Perception and bias in the processing of English compound versus phrasal stress:
The effect of synthetic versus natural speech

Irene Vogel\textsuperscript{a}, Arild Hestvik\textsuperscript{b}, Nadya Pincus\textsuperscript{c}

\textsuperscript{a,b,c} Department of Linguistics and Cognitive Science
University of Delaware
125 E. Main Street
Newark, DE 19716
USA

\textsuperscript{a} ivogel@udel.edu
\textsuperscript{b} arild@udel.edu
\textsuperscript{c} npincus@udel.edu

Corresponding author: Irene Vogel
Phone number: (302) 831-6887

KEYWORDS: Speech perception, stress, compound stress, phrasal stress, psycholinguistics
Abstract

The distinction between English compound and phrasal stress has received much attention both theoretically and experimentally. Experimental findings show that listeners attribute a compound meaning to a phrasal stress pattern more often than the other way around. We further investigate the difference by modifying the signal-to-noise ratio in our stimuli to determine the effect of noise on compound vs. phrasal stress interpretation. The responses to different types of natural speech and synthetic speech reveal that increased noise does in fact lead to poorer discriminability and favors compound interpretation as a type of default. An acoustic examination of the stimuli shows how the manifestation of the stress patterns in the different speech types may affect discriminability. Finally, our findings suggest that the compound response preference may be further bolstered by the fact that it involves identification of a single lexical item rather than the two items needed for phrase identification.
Perception and bias in the processing of English compound versus phrasal stress:

The effect of synthetic versus natural speech

1. Introduction

Compound and phrasal stress in English have received much attention from both theoretical and different types of empirical/experimental perspectives since the introduction of the distinction between the Compound and Nuclear Stress Rules in *The Sound Pattern of English* (Chomsky & Halle, 1968). The prosodic difference between these patterns allows us to distinguish otherwise ambiguous Adjective-Noun (A-N) sequences such as *green-house*, where placing stress on the first element yields a compound meaning and placing stress on the second element yields a phrasal meaning.¹

In tasks requiring the selection of a picture corresponding to an auditory stimulus with either compound or phrasal stress, adults are quite successful in disambiguating A-N strings on the basis of their prosody, although when they make errors, they tend to be in the direction of interpreting phrasal stress as compound stress (Vogel & Raimy, 2002; McCauley et al., 2012). Children, on the other hand, consistently interpret phrasal stress as compound stress and do not show adult behavior until about 11 years of age (e.g. Atkinson-King, 1973; Vogel & Raimy, 2002). The fact that the greater accuracy in identifying compound stress extends into adulthood raises the question of whether phrasal stress is simply harder to perceive, resulting in more misinterpretations than compound stress, or whether there is some more basic response bias that leads to a preference for a compound interpretation and thus more errors with phrasal stress. If

¹ Not all compounds in English are stressed on the first syllable (e.g. *Times Square, city planner*) as discussed in Plag (2006) and Plag et al. (2008), among others. The focus here, however, is on those that do exhibit stress on the first element since these are the ones that permit investigation of the role of prosody in the disambiguation of pairs of A-N sequences.
the former is true, increasing “noise” in the stimuli would be expected to result in increased reliance on the compound preference. By contrast, if the latter is true, the overall preference for compound interpretations should be observed independently of “noise”.

In this paper, we examine the relationship between noise in the signal and the bias towards compound stress interpretation. Specifically, we investigate the nature of the preference for compound interpretation of ambiguous A-N sequences by comparing responses in a picture-selection task where the stimuli are presented with four different voices, two natural voices and two synthetic voices. An acoustic analysis of the stimuli furthermore permits us to assess the nature and the role of the signal-to-noise ratio in the interpretation of the ambiguous strings, and ultimately, to assess the source of the observed compound response preference.

