65$, $SE = 22.11$, $\chi^2(1) = 10.76$, $p = .001$. F2 is lower in ADS ($M_{\text{ADS}} = 1868.03$, $SD = 126.59$) than IDS ($M_{\text{IDS}} = 1949.55$, $SD = 128.04$), which is consistent with previous literature that formant frequencies increase in IDS (Benders, 2013; Englund & Behne, 2005). Of particular interest to the current paper is the interaction term of Addressee and Assimilation Licensing Context. If IDS contains more canonical tokens, and fewer assimilated tokens, in an assimilation-licensing context, then we expect the difference in F2 between the assimilating and non-assimilating context to be greater for IDS than ADS. This

is not the case, and there is no significant interaction of Addressee and Assimilation Licensing Context ($\beta_{\text{Addressee:Assimilation}}$ = -21.78, $SE$ = 44.02, $\chi^2(1)$ = 0.24, $p$ = .62; $M_{IDS,\ NoLicensingContext}$ = 1899.23, $SD$ = 193.24, $M_{IDS,\ LicensingContext}$ = 2026.29, $SD$ = 234.83; $M_{ADS,\ NoLicensingContext}$ = 1839.69, $SD$ = 199.28, $M_{ADS,\ LicensingContext}$ = 1899.03, $SD$ = 131.5). This data is presented graphically in Figure 3. No other effects or interactions were significant.

Note to Publisher: Insert Figure 3 about here

In the analysis of the intensity of the second formant at the end of the vowel (A2) we find, as expected, an effect of Assimilation Licensing Context, $\beta_{\text{Assimilation}}$ = -1.49, $SE$ = 0.48, $\chi^2(1)$ = 9.62, $p$ = .002. The intensity of the second formant is higher in a context that does not license assimilation ($M_{\text{NoLicensingContext}}$ = 14.97, $SD$ = 9.42) than a context where assimilation is licensed ($M_{\text{LicensingContext}}$ = 13.4, $SD$ = 8.63). The effect of Addressee is marginally significant ($\beta_{\text{Addressee}}$ = -0.88, $SE$ = 0.48, $\chi^2(1)$ = 3.31, $p$ = .07), with slightly higher intensity for tokens uttered in IDS than ADS ($M_{\text{IDS}}$ = 15, $SD$ = 8.91; $M_{\text{ADS}}$ = 13.67, $SD$ = 9.17). The interaction of Addressee and Assimilation Licensing Context is not significant, again indicating similar differences in intensity between a context that licenses assimilation and a context that does not in IDS and ADS ($\beta_{\text{Addressee:Assimilation}}$ = -0.3, $SE$ = 0.96, $\chi^2(1)$ = 0.1, $p$ = .75; $M_{IDS,\ NoLicensingContext}$ = 15.78, $SD$ = 10.08, $M_{IDS,\ LicensingContext}$ = 14.16, $SD$ = 8.31; $M_{ADS,\ NoLicensingContext}$ = 14.58, $SD$ = 9.31, $M_{ADS,\ LicensingContext}$ = 12.77, $SD$ = 9.34).

IDS typically has a higher pitch and slower speech rate than ADS, and that is also the case in our data. In the analysis of mean pitch of the vowel we find a significant effect of Addressee ($\beta_{\text{Addressee}}$ = -7.65, $SE$ = 2.44, $\chi^2(1)$ = 9.75, $p$ = .002), with a higher pitch in IDS than ADS ($M_{\text{IDS}}$ = 215.54, $SD$ = 11.14; $M_{\text{ADS}}$ = 207.81, $SD$ = 14.64). Tokens with an assimilation-licensing context are higher in pitch than tokens that do not license assimilation ($\beta_{\text{Assimilation}}$ = 5, $SE$ = 2.48, $\chi^2(1)$ = 4.04, $p$ = .04; $M_{\text{LicensingContext}}$ = 212.59, $SD$ = 15.32; $M_{\text{NoLicensingContext}}$ = 208, $SD$ = 11.98), and there is a significant interaction of Addressee and Assimilating Licensing Context, $\beta_{\text{Addressee:Assimilation}}$ = 11.64, $SE$ = 4.86, $\chi^2(1)$ = 5.71, $p$ = .02. This interaction reflects the difference

in pitch between tokens with an assimilation context and a non-assimilating context is larger in ADS than IDS ($M_{IDS,\ NoLicensingContext}$ = 216.06, $SD$ = 12.6, $M_{IDS,\ LicensingContext}$ = 214.17, $SD$ = 16.14; $M_{ADS,\ NoLicensingContext}$ = 201.97, $SD$ = 15.09, $M_{ADS,\ LicensingContext}$ = 212.76, $SD$ = 21.5).

Vowel duration is longer in IDS than ADS, reflecting the slower speech rate of IDS ($\beta_{\text{Addressee}}$ = -18.76, $SE$ = 1.85, $\chi^2(1)$ = 99.6, $p$ < .001; $M_{\text{IDS}}$ = 118.76, $SD$ = 20.23; $M_{\text{ADS}}$ = 100.06, $SD$ = 12.63). Vowel duration is also longer in an assimilation-licensing context than a non-licensing context ($\beta_{\text{Assimilation}}$ = 4.78, $SE$ = 1.84, $\chi^2(1)$ = 6.75, $p$ = .009; $M_{\text{LicensingContext}}$ = 110.08, $SD$ = 15.11; $M_{\text{NoLicensingContext}}$ = 106.65, $SD$ = 17.84). There is no significant interaction of Addressee and Assimilation Licensing Context, $\beta_{\text{Addressee:Assimilation}}$ = 0.07, $SE$ = 3.68, $\chi^2(1)$ = .0003, $p$ = .99). There is a significant effect of Scriptedness on vowel duration, and, as expected, scripted, read tokens have a longer vowel duration than unscripted tokens ($\beta_{\text{Scriptedness}}$ = -9.29, $SE$ = 1.87, $\chi^2(1)$ = 24.67, $p$ < .001; $M_{\text{Scripted}}$ = 104.94, $SD$ = 14.34; $M_{\text{Unscripted}}$ = 114.5, $SD$ = 18.8). There is no significant interaction of Addressee and Scriptedness, indicating that the change in speech rate between scripted and unscripted speech is similar in IDS and ADS, $\beta_{\text{Addressee:Scriptedness}}$ = 3.32, $SE$ = 3.68, $\chi^2(1)$ = 0.82, $p$ = .37.

In summary, we find expected differences in the speech patterns of mothers' IDS and ADS. Specifically, when talking to their children they speak in a higher pitch and with a slower speech rate. We also find expected differences between read and spontaneous speech, and read speech is slower than unscripted speech. Regarding assimilation patterns, we also find patterns in the frequency and intensity of the second formant that are consistent with previous literature. However, in neither of these measures do we find evidence to indicate that parents' use of assimilation differs depending on whether they are speaking to their child or another adult. That is, we do not find support for the hypothesis that parents are reducing their use of connected speech processes, or using more canonical pronunciations, in IDS. This supports the findings of the perceptual experiments reported in Section 3, and indicates that the subset of tokens used as stimuli in those experiments were representative of the corpus as a whole.

## 5.0 Corpus of Spontaneous Speech

All data reported so far converge on the finding that parents are not spoon-feeding their children canonical, or citation, forms, but presenting children with a complex linguistic input that includes a variety of pronunciation variants. Taken together with the findings of Dilley et al., (2014), one could conclude that there is little difference in the use of connected speech processes in IDS and ADS. However, both our study and that of Dilley et al., (2014) analyzed recordings of parents reading to their children in a laboratory. Although we included unscripted speech, the manner in which it was recorded did not necessarily elicit truly spontaneous speech. There are important reasons for conducting analyses on elicited speech, primarily because it allows for greater experimental control (e.g. enabling elicitation of lexically ambiguous stimuli, and controlling the environment to obtain high-quality audio recordings). Nevertheless, the control gained by eliciting speech in the lab comes at a cost; namely, this does not resemble a natural speech situation and spontaneous interactions between parents and their children. In the next section we supplement our analysis of elicited speech with an analysis of how connected speech processes are realized in a corpus of spontaneous mother-child interactions. As well as complementing our previous data, this is one of the first descriptive analyses of how place assimilation is realized in spontaneous IDS in English.

### *5.1 Methods*

#### *5.1.1 The Corpus*

Data was taken from the Providence Corpus (Demuth, Culbertson, & Alter, 2006), accessed through the CHILDES database (MacWhinney, 2000). This corpus includes audio and video recordings from 6 children (3 boys, 3 girls) in spontaneous interactions with their mothers in their own homes. Recordings were made every two-weeks between the ages of 1 and 3 years. For the current study we restricted the corpus to sessions where the child was between 1 year 5 months, and 1 year and 8 months old. This age-range was comparable to the children whose mothers we recorded in a laboratory setting. The resulting subset of the corpus contained 39 recording sessions of approximately 1 hour in duration, from 6 children and their mothers. There were an average of 6.5 hours of recordings for each child.

*5.1.2 Analysis*

Using the orthographic transcriptions available for the corpus, we used the CLAN software (MacWhinney, 2000) to search for instances where the mother uttered two words in succession that would create a context where place assimilation could occur. That is, we identified all words that had a final coronal segment (/t, d, n/), followed by a word with an initial labial segment (/p, b, m/). We restricted our search to labial place of assimilation and did not include velar segments, which may also trigger place assimilation in English. We further restricted our search to contexts where the first word of the pair ended in a VC segment and avoided word-final clusters. Once the target contexts had been identified, orthographic transcriptions were aligned with time-stamps in the audio recordings to identify tokens for analysis.

There were 1463 tokens identified from the orthographic transcriptions of the corpus. The first author listened to all identified tokens to identify which were suitable for analysis based on the mothers' speech style, whether the phonological context was met, and audio quality. Tokens were excluded from analysis if the mother was reading to the child ($N$ = 170), if they were sung or whispered ($N$ = 32), if the mother was talking to another adult present and not the child (i.e. ADS and not IDS) ($N$ = 136), if the context of interest was not met, for example due to the presence of a phrase boundary or pause between the words of interest or inaccuracies in the orthographic transcription ($N$ = 81), or if the audio quality was not sufficient to analyse the speech, for example due to noise, such as a child's toy, a child's vocalisation, or environmental noise (e.g. traffic or the radio playing) ($N$ = 208). Of the original tokens identified, 627 were excluded, providing 836 tokens of spontaneous IDS from six mothers with reasonable audio quality. Mean number of tokens from each speaker was 126, with a range of 55-247. All tokens were a two-word phrase containing a context that licensed place assimilation.

Two phonetically-trained coders analysed the 836 tokens independently, adhering to the same categories established by Dilley and Pitt (2007) and Dilley et al. (2014). Tokens were

classified as Canonical, Assimilated, Glottalized or Deleted. Tokens were classified as Canonical if the final segment of the first word was present, without voicing irregularity. Tokens were classified as Assimilated if there was evidence of a downward trajectory in the preceding F2. Tokens were classified as Deleted if there was no spectrographic evidence for the final segment of the word being present, for example, a very short closure phase in a C#C sequence. Finally, a token was given the classification of Glottalized if there was irregularity in the timing of pitch pulses in the waveform. Examples of each of these variants are presented in Figure 4.

Once the two coders had classified all tokens, the two sets of classification data were compared. Where coders disagreed, the tokens with disagreement were identified and the coders given the opportunity to reassess their classification. There was almost perfect agreement between the two coders (90%, $\kappa = 0.87$; Landis & Koch, 1977). Of the 836 tokens classified, coders disagreed on 83 tokens. We report data from the 753 tokens that coders agreed upon.

Note to Publisher: Insert Figure 4 about here

*5.2 Results & Discussion*

Within the whole sample, 182 tokens (24%) were realised with a Canonical pronunciation. Some three-quarters of the pronunciations in this sample of spontaneous IDS, therefore, were produced with a non-canonical pronunciation. Deleted variants were produced 249 times (33%), Assimilated 168 (22%) and Glottalized 154 (21%).

The corpus of spontaneous IDS included phrases in which the first word was either a function word ($N = 475$, 63%) or a content word ($N = 278$, 37%). Function words and content words differ in their phonological and phonetic behaviour, and function words are more susceptible to pronunciation variation (e.g. Ogden, 1999; Zimmerer, Reetz, & Lahiri, 2009). Given that 63% of our tokens were function words the number of canonical pronunciations is unsurprising. Table 4 presents the distribution of pronunciation variants in canonical and function words separately. As expected, canonical pronunciations were more frequent in content

words (*N* = 103, 37%) than function words (*N* = 79, 16%). However, even when we look only at content words, 63% of the tokens that the child is hearing do not have a canonical realization.

*Table 4. Pronunciation variation in tokens that license assimilation in a corpus of spontaneous mother-child interactions. Tokens were classified as having a Canonical, Assimilated, Glottalized or Deleted pronunciation. Results are broken down by whether the first word of the phrase was a content word or a function word.*

| | Canonical (% & *N*) | Assimilated (% & *N*) | Glottalized (% & *N*) | Deleted (% & *N*) |
|---|---|---|---|---|
| Content words | 37 (103) | 18 (50) | 15 (41) | 30 (84) |
| Function words | 16 (79) | 25 (118) | 24 (113) | 35 (165) |
| Total | 24 (182) | 22 (168) | 21 (154) | 33 (249) |

The data indicate that children do not encounter a high proportion of canonical pronunciations in an assimilation-licensing context in spontaneous IDS; in fact, canonical pronunciations account for less than a quarter of tokens heard. Parents are not using fewer non-canonical pronunciations in IDS in order to break the learning problem down for their children, allowing them first to learn canonical forms and then later learn about pronunciation variation. This result is in line with the findings of the laboratory corpus (Section 3).

Children are exposed to many variable pronunciations. Parents do not only use either a canonical or an assimilated pronunciation in a context where assimilation is licensed; actually, around half of the time they use a different pronunciation, either glottalizing the final obstruent or deleting it entirely. These pronunciations are typical of spontaneous speech (Dilley & Pitt, 2007), further indicating that parents are not modifying their speech and making an effort to speak clearly to their children.