1.1. Previous Studies: the compound bias and perception

In previous studies (e.g. Atkinson-King, 1973; Vogel & Raimy, 2002), potentially ambiguous Adjective-Noun (A-N) compounds and phrases (e.g. green-house) were presented to subjects with one of the two stress patterns. For each auditory stimulus, the subjects saw two pictures corresponding to the potential meanings, and were required to select the picture that corresponded to what they believed they heard. This research found that young children (around age 6) overwhelmingly interpret all strings as compounds regardless of stress. Over time, they show developing ability to identify the phrasal pattern, reaching the adult behavior around age 11 (Atkinson-King, 1973; Vogel & Raimy, 2002). While the older children and adults mostly identify phrases as such, they nevertheless continue to show a bias towards the compound interpretation. That is, they still tend to make more errors in which phrasal stress is
misinterpreted as the compound meaning than errors in which compound stress is misinterpreted as the phrasal meaning, a pattern that has also been observed elsewhere (e.g. Farnetani et al., 1988).

Using a different paradigm, McCauley et al. (2012) presented a single picture, and subjects then made a congruency decision based on an auditory labeling of the picture that was either matched or mismatched in terms of phrasal or compound stress pattern. For example, a picture of a greenhouse was labeled with the (auditory) string “green-house”, carrying either the congruent compound stress or the incongruent phrasal stress. Here too, the behavioral responses showed more errors with phrasal stress: subjects more often characterized as congruent those trials where a phrasal stress auditory stimulus was paired with a “compound” picture, than when a compound stress auditory stimulus was paired with a “phrasal” picture. A Signal Detection Theory (SDT) analysis revealed that subjects had good discrimination of the stress patterns but also a significant bias towards compound stress in response selection.

McCauley et al (2012) also measured event-related brain potentials (ERPs) to the incongruities, with the aim of testing whether the bias was reflected in poor auditory perception of the phrasal stress pattern. The ERP findings converged with the SDT findings: both incongruities elicited distinct and statistically significant ERP responses in comparison with congruent trials, even when subjects erroneously reported that the sound and the picture were congruent. In other words, the brain’s perceptual system recognized that the sound and the picture was incongruent, but the subject’s behavioral system incorrectly reported it to be congruent. This led McCauley et al. to conclude that the bias towards interpreting phrasal stress as compound stress must be due to a post-perceptual response selection bias.
In the present study, we asked whether this post-perceptual bias can be modulated by changing the signal-to-noise ratio, or whether the bias is independent of noise and due to other underlying factors, such as the fact that compounds, as single lexical items, are inherently much more frequent than any non-idiomatic A-N phrase. We follow the paradigm of Vogel & Raimy (2002) where subjects must decide which of two pictures corresponds to an auditory stimulus. In addition, we add a between-subjects variable by presenting different groups with either synthetic speech or natural speech, with two different degrees of clarity within each type of voice. If the “compound stress effect” is due to noise, then we would expect that the more “noise” we introduce into the auditory signal, the more likely subjects will be to respond that they heard compound stress in cases in which they, in fact, heard phrasal stress. On the other hand, if there is still a bias when the signal is not degraded and subjects have good discrimination, the bias must be due to other types of factors.

2. Method

2.1 Participants

A total of 124 university undergraduates participated in the study for course credit. Based on questionnaire information, 27 participants were excluded because they (i) were bilingual or non-native speakers of English (5 subjects), (ii) had parents speaking a language other than English (17 subjects), or (iii) had a history of speech disorders (5 subjects). One participant was excluded for failing to follow task instructions. Of the remaining 97 subjects, only 15 were males due to the skew in the population of students taking an introductory linguistics course. The average age was 18.6 (SD 1.2).
2.2 Design and materials

The study employed a 2 x 4 factorial design with two stress patterns (compound and phrasal) as the within-subjects factor and with four different speaker voices (two natural and two synthetic) as the between-subjects factor. The dependent measures were reaction time and accuracy in picking the correct picture as a function of stress.

Sixteen ambiguous strings were selected in which only stress distinguishes between their interpretation as compound nouns or phrases. Four items were used for practice (along with 8 unambiguous trials) and the remaining items were used as targets. A pair of pictures was constructed for each expression (e.g. a belt colored black for the phrase “black belt”, and a black-belt karate practitioner for the compound version; a green painted house for the phrase “green house” or a glass building for growing vegetables for the compound “greenhouse”). In several items, the first word could be considered a noun; however, the picture we used showed it as a modifying property of the following word. For example, the pictures for “paper boy” showed a newspaper carrier for the compound meaning and a string of paper cut-out boys for the phrasal meaning. See Appendix B for the full list of targets.