Although both the laboratory and spontaneous corpus data converge on the conclusion that parents are not simplifying their speech to their infants by reducing the variability in the

pronunciations they use, there are differences in the distribution of pronunciation variants in each data set. In the laboratory, parents used many more canonical pronunciations (i.e. citation form pronunciations) and fewer assimilated pronunciations than in spontaneous speech. This difference is likely a reflection of the recording situation and the speech material analysed. In the laboratory, even the spontaneous speech we recorded was more careful speech than the speech of parents in their home environment. Furthermore, in the laboratory we recorded parents reading and retelling a story where the phrases of interest were important for the narrative and the first word of the phrase was a content word. This encourages clearer pronunciation than in the spontaneous speech recordings that contained both function and content words, and often in less prosodically marked positions. That is, in the laboratory we elicited more careful speech than in parents' own homes, which in itself is not unexpected. However, even when parents are speaking more carefully in the laboratory they do not modify their IDS more than their ADS.

## 6.0 General Discussion

In order to become a competent user of their native language, children must learn to cope with connected speech processes, yet little is known about how this ability develops. This study examined how connected speech processes, specifically cases of place assimilation, are realized in speech addressed to infants. We elicited both IDS and ADS from parents of 18-month-old children that included many tokens of phrases that license place assimilation (e.g. *cat box*). Both perceptual and acoustic measures support the conclusion that IDS is not less ambiguous than ADS, it does not contain more canonical forms and parents do not use fewer connected speech processes. The generalizability of this conclusion was supported by a corpus analysis of spontaneous mother-child interactions. By gaining a better understanding of the nature of the child's linguistic input, we are in a better position to characterize the learning problem faced by the child, and lay groundwork for further research into how children's ability to cope with connected speech processes may develop.

It is important to consider how our results fit into the previous literature. In the classification analysis of tokens used in the experimental task approximately half of the IDS tokens were produced with a canonical pronunciation. This result is comparable to the one previous study investigating pronunciation variation in place of assimilation in English IDS (Dilley et al., 2014). Our data, and that of Dilley et al., (2014), therefore suggest that although canonical pronunciations are not the only form children hear, they are encountered relatively often in the child's input. However, both of these data sets are based on elicited speech recorded in a laboratory, and there are a number of reasons to believe that this may be a conservative measure of how often canonical forms are encountered. When we look at more naturalistic recordings, as we did in our analysis of the corpus of spontaneous mother-child interactions, we find closer to one quarter of tokens produced canonically. The difference in proportion of canonical tokens found in speech elicited in the laboratory and spontaneous utterances is striking, and highlights the potential limitations in the ecological validity of studying only elicited speech when investigating the acoustic properties of IDS. In order to get a true picture of the speech that infants hear it is important to combine evidence from laboratory elicited speech, where the experimenter has greater control, and spontaneous speech corpora in a more natural setting. Both types of analysis were included here, and despite differences, both point to the same conclusion that, with regard to connected speech processes, the child's input contains many non-canonical pronunciations. Having a better understanding of the child's input allows us to consider how it constrains our theoretical outlook. In what follows we speculate on the role of IDS in acquiring connected speech processes, how being exposed to much acoustic-phonetic variation may be beneficial to learning about connected speech processes, and what form the learning process may take.

IDS has often been argued to be clearer than ADS, and authors have used this as evidence that IDS has a didactic function (e.g. Burnham et al., 2002; Englund, 2005; Ferguson, 1964; Kuhl et al., 1997). In the domain of place assimilation in English we find no evidence that parents are increasing the clarity of their speech by increasing the number of canonical pronunciations. This is not inconsistent with the claim that IDS is a didactic device. Studies

where evidence for hyperarticulation have been attested typically investigate how parents mark phonemic contrasts in their IDS and ADS, for example by contrasting the size of the vowel space used (e.g. Kuhl et al., 1997), or the difference in VOT contrast for voiced or voiceless stops (e.g. Englund, 2005). It is somewhat intuitive that learning categories may be easier if the two distributions were further apart in acoustic space, and there was less overlap between the two categories. However, in the case of connected speech processes, it is not clear whether increasing the clarity of speech by increasing the number of canonical forms used would create a more optimal learning situation for the child. Parents would be helping their child learn the citation forms of words, but not about connected speech processes. Effectively they would break the learning process down into stages for their child, where the child first learns canonical forms of words, and then at a later stage learns about possible pronunciation variants in different contexts in connected speech. This could potentially speed the acquisition of specific lexical forms, but would not help children learn about connected speech processes in their language. Simultaneous exposure to both canonical forms and variations that occur in connected speech may be beneficial to learning about connected speech processes, as the learning problem does not have to comprise of discrete stages.

If parents are not reducing the learning problem into bite-sized chunks for their children, the question arises as to how children learn to compensate for connected speech processes from the input they receive. For a given lexical item, how do children learn which form is the canonical form, which is a context-dependent variant, and which is a non-context-dependent variant? One likely possibility is that infants use distributional statistics. Infants are highly sensitive to statistical and distributional information in their input, and are able to make use of it during the early stages of language acquisition (Anderson, Morgan, & White, 2003; Chambers, Onishi, & Fisher, 2003; Maye, Weiss, & Aslin, 2008; Maye, Werker, & Gerken, 2002; Saffran, Aslin, & Newport, 1996). One study particularly relevant to learning about connected speech processes, showed that children as young as 12-months-old can use distributional information alone to learn phonological alternations when they occur in complementary distribution (White, Peperkamp, Kirk, & Morgan, 2008). How exactly the infant

uses distributional statistics to learn about connected speech processes depends somewhat on what the initial state is with regard to connected speech processes, as this dictates the nature of the learning task.

One view is that infants initially have no knowledge of connected speech processes and must learn canonical forms, legitimate variations, and information about the context that governs the change. Once these are acquired the child has the knowledge to compensate for the variation that they encounter. Peperkamp and Dupoux (2002) describe a theoretical account of how infants may be able to use distributional information to learn about alternations such as place assimilation. They predict that infants will note that clusters of a coronal segment followed by a non-coronal segment never occur within an intonational phrase, and from this derive a rule of assimilation. For instance, infants hear both *green* and *greem* phrase-medially depending on the following word, but only *green* phrase finally. A word such as *arm*, however, surfaces with a final [m] in all positions. From this generalization the child can hypothesize that *greem* is the assimilated form of the underlying *green*. This account seems credible if the infant were faced with just two pronunciation variants that depend on the context. However, we have established that this is not the case. In the present study, we have found that the input the child receives does not contain complete categorical shifts and contains a great deal of variation. More research is needed to address how well this hypothesised learning mechanism copes with ecologically valid input, and whether it scales-up to a real-world learning situation.