A recorded sentence of the form “Show me where the A-N is” was matched to each picture of a pair. In the experimental trials, one stress version of the A-N sequence was presented along with a pair of pictures. Thus, the reference of the two-word sequence in the experimental trials would be ambiguous if stress was not taken into consideration (i.e. the sequence “black” and “belt” could match either a black-belt level karate practitioner or simply a black belt). In addition, 24 filler pairs of pictures and recordings were constructed with objects
unambiguously labeled by an auditory stimulus. For example, a pair of pictures showing a jump-rope and a blue rope would be matched with either the compound stress word *jump-rope* or the phrasal stress expression *blue rope*. These provided both a distraction from the targets and a check to see if subjects were paying attention to obvious mismatches.

Two lists of stimuli were assembled with target, filler and practice items. In one list, half of the items were matched to the recorded sentence with compound stress and the other half to the sentence with phrasal stress. In the other list, the opposite pairing was used. Each participant was presented with one list, so that s/he only heard one version of a picture-sound pair; the use of the two lists was counterbalanced across subjects.

In order to compare the effect of different speech qualities with different degrees of signal-to-noise ratio, four different voices were used for the presentation of the stimuli: two with natural speech and two with synthesized speech. The two types of natural voice recordings were (i) that of one of the female authors speaking in a comfortable way without specifically attempting to distinguish between the compound and phrasal interpretations of the sentences and (ii) that of a female voice-over professional who was instructed to make the compound or phrasal meaning clear without exaggerating in an unnatural way. These are referred to below as NAT1 and NAT2. The two synthesized voices, which we refer to as SYN1 and SYN2, were (i) a low-quality female synthetic voice and (ii) a higher-quality male synthetic voice. The synthesized stimuli were generated using the ModelTalker TTS system (Bunnell & Pennington, 2010), a concatenative synthesizer that also controls intonation and timing, thus offering a good comparison with the natural voice stimuli. The voice used for the female voice (“Kate”) was based on a relatively sparse corpus comprising fewer than 150 short sentences, whereas the
synthetic male voice (“Tom”) was based on a 1650 utterance corpus, the size typically used for ModelTalker synthesis.²

The natural voices were considered to involve less “noise” than the synthetic voices. Moreover, the intentionally clarificational NAT2 was considered to involve less “noise” than NAT1. The higher quality synthesized voice, SYN2, was considered to involve less “noise” than SYN1. The voices can therefore be ordered from most to least noise as SYN1 > SYN2 > NAT1 > NAT1.³

2.3 Procedure

Stimulus presentation and data collection were programmed in E-Prime (Schneider et al., 2002). The stimulus presentation was counterbalanced using the two lists discussed above, such that each subject saw each picture pair once, and either heard a sentence with phrasal stress or compound stress corresponding to that picture pair. The two lists were in different fixed pseudo-randomized orders. In addition, whether the compound picture was presented on the left or right side was counterbalanced, as was the presentation of the correct picture on the left or right side.

² The laboratory version of the ModelTalker system (Bunnell & Pennington, 2010) controls both intonation and timing via PSOLA processing. Thus, while F0 and timing effects associated with pitch and phrase accents are synthesized, other factors associated with prominence in natural speech (e.g. amplitude, voice quality, spectral tilt) are represented in the synthetic output only to the extent they are present in the concatenation units the system selected from its speech database (cf. Vogel et al., 2009).

³ We are using the term “noise” in a general sense here to refer to some property of the speech that detracts from its clarity, not the addition of some constant type of “white noise”. Section 3.4 below on the acoustic properties of the voices, which discusses the nature of the “noise” in terms of two main correlates of stress: duration and F0.
Each subject heard the stimuli presented in only one voice. Before beginning the actual experiment, subjects were given 12 practice trials, with accuracy feedback after each trial.