An alternative view is that infants are initially able to compensate for connected speech processes present in all languages, and in learning their native language they 'unlearn' the ability to compensate for patterns not present in their language. This view is grounded in the argument that assimilation patterns attested in the world's languages are natural, reflecting universal tendencies to produce simultaneous sounds that are more acoustically similar to one another (Donegan & Stampe, 1979; Smolensky, 1996). In this case, it would be beneficial to the child to hear a variety of pronunciations that are licensed in a given context in the input they receive from their caregivers. By hearing both canonical and assimilated forms, as well as other legitimate variations, the child can track which sounds co-occur and which do not, and the

patterns of the native language will be reinforced. Thus, the ability to compensate for native language patterns is retained, while non-native language patterns are not supported and the ability to compensate for them will eventually be lost. The ability to compensate for assimilation patterns in the native language but not from other languages is argued to already be in place by two years of age (Skoruppa, Mani, Plunkett, et al., 2013). Further data is needed from younger infants to know whether this is a result of having learned the native language pattern or unlearning the non-native language pattern.

To conclude, the present study has shown that IDS is not always clearer than ADS with regard to connected speech processes. Parents do not reduce the amount of assimilation they use in their speech to their infants in order to present them with many canonical, or citation forms of words. Remarkably little is known about how children learn to cope with connected speech processes in their native language, although it is widely accepted that the child's linguistic input is the primary source of information that they learn from. As such, determining the nature of connected speech processes in IDS forms an important foundation for further work and can tell us much about how the ability to compensate for assimilation may develop.

**Acknowledgements**

This research was funded by an NWO Rubicon grant awarded to Helen Buckler and SSHRC and NSERC grants awarded to Elizabeth K. Johnson. We thank Julie Kow for help designing stimuli and collecting recordings, and Kazuya Bamba and Natalie Fecher for assistance with classifying tokens.

**Reference List**


Albin, D. D., & Echols, C. H. (1996). Stressed and word-final syllables in infant-directed speech. *Infant Behavior and Development*, *19*(4), 401–418. doi:10.1016/S0163-6383(96)90002-8

Anderson, J. L., Morgan, J. L., & White, K. S. (2003). A statistical basis for speech sound discrimination. *Language and Speech*, *46*(2–3), 155–182. doi:10.1177/00238309030460020601

Bard, E. G., & Anderson, A. H. (1983). The unintelligibility of speech to children. *Journal of Child Language*, *10*(2), 265–292. doi:10.1017/S0305000900007777

Bard, E. G., & Anderson, A. H. (1994). The unintelligibility of speech to children: effects of referent availability. *Journal of Child Language*, *21*(3), 623–648. doi:10.1017/S030500090000948X

Barr, D. J. (2008). Analyzing "visual world" eyetracking data using multilevel logistic regression. *Journal of Memory and Language*, *59*(4), 457–474. doi:10.1016/j.jml.2007.09.002

Barr, D. J., Levy, R., Scheepers, C., & Tily, H. J. (2013). Random effects structure for confirmatory hypothesis testing: Keep it maximal. *Journal of Memory and Language*, *68*(3), 255–278. doi:10.1016/j.jml.2012.11.001

Bates, D., Maechler, M., & Bolker, B. (2015). Package "lme4." Retrieved from http://lme4.r-forge.r-project.org/

Benders, T. (2013). Mommy is only happy! Dutch mothers' realisation of speech sounds in infant-directed speech expresses emotion, not didactic intent. *Infant Behavior and Development*, *36*(4), 847–862. doi:10.1016/j.infbeh.2013.09.001

Bernstein Ratner, N. (1984). Patterns of vowel modification in mother-child speech. *Journal of Child Language*, *11*(3), 557–578. Retrieved from internal-pdf:/BernsteinRatner_1984.pdf

Blomert, L., Mitterer, H., & Paffen, C. (2004). In search of the auditory, phonetic, and/or phonological problems in dyslexia. *Journal of Speech, Language, and Hearing Research*, *47*(5), 1030–1047. doi:10.1044/1092-4388(2004/077)

Boersma, P., & Weenink, D. (2011). Praat: Doing Phonetics by Computer (version 5.2.33).

Burnham, D., Kitamura, C., & Vollmer-Conna, U. (2002). What's new, pussycat? On talking to babies and animals. *Science*, *296*, 1435.

Chambers, K. E., Onishi, K. H., & Fisher, C. (2003). Infants learn phonotactic regularities from brief auditory experience. *Cognition*, *87*, B69-77. doi:10.1016/S0

Chevrot, J.-P., Dugua, C., & Fayol, M. (2009). Liason acquisition word segmentation and construction in French: a usage-based account. *Journal of Child Language*, *36*(3), 557–596. doi:10.1017/S0305000908009124

Cho, T. (2004). Prosodically conditioned strengthening and vowel-to-vowel coarticulation in English. *Journal of Phonetics*, *32*(2), 141–176. doi:10.1016/S0095-4470(03)00043-3

Cho, T., Kim, D., & Kim, S. (2017). Prosodically-conditioned fine-tuning of coarticulatory vowel nasalization in English. *Journal of Phonetics*, *64*, 71–89. doi:10.1016/j.wocn.2016.12.003

Cristia, A., & Seidl, A. H. (2014). The hyperarticulation hypothesis of infant-directed speech. *Journal of Child Language*, *41*(4), 913–935. doi:10.1017/S0305000914000105

Darcy, I. (2002). Online processing of phonological variation in speech comprehension: The case of assimilation. In S. Hawkins & N. Nguyen (Eds.), *ISCA Tutorial and*

*Research Workshop (ITRW) on Temporal Integration in the Perception of Speech* (p. 32). ISCA.
Darcy, I., Peperkamp, S., & Dupoux, E. (2007). Bilinguals play by the rules: perceptual compensation for assimilation in late L2-learners. In J. Cole & J. Hualde (Eds.), *Papers in Laboratory Phonology 9* (pp. 411–442). Berlin: Mouton de Gruyter.
Darcy, I., Ramus, F., Christophe, A., Kinzler, K., & Dupoux, E. (2009). Phonological knowledge in compensation for native and non-native assimilation. In F. Kügler, C. Féry, & R. Van de Vijver (Eds.), *Variation and Gradience in Phonetics and Phonology* (pp. 265–309). Berlin: Mouton de Gruyter.
Demuth, K., Culbertson, J., & Alter, J. (2006). Word-minimality, epenthesis and coda licensing in the acquisition of English. *Language & Speech*, *49*(2), 137–174. doi:10.1177/00238309060490020201
Dilley, L. C., Millett, A. L., McAuley, J. D., & Bergeson, T. R. (2014). Phonetic variation in consonants in infant-directed and adult-directed speech: the case of regressive place assimilation in word-final alveolar stops. *Journal of Child Language*, *41*(1), 153–175. doi:10.1017/S0305000912000670
Dilley, L. C., & Pitt, M. A. (2007). A study of regressive place assimilation in spontaneous speech and its implications for spoken word recognition. *The Journal of the Acoustical Society of America*, *122*, 2340–2353. doi:10.1121/1.2772226
Donegan, P. J., & Stampe, D. (1979). The study of natural phonology. In D. A. Dinnsen (Ed.), *Current Approaches to Phonological Theory* (pp. 126–173). Bloomington, Indiana: Indiana University Press.
Dugua, C., Spinelli, E., Chevrot, J.-P., & Fayol, M. (2009). Usage-based account of the acquisition of liaison: Evidence from sensitivity to the singular/plural orientation of nouns. *Journal of Experimental Child Psychology*, *102*(3), 342–350. doi:10.1016/j.jecp.2008.07.006
Durvasula, K., & Kahng, J. (2016). The role of phrasal phonology in speech perception: What perceptual epenthesis shows us. *Journal of Phonetics*, *54*, 15–34. doi:10.1016/j.wocn.2015.08.002
Englund, K. (2005). Voice onset time in infant directed speech over the first six months. *First Language*, *25*(2), 219–234. doi:10.1177/0142723705050286
Englund, K., & Behne, D. (2005). Infant Directed Speech in Natural Interaction—Norwegian Vowel Quantity and Quality. *Journal of Psycholinguistic Research*, *34*(3), 259–280. doi:10.1007/s10936-005-3640-7
Englund, K., & Behne, D. (2006). Changes in infant directed speech in the first six months. *Infant and Child Development*, *15*(2), 139–160. doi:10.1002/icd.445
Fenson, L., Dale, P. S., Reznick, J. S., Bates, E., Thal, D. J., Pethick, S. J., … Stiles, J. (1994). Variability in Early Communicative Development. *Monographs of the Society for Research in Child Development*, *59*(5), 1–185.
Ferguson, C. A. (1964). Baby talk in six languages. *American Anthropologist*, (66), 103–114.
Fernald, A., Taeschner, T., Dunn, J., Papousek, M., de Boysson-Bardies, B., & Fukui, I. (1989). A cross-language study of prosodic modifications in mothers' and fathers' speech to preverbal infants. *Journal of Child Language*, *16*(3), 477–501. doi:10.1017/S0305000900010679
Fish, M. S., García-Sierra, A., Ramírez-Esparza, N., & Kuhl, P. K. (2017). Infant-directed speech in English and Spanish: Assessments of monolingual and bilingual caregiver VOT. *Journal of Phonetics*, *63*, 19–34. doi:10.1016/j.wocn.2017.04.003
Gaskell, M. G., & Marslen-Wilson, W. D. (2001). Lexical ambiguity resolution and spoken word recognition: Bridging the gap. *Journal of Memory and Language*, *44*,

325–349. doi:10.1006/jmla.2000.2741
Golinkoff, R. M., & Alioto, A. (1995). Infant-directed speech facilitates lexical learning in adults hearing Chinese: implications for language acquisition. *Journal of Child Language*, *22*, 703–726. doi:10.1017/S0305000900010011
Gow, D. W. (2001). Assimilation and Anticipation in Continuous Spoken Word Recognition. *Journal of Memory and Language*, *45*(1), 133–159. doi:10.1006/jmla.2000.2764
Gow, D. W. (2002). Does English coronal place assimilation create lexical ambiguity? *Journal of Experimental Psychology: Human Perception and Performance*, *28*(1), 163–179. doi:10.1037/0096-1523.28.1.163
Gow, D. W. (2003). Feature parsing: Feature cue mapping in spoken word recognition. *Perception & Psychophysics*, *65*(4), 575–590. Retrieved from internal-pdf:/Gow_2003.pdf
Jesse, A., & Johnson, E. K. (2012). Prosodic Temporal Alignment of Co-speech Gestures to Speech Facilitates Referent Resolution. *Journal of Experimental Psychology: Human Perception and Performance*, *38*(6), 1567–1581. doi:10.1037/a0027921
Kuhl, P. K., Andruski, J. E., Chistovich, I. A., Chistovich, L. A., Kozhevnikova, E. V, Ryskina, V. L., … Lacerda, F. (1997). Cross-Language Analysis of Phonetic Units in Language Addressed to Infants. *Science*, *277*, 684–686. doi:10.1126/science.277.5326.684
Kuhl, P. K., Conboy, B. T., Coffey-Corina, S., Padden, D., Rivera-Gaxiola, M., & Nelson, T. (2008). Phonetic learning as a pathway to language: new data and native language magnet theory expanded (NLM-e). *Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences*, *363*(1493), 979–1000. doi:10.1098/rstb.2007.2154
Lahey, M., & Ernestus, M. (2013). Pronunciation variation in infant-directed speech: Phonetic reduction of two highly frequent words. *Language Learning and Development*, *10*(4), 308–327. doi:10.1080/15475441.2013.860813
Lahiri, A., & Marslen-Wilson, W. D. (1991). The mental representation of lexical form: A phonological approach to the recognition lexicon. *Cognition*, *38*, 245–294. doi:10.1016/0010-0277(91)90008-R
Landis, J. R., & Koch, G. G. (1977). The measurement of observer agreement for categorical data. *Biometrics*, *33*(1), 159–174.
Liu, H.-M., Kuhl, P. K., & Tsao, F.-M. (2003). An association between mothers' speech clarity and infants' speech discrimination skills. *Developmental Science*, *6*(3), F1–F10.
MacWhinney, B. (2000). *The CHILDES Project: Tools for analyzing talk.* (Third Edit). Mahwah, NJ: Lawrence Erlbaum Associates.
Marshall, C. R., Ramus, F., & van der Lely, H. (2011). Do children with dyslexia and/or specific language impairment compensate for place assimilation? Insight into phonological grammar and representations. *Cognitive Neuropsychology*, *27*(7), 563–586. doi:10.1080/02643294.2011.588693
Maye, J., Weiss, D. J., & Aslin, R. N. (2008). Statistical phonetic learning in infants: facilitation and feature generalization. *Developmental Science*, *11*(1), 122–134. doi:10.1111/j.1467-7687.2007.00653.x
Maye, J., Werker, J. F., & Gerken, L. (2002). Infant sensitivity to distributional information can affect phonetic discrimination. *Cognition*, *82*(3), B101–B111. doi:10.1016/S0010-0277(01)00157-3
Mitterer, H., & Blomert, L. (2003). Coping with phonological assimilation in speech