After the practice, subjects were presented with 12 blocks of 3 trials (each block always contained 2 fillers). After each series of 3 trials, a feedback slide was provided with the cumulative accuracy to that point (i.e. a message appeared saying “Your cumulative accuracy is X%”). In this way, subjects were encouraged to attend fully to the task, but they did not receive feedback on specific trials, and could not use the feedback information to infer correct vs. incorrect interpretation for any one experimental trial.

In sum, a single trial appeared to the subject as follows. First, a fixation cross appeared in the center of the computer screen for 1000 ms. This was followed by the spoken sentence (e.g. “Show me where the greenhouse is”), which was presented within a maximum time frame of 2.55 seconds, with 500 ms lead-in and 500 ms pause after the end of the sentence. This was followed by the presentation of the two pictures (e.g. a green-colored house, and a growing structure for plants), one on the left and one on the right of the screen, as illustrated in Figure 1:

(INSERT FIGURE 1 HERE)

Subjects responded by pressing the number “0” to select the right picture and “1” to select the left picture as the correct match to what they heard. If no selection was made, the picture timed out after 3000 ms. The whole experiment lasted about 10 minutes per subject.
3. Results

We first consider the reaction times and then the accuracy in the subjects’ responses. Finally, we test for bias with a Signal Detection Theory analysis. We analyze the effects of the within-subject variable STRESS (2 levels: compound vs. phrasal stress) and the between-subject variable VOICETYPE (four levels: NAT1, NAT2, SYN1, SYN2). Although four voices were used in the experiment, we also present the results in some cases as a combination of the two natural voices and the two synthetic voices simply as a NATURAL voice type versus a SYNTHETIC voice type.

3.1 Reaction time analysis

Figure 2 shows the distribution of mean reaction times and dispersion for correct answers for each voice type with the two stress patterns.

A repeated measures ANOVA on RT for correct responses revealed a main effect of STRESS (F(1,93)=7.23, p=0.008), but no main effect or interactions involving VOICETYPE.\footnote{Three subjects were excluded from this analysis because of missing data in one cell, due to low accuracy.} Using orthogonal contrast analysis, we determined whether there were differences among the natural vs. synthetic voices. No significant difference was found between the compound and phrasal stress patterns as a main effect, between the two synthetic voices (t=-0.17, p=.86), or between the
two natural voices (t=-0.04, p=.92). Furthermore, while the RT difference between the two stress patterns is greater for the natural voices, the interaction between the two broader voice categories was not significant when the difference between compound and phrasal stress was compared (F(1,93)=0.75, p=.38).

Similarly, a repeated measures ANOVA with items as a random factor, and STRESS and VOICETYPE as within-item factors revealed a main effect of STRESS (F(1,11)=5.62, p=0.04), but no interaction with VOICETYPE. Furthermore, orthogonal contrast analysis revealed no difference between the two synthetic voices or between the two natural voices, nor between the two sets of voices.

3.2. Accuracy analysis

The accuracy per voice type is given in Table 1:

<table>
<thead>
<tr>
<th>VOICETYPE</th>
<th>Voice</th>
<th>compound</th>
<th>phrasal</th>
</tr>
</thead>
<tbody>
<tr>
<td>synthetic speech</td>
<td>SYN1</td>
<td>78%</td>
<td>41%</td>
</tr>
<tr>
<td></td>
<td>SYN2</td>
<td>65%</td>
<td>48%</td>
</tr>
<tr>
<td>natural speech</td>
<td>NAT1</td>
<td>80%</td>
<td>69%</td>
</tr>
<tr>
<td></td>
<td>NAT2</td>
<td>76%</td>
<td>69%</td>
</tr>
</tbody>
</table>

Table 1: Accuracy per voice and stress pattern
For analysis of accuracy, we used hierarchical linear regression models (HLM) with the logit link. Logistic regression avoids problems associated with using ANOVAs on proportions, which would violate assumptions about continuity, as well as ceiling/floor effect problems (Jaeger, 2008). HLM also allows for the simultaneous analysis of items as random factors and subjects as random factors; hence, no separate analysis of items is necessary (Clark, 1973).