perception: evidence for early compensation. *Perception & Psychophysics*, *65*(6), 956–969. doi:10.3758/BF03194826
Mitterer, H., Cho, T., & Kim, S. (2016). How does prosody influence speech categorization? *Journal of Phonetics*, *54*, 68–79. doi:10.1016/j.wocn.2015.09.002
Mitterer, H., Csépe, V., & Blomert, L. (2006). The role of perceptual integration in the recognition of assimilated word forms. *Quarterly Journal of Experimental Psychology (2006)*, *59*(8), 1395–1424. doi:10.1080/17470210500198726
Mitterer, H., Csépe, V., Honbolygo, F., & Blomert, L. (2006). The recognition of phonologically assimilated words does not depend on specific language experience. *Cognitive Science*, *30*(3), 451–479. doi:10.1207/s15516709cog0000_57
Mitterer, H., & Tuinman, A. (2012). The role of native-language knowledge in the perception of casual speech in a second language. *Frontiers in Psychology*, *3*, 1–13. doi:10.3389/fpsyg.2012.00249
Nakamura, M., Iwano, K., & Furui, S. (2008). Differences between acoustic characteristics of spontaneous and read speech and their effects on speech recognition performance. *Computer Speech & Language*, *22*(2), 171–184. doi:10.1016/j.csl.2007.07.003
Ogden, R. (1999). A declarative account of strong and weak auxiliaries in english. *Phonology*, *16*(1), 55–92. doi:10.1017/S095267579900370X
Otake, T., Yoneyama, K., Cutler, A., & van der Lugt, A. (1996). The representation of Japanese moraic nasals. *The Journal of the Acoustical Society of America*, *100*(6), 3831–3842. doi:10.1121/1.417239
Peperkamp, S., & Dupoux, E. (2002). Coping with phonological variation in early lexical acquisition. In I. Lasser (Ed.), *The Process of Language Acquisition* (pp. 359–385). Berlin: Peter Lang Verlag.
R Core Team. (2012). R: A language and environment for statistical computing. Vienna, Austria: R Foundation for Statistical Computing. Retrieved from http://www.r-project.org
Saffran, J. R., Aslin, R. N., & Newport, E. L. (1996). Statistical learning by 8-month-old infants. *Science*, *274*(5294), 1926–1928. doi:10.1126/science.274.5294.1926
Scott, D. R., & Cutler, A. (1984). Segmental phonology and the perception of syntactic structure. *Journal of Verbal Learning and Verbal Behavior*, *23*(4), 450–466. doi:10.1016/S0022-5371(84)90291-3
Shattuck-Hufnagel, S., & Turk, A. E. (1996). A prosody tutorial for investigators of auditory sentence processing. *Journal of Psycholinguistic Research*, *25*(2), 193–247. doi:10.1007/BF01708572
Shockey, L., & Bond, Z. S. (1980). Phonological processes in speech addressed to children. *Phonetica*, *37*, 267–274.
Skoruppa, K., Mani, N., & Peperkamp, S. (2013). Toddlers' processing of phonological alternations: Early compensation for assimilation in English and French. *Child Development*, *8*(1), 313–330.
Skoruppa, K., Mani, N., Plunkett, K., Cabrol, D., & Peperkamp, S. (2013). Early word recognition in sentence context: French and English 24-month-olds' sensitivity to sentence-medial mispronunciations and assimilations. *Infancy*, *18*(6), 1007–1029. doi:10.1111/infa.12020
Smolensky, P. (1996). The initial state and "Richness of the Base" in Optimality Theory. *Rutgers Optimality Arhive*, *293*. Retrieved from http://roa.rutgers.edu
Tuinman, A., Mitterer, H., & Cutler, A. (2011). Perception of intrusive /r/ in English by native, cross-language and cross-dialect listeners. *The Journal of the Acoustical*

*Society of America*, *130*(3), 1643–52. doi:10.1121/1.3619793
Uther, M., Knoll, M. A., & Burnham, D. (2007). Do you speak E-NG-L-I-SH? A comparison of foreigner- and infant-directed speech. *Speech Communication*, *49*, 2–7. doi:10.1016/j.specom.2006.10.003
Warner, N., & Tucker, B. V. (2011). Phonetic variability of stops and flaps in spontaneous and careful speech. *The Journal of the Acoustical Society of America*, *130*(3), 1606–1617. doi:10.1121/1.3621306
Weber, A. (2001). Help or hindrance: How violation of different assimilation rules affects spoken-language processing. *Language and Speech*, *44*(1), 95–118. doi:10.1177/00238309010440010401
Werker, J. F., Pons, F., Dietrich, C., Kajikawa, S., Fais, L., & Amano, S. (2007). Infant-directed speech supports phonetic category learning in English and Japanese. *Cognition*, *103*(1), 147–162. doi:10.1016/j.cognition.2006.03.006
White, K. S., Peperkamp, S., Kirk, C. J., & Morgan, J. L. (2008). Rapid acquisition of phonological alternations by infants. *Cognition*, *107*(1), 238–265. doi:10.1016/j.cognition.2007.11.012
Xu Rattanasone, N., Burnham, D., Kitamura, C., & Vollmer-Conna, U. (2013). Vowel hyperarticulation in parrot-, dog- and infant-directed speech. *Anthrozoos*, *26*(3), 373–380.
Zimmerer, F., Reetz, H., & Lahiri, A. (2009). Place assimilation across words in running speech: corpus analysis and perception. *The Journal of the Acoustical Society of America*, *125*(April 2009), 2307–2322. doi:10.1121/1.3021438

## Figures

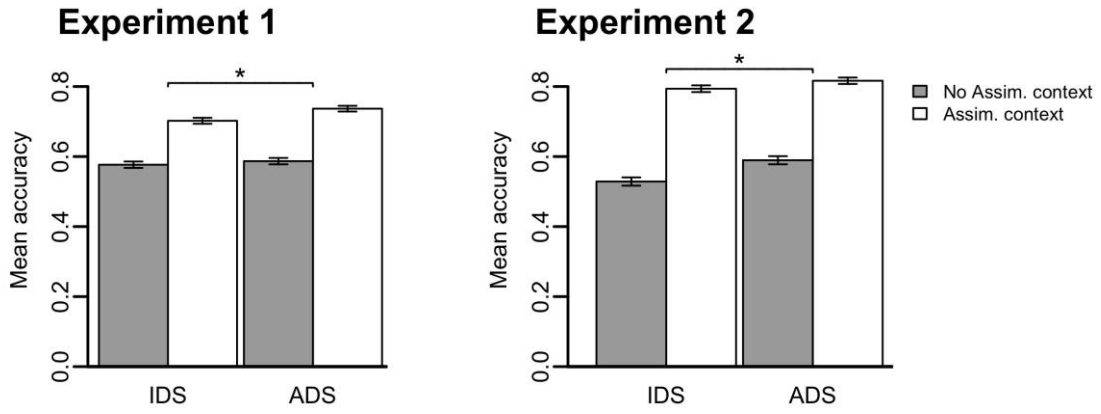


Figure 1

Mean identification accuracy of phrases spoken in infant-directed speech (IDS) and adult-directed speech (ADS) in Experiments 1 and 2. Targets were potentially ambiguous two-word phrases that did or did not license place assimilation (e.g. *cat box* and *cap box*), presented in isolation (Experiment 1) or whole sentences (Experiment 2). In both experiments adult listeners were more accurate in identifying the intended target in ADS than IDS.

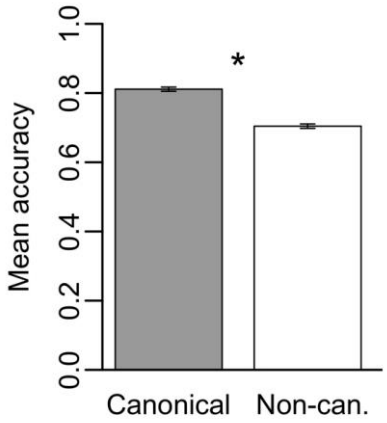


Figure 2

Mean identification accuracy of two-word phrases containing an assimilation-licencing context by participants in Experiments 1 and 2. Results are divided by pronunciation, either Canonical or Non-canonical, as identified by three phonetically trained classifiers.