For the full set responses, setting the intercept at compound stress yielded a mean accuracy of 74% correct; the 95% confidence interval (CI) was 70%-79%. The coefficient for phrasal stress was highly significant (t(96)=−6.531, p<.0001), indicating that accuracy for phrasal stress was significantly lower at 56% (CI: 50%-62%, just shy of chance).

---

5 HLM 6 software (Raudenbush et al., 2005) was used for the statistical analysis. In this type of analysis, the single trial data within each subject is first modeled by a regression equation that estimates the logit—the logarithm of the odds of correct picture selection in a baseline condition (the intercept)—and then computes the regression coefficient for the effect of changing stress from compound to phrasal stress. At the level of a single subject, the equation that models that subject’s data is given by (i).

\[ \text{logit}_i = \log \left( \frac{p}{1-p} \right) = \beta_0 + \beta_1 \times \text{PHRASAL} \]

Specifically, \( \beta_0 \) is the log-odds of correct picture selection in the compound condition, and \( \beta_0 + \beta_1 \) is the log-odds of correct picture selection in the phrasal condition. Because each value is the proportion or probability correct in each condition (based on six trials in each condition), there is no error term at this level. At the level of the set of all subjects, each subject’s intercept and coefficient are modeled by two new regression equations, where \( \gamma_{00} \) is the mean of all intercepts, and \( \gamma_{10} \) is the mean of all slope coefficients. This provides variance estimates of the baseline accuracy as well as of the coefficients for phrasal stress, as in (ii).

\[ (a) \quad \beta_0 = \gamma_{00} + \gamma_{01} \times \text{SYNTHETIC} + u_0 \]
\[ (b) \quad \beta_1 = \gamma_{10} + \gamma_{11} \times \text{SYNTHETIC} + u_1 \]

Between-subject variables can be modeled at this level as coefficients for both the intercept regression models as well as the phrasal slope regression models. The combined model allows for tests of significance of whether voice type (natural vs. synthetic) has a significant effect on the basic accuracy of compound stress or on the phrasal slope coefficients, effectively modeling interactions between i) between-subject variables and ii) within-subject variables, as shown in the combined model in (iii):

\[ (c) \quad \text{logit}_i = \log \left( \frac{p}{1-p} \right) = \gamma_{00} + \gamma_{01} \times \text{SYNTHETIC} + \gamma_{10} \times \text{PHRASAL} + \gamma_{11} \times \text{SYNTHETIC} \times \text{PHRASAL} + u_0 + u_1 \times \text{PHRASAL} \]
To examine main effects and interactions involving VOICETYPE, we first used a binary predictor (natural vs. synthetic) at the level of subjects (collapsing the two natural and synthetic voices into two categories). Using natural voices as the intercept, the accuracy for compound stress was 78% (CI: 72%-83%). The effect of synthetic speech lowered the accuracy for compounds by 6% to 71% (CI: 60%-80%), but this change was not significant (t(95)=-1.54, p=0.13). Accuracy in the phrasal stress condition with natural speech was 69% (CI: 60%-76%); this lowering of accuracy in the phrasal stress condition was significant (t(95)=-2.41, p=0.018). The interaction between STRESS and VOICETYPE was also significant (t(95)=-2.57, p=.012): changing the coefficients from compound to phrasal stress and from natural to synthetic speech lowered accuracy in the phrasal condition by another 24%, bringing it down to 45% (CI: 29%-53%), shown in Figure 3:

This interaction means that a change in coefficients from natural to synthetic speech results in a significant decrease in accuracy for phrasal stress, compared to the difference in accuracy with compound stress.

Separate HLM analyses were also conducted for the individual synthetic and natural voices. Comparison of the two natural voices revealed no significant difference with respect to the intercept (compound stress). The coefficient for phrasal stress was significant (t(45)=-2.287, p=0.027); however, there was no interaction, indicating that there was no difference between NAT1 and NAT2 with regard to phrasal stress. Comparison of the synthetic voices revealed
marginally lower accuracy with SYN2 for compound stress (t(48)=-1.937, p=0.058). The coefficient for phrasal stress was again significant (t(48)=-8.148), p<.001); in addition, there was an interaction between synthetic voice type and the phrasal stress coefficient (t(48)=3.4, p=.002), caused by lower accuracy with SYN1 in the phrasal stress condition.

To summarize, these analyses revealed that phrasal stress resulted in lower accuracy overall, and that this lowering effect was greater with synthetic speech. Considering that synthesized speech introduces “noise”, these findings indicate that increased noise results in lower accuracy in pairing compound and phrasal stress with their corresponding pictures. Moreover, the two synthetic voices exhibited a further interaction related to accuracy and “noise”. Finally, when the two natural and the two synthetic voices were grouped and compared, there was also an interaction, with “errors” increasing dramatically in the phrasal stress condition.

3.3. Signal Detection Theory analysis

To further examine whether the above effects could be due to response bias in favor of compound stress or to perceptual difficulties with the phrasal stress pattern itself, the data were submitted to a Signal Detection Theory analysis. This requires determination of the “signal”, which we (arbitrarily) defined as a trial with phrasal stress. The task can then be conceived of as “Be on the look-out for phrasal stress!” If the subject actually hears phrasal stress and correctly picks the phrasal picture, the trial is coded as a HIT. If the subject hears phrasal stress and picks the compound picture, it is a MISS. If compound stress is presented (the “no-signal condition”), and the subject picks the compound picture, it is a CORRECT REJECTION. If compound stress
is presented and the subject selects the phrasal picture (thinking there was a signal when in fact there was none), it is a FALSE ALARM. This provides the $d'$ score, an indication of the sensitivity associated with each voice type. The corresponding C score indicates the bias observed in each case.

Figure 4 shows the mean $d'$ and C scores for VOICETYPE and STRESS.\(^6\)

(INSERT FIGURE 4 HERE)

As can be seen in Figure 4, sensitivity ($d'$) is high for the two natural voices and relatively low for the synthetic voices. In addition, bias (C) is negative for all voice types, but minimally so for the natural voices. The negative C scores mean that subjects were biased towards identifying the stimuli as compound stress, and this bias increases with more “noise”. In order to assess these patterns, we first compared the voices to each other to see if the differences in $d'$ and C were significant, and then tested whether $d'$ and C were significantly different from 0 for each voice.

Analyzing sensitivity across voices first, a one-way ANOVA with $d'$ as the dependent measure and VOICETYPE as the categorical factor revealed a main effect of voice on discrimination sensitivity, evidently driven by the high sensitivity in the natural voice groups. This effect was further examined with orthogonal contrast analysis, which showed that the two synthetic voices did not differ significantly, nor did the two natural voices; however, the two

\(^6\) Note that if compound stress is taken to be the signal, the $d'$ scores will be the same, and the C scores for bias (see below) will be inverted. The interpretation is the same: there will be a positive bias towards interpreting phrasal stress as compound stress – a false alarm bias.
natural voices as a group resulted in significantly higher sensitivity than the two synthetic voices as a group (t=3.16, p<.001).

We also tested whether each voice’s sensitivity score was significantly greater than 0 to determine whether subjects performed better than if they had simply guessed about which stress pattern they heard. Again, the two natural voices fared better, showing high sensitivity scores. The $d'$ of the lower quality synthetic voice (SYN1) was also significantly different from 0 (t(23)=2.79, p=.01). The $d'$ of the higher quality synthetic voice (SYN2), however, was not different from 0 (t(27)=1.35, p=.19), indicating that subjects were not able to perceive the stress difference, despite our assumption that this voice was of better “quality”. In sum, with respect to perception or discriminability, the synthetic voices made it more difficult to distinguish between compound and phrasal stress. This converges with the results from the HLM analysis of raw accuracy.