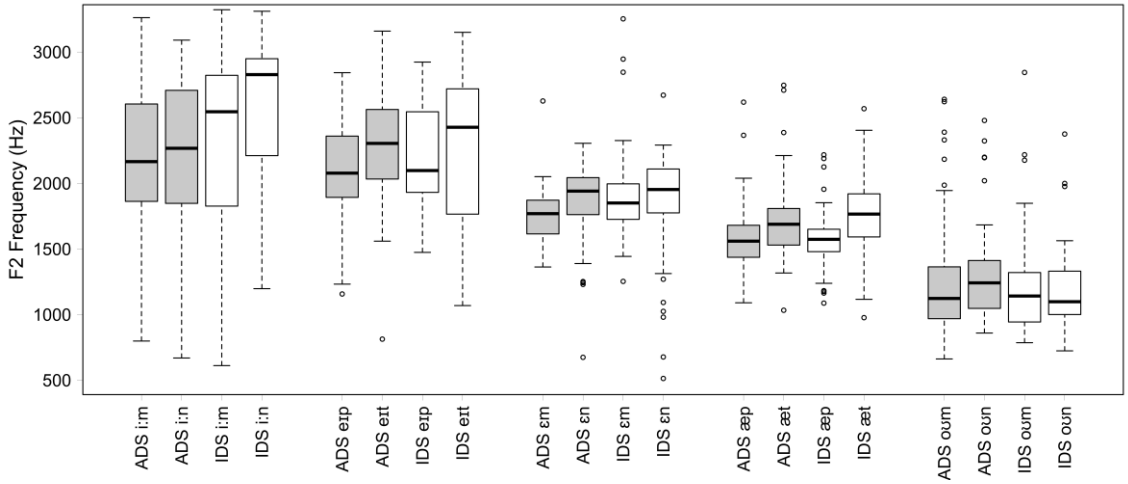


Figure 3

Frequency of second formant at the end of the vowel in the first word of the two-word phrase. Data is grouped by vowel, and divided by Addressee (IDS or ADS) and whether the context licenses assimilation (e.g. *cat box*) or not (e.g. *cap box*). An ANOVA analysis of the data by vowel including the factors Addressee, Licencing Context and their interaction, revealed a significant main effect of Licencing Context ($F(1,4) = 9.75$, $p = .04$) with higher F2 frequency in a context that licenses assimilation. There was no significant effect of Addressee ($F(1,4) = 1.52$, $p = .28$), and no interaction of the two factors ($F(1,4) = 0.01$, $p = .92$).

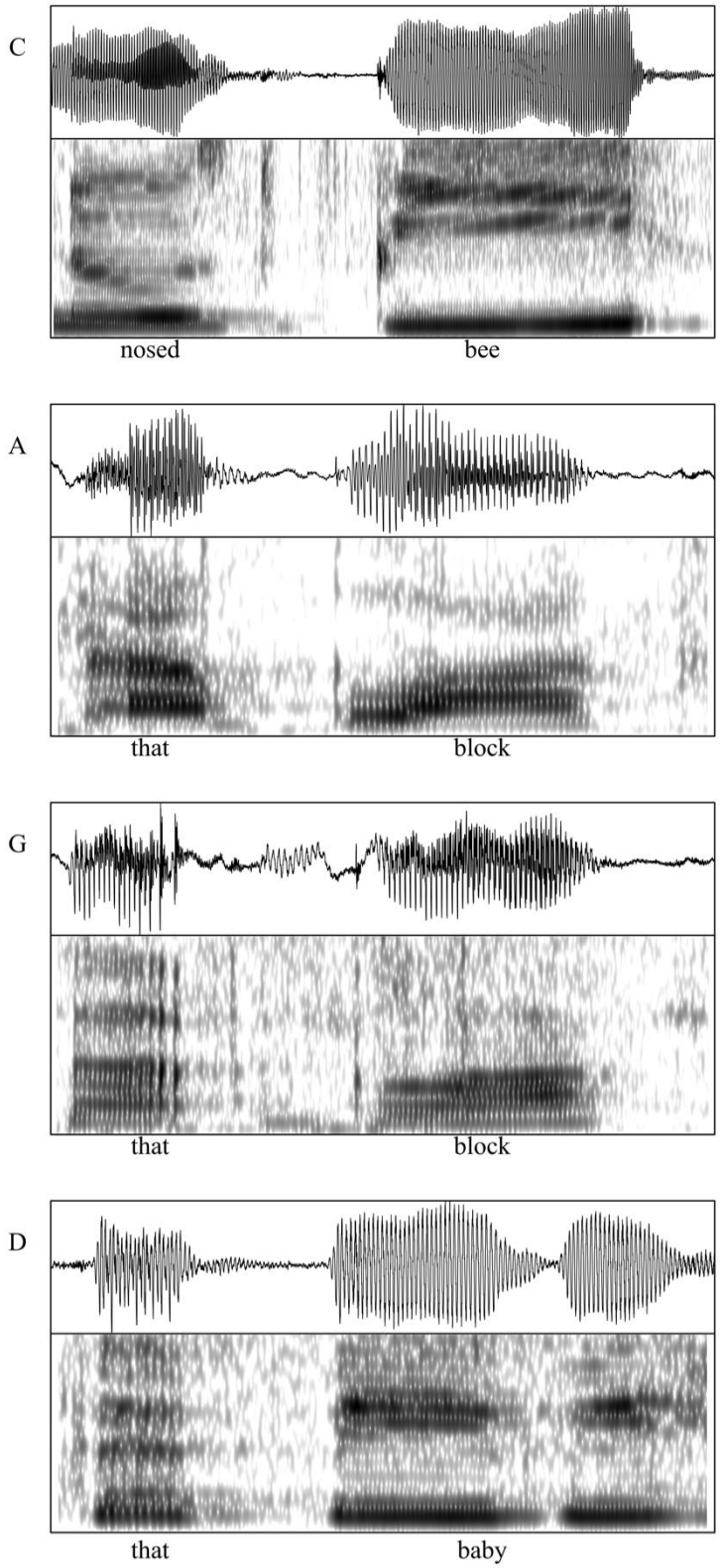


Figure 4

Example waveforms and spectrograms of the four classification types used to analyse the corpus of spontaneous speech. The categories are Canonical (C), Assimilated (A), Glottalized (G), and Deleted (D).

Appendix A

Results from analysis of the frequency (F2) and intensity (A2) of the second formant at the end of the vowel, mean pitch (F0), and duration of the vowel presented in Section 4.

| | F2 | | | A2 | | | F0 | | | Duration | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | *β* | *t* | *p* | *β* | *t* | *p* | *β* | *t* | *p* | *β* | *t* | *p* |
| Intercept | 1860.68 | 10.51 | | 14.76 | 4.23 | | 210.15 | 40.94 | | 110.67 | 17.04 | |
| Addressee | -72.65 | -3.29 | .001 | -0.88 | -1.82 | .07 | -7.65 | -3.13 | .002 | -18.76 | -10.16 | < .001 |
| Scriptedness | 20.19 | 0.91 | .37 | -2.65 | -5.43 | < .001 | 4.16 | 1.69 | .09 | -9.29 | -4.99 | < .001 |
| Assimilation | 90 | 4.09 | < .001 | -1.49 | -3.11 | .002 | 5 | 2.01 | .04 | 4.78 | 2.6 | .009 |
| Add.*Script. | -14.58 | -0.33 | .74 | -1.21 | -1.26 | .21 | -5.04 | -1.04 | .3 | 3.32 | 0.9 | .37 |
| Add.*Assim. | -21.78 | -0.5 | .62 | -0.3 | -0.32 | .75 | 11.64 | 2.4 | .02 | 0.07 | 0.02 | .99 |
| Script.*Assim. | 49.06 | 1.11 | .27 | -0.04 | -0.04 | .97 | -4.88 | -1 | .32 | -7.73 | -2.1 | .04 |
| Add.*Script.*Assim | -54.25 | -0.62 | .54 | -0.31 | -0.16 | .87 | -0.96 | -0.1 | .92 | 6.93 | 0.94 | .35 |

Appendix B

List of items analysed in corpus of spontaneous speech. The number in parentheses indicates the token frequency.

about bears (1)
about being (2)
about Birthday_Bear (1)
about boats (1)
about mice (1)
about mommy(s) (2)
about my (1)
about pigs (1)
about playing (1)
about Potato's (1)
again because (1)
alphabet blocks (1)
alright mommy's (1)
alright put (1)
at both (1)
at me (4)
at Mommy (3)
at pictures (1)
bad bird (1)
baked beans (2)
baked bread (1)
basket ball (1)
bat bat (1)
bat made (1)
bed book (1)
bed platform (1)
been making (1)
been playing (1)
bit big (1)
bit Mommy (1)
bit more (4)
blood pressure (2)
bonked me (1)
bonked Mommy (1)
bought mommy (2)
broken book (1)
brought books (1)
brown bear (1)
brown bird (1)
but Boom_Shaka_Laka_Laka (1)
can bark (1)
can be (4)
can bring (1)
can build (5)
can make (11)
can Mama (1)
can Mommy (5)
can move (3)
can paint (1)
can play (4)
can practically (1)
can press (1)
can pull (3)
can put (15)
cannot push (1)
cat black (1)
cat brown (1)
cat buttons (1)
children painting (1)
clean piece (3)
clean puppy (1)
could be (6)
could make (1)
could play (1)
could put (3)
crayon back (1)
did Birthday_Bear (1)
did miss (1)
did mommy(s) (2)
down by (1)
down please (4)
eat big (1)
eat breakfast (1)
eighteen months (2)
Ethan made (1)
even bigger (1)
even make (1)
even more (1)
even put (1)
fit because (1)
fit better (1)
flowered bathing (1)
folded pajamas (1)
food bag (1)
fooled me (1)
forgot mommy (1)
get big (1)
get bundled (1)
get Max(s) (5)
get me (1)
get miss (1)
get more (3)
get motorcycle (1)
get muscles (1)
get pen (1)
get puppy (2)
goat baby (2)
goat but (1)
God bless (10)
good balancing (1)
good boy (10)
good breakfast (2)
good maybe (1)
good morning (10)
good muffin (1)
good pictures (1)
good place (2)
good_night book (3)
got baby (1)
got more (1)
got peepee (1)
got plenty (1)
got poopy (2)
got pretty (1)
great manners (1)
great pictures (1)
green ball (1)
green marks (1)
green mountain (1)
green pants (1)
green pepper(s) (2)
had put (1)
head broken (1)
hold Mommy's (1)
hot baby (1)
in back (1)
in bed (2)
in between (2)
in black (1)
in blue (1)
in Maine (2)
in Max's (1)
in mine (1)
in Mommy(s) (5)
in moo (1)
in my (6)
in Pikachu's (1)
it back (24)
it be (2)
it belong(s) (8)
it bigger (2)
it black (1)
it bothering (3)
it break (1)
it broke (1)
it by (2)
it makes (2)
it matter (1)
it means (1)
it might (8)
it mom (1)
it must (3)
it plugs (1)
it pop (2)
kitchen messy (1)
let me (43)
let Mommy (4)
lid back (1)
light brown (1)
light bulb (9)
lion book (1)
magnet board (1)
might be (17)
need batteries (1)

need me (1)
need more (1)
night biting (1)
nineteen months (1)
nosed bee (2)
not be (1)
not been (2)
not being (2)
not blue (1)
not break (1)
not broken (1)
not by (2)
not move (1)
not play (1)
not playing (1)
not Pooh (1)
not put (1)
old McDonald (1)
on back (2)
on Ben (1)
on Birthday_Bear(s) (3)
on but (1)
on me (2)
on missy (1)
on mommy(s) (12)
on my (13)
on paper (2)
on Peter's (1)
on please (1)
on Pooh (1)
on purpose (1)
parrot (1)
parrot puppet (2)
peanut butter (2)
person might (1)
picked blueberries (1)
pumpkin pie (1)
put moisturizer (1)
put paint (1)
put Pooh (1)
queen bee (1)
raisin boy (1)
read books (3)
read Maisy (1)
read Max (1)
read moo (1)
read more (1)
read Pooh (1)
red ball (1)
red baseball (1)
red bird (1)
red block (3)
red bow (1)
red mouth (2)
red polka (1)
right back (4)
right behind (1)
right people (1)
right place (1)
right pumpkin (1)
sad baby (1)
said byebye (1)
said Mommy (1)
said more (1)
said please (1)
scared me (1)
seaweed but (1)
seen my (1)
should be (5)
should bring (1)
should mommy (4)
should move (1)
should probably (1)
should put (2)
sign means (1)
sit by (1)
soon because (1)
spilled porridge (1)
spot book (1)
sun bath (1)
sweet peas (1)
sweet potato (1)
that baby(s) (28)
that back (5)
that bag (1)
that ball (1)
that banana (1)
that Barney (1)
that be (1)
that bear (3)
that beautiful (1)
that because (1)
that bed (2)
that bellows (1)
that belong(s) (3)
that better (4)
that big (5)
that bin (1)
that bird (2)
that birdie (1)
that black (1)
that block (6)
that boat (1)
that book (14)
that bothering (1)
that box (2)
that boy (1)
that bracelet (1)
that brown (1)
that bucket (1)
that bumble (1)
that bunny (1)
that bus (1)
that by (1)
that made (1)
that magazine (3)
that Maisy's (1)
that make(s) (3)
that Manuela's (1)
that many (2)
that Max (1)
that may (1)
that mean(s) (5)
that might (2)
that Mommy(s) (4)
that money (1)
that monkey (2)
that mouse (3)
that move (1)
that much (1)
that musketeer (1)
that must (1)
that my (2)
that page (1)
that paper (2)
that part (1)
that person (9)
that picture (20)
that piece (1)
that pig (1)
that pillow (1)
that play (1)
that please (1)
that poor (1)
that present (1)
that pretty (1)
that puppy(s) (4)
that purple (1)
that puzzle (2)
then put (2)
told B (1)
told me (1)
train book (2)
train massage (1)
train might (1)
train passes (1)
vacation but (1)
what belongs (1)
what book (4)
what Maisy (1)
what makes (1)
what Mommy (2)
what music (1)
what my (1)
when Badega (2)
when Max (1)
without me (1)
without Mommy (1)
wood block (1)
would be (12)
would Birthday_Bear (1)
would pull (1)