With respect to bias, we first compared C across all groups. A main effect of voice was observed (F(3,93)=6.5, p<.001), driven by the strong negative bias with SYN1. Orthogonal contrast analysis showed that the bias observed with SYN1 was significantly greater than the bias seen with SYN2 (t=3.37, p=.002). Separate t-tests against 0 for each voice showed that two natural voices differed marginally (t=1.79, p=.076), and the two natural voices together differed from the two synthetic voices (t=2.92, p<.01). Bias for NAT1 (t(23)=-1.69, p=.10) and NAT2 (t(24)=-0.57, p=.57) did not differ significantly from 0, indicating an absence of bias for the natural voices. However, SYN2 (t(23)=-7.7, p<.001) and SYN1 (t(27)=-2.11, p=.043) did differ significantly from 0, indicating the presence of a bias for the synthetic voices.
3.4. Acoustic properties of the stimuli

As just seen, a number of results show effects due to the voice the subjects heard; therefore, we conducted a basic acoustic analysis of the stimuli to assess the differences among the voices. We did not systematically vary individual acoustic properties of the stimuli as part of the stimulus preparation; however, the four voices we used could globally be ranked as less or more clear, ranging from the poor quality synthesized voice (SYN1) to the intentionally clarificational voice (NAT2). This, in turn, can be associated with more or less “noise”. As mentioned above, we have been using the term “noise” in a broad sense here to refer to some aspect of the stimuli that results in poorer performance. While it is subjectively easy to recognize such “noise” in the synthetic voices, it is not as easy to recognize what might constitute noise in the natural voices. We thus examine below two properties strongly correlated with stress in English, duration and fundamental frequency (F0), which corresponds to pitch. Given that all of our stimulus words have some degree of stress (compound or phrasal), other properties more related to vowel quality and vowel reduction (cf. Sluijter et al., 1997) were not examined.


Since the present experimental design closely follows that of Vogel & Raimy (2002) and uses many of the same materials, we first briefly consider the acoustic properties reported for the stimuli in that study. Given that Vogel & Raimy (2002) observed 92% accuracy for compound 

7 Systematically manipulating the stimuli would have resulted in all of the voices being “unnatural” to some extent, and moreover, it would have involved a different type of experiment altogether, where the effects of individual acoustic features are examined rather than the effects of the more comprehensive voice types investigated here.
stress and 77% for phrasal stress (for known words), we may consider the lower accuracy with both our natural voices (compounds 78%, phrases 69%) and our synthetic voices (compounds 71%; phrases 45%) to be a result of noise in comparison with the earlier study.

As can be seen in Figure 5, in Vogel & Raimy’s study, the duration of both words of the compounds is essentially the same. The duration of the first word (W1) of phrases is also essentially the same as these words. Thus, as far as duration is concerned, the only distinguishing property between compounds and phrases is the substantially longer second word (W2) in phrases, precisely what is expected due to phrase-final lengthening.

(INSERT FIGURE 5 HERE)

Figure 6 shows the pitch patterns of compounds and phrases; a positive value indicates that F0 has risen from the beginning to the end of a word while a negative value indicates that it has decreased.\(^8\) As can be seen, there is a rise only on the first word of compounds. Although there is a greater drop in pitch on both words of phrases than on W2 of compounds, the rise on the first word of compounds appears to be the main pitch cue to the distinction between compound and phrasal stress, since it is the opposite pattern from that observed in phrases.

(INSERT FIGURE 6 HERE)

Taken together, these duration and F0 patterns may be considered to involve the least noise of all of the stimulus voices, given that they led to the greatest response accuracy. The

\(^8\) The sonorant part of the rhyme of the target words was used for the calculations of F0. In disyllabic words, the stressed syllable was used for measurement.
properties of our natural and synthetic voices can be compared to the patterns in Vogel & Raimy in assessing the manifestation of noise in the different voices.

3.4.2. Acoustic patterns of the two natural and two synthetic voices

Figure 7 shows the duration patterns of the natural and synthetic voices, respectively. As can be seen, all four voices show the same pattern, a longer duration of the second word with respect to the first word in both the compound and phrasal stimuli. This pattern is different from that observed by Vogel & Raimy in several respects. First, it fails to provide a clear difference between compounds and phrases, and specifically, it does not show the substantial rise on W1 of compounds that we suggested may have been the primary duration cue in the earlier study. More subtly, it can be noticed that the duration difference between the two words of compounds in SYN1 is essentially the same as the duration difference between the two words of phrases in Vogel & Raimy. Given the higher accuracy in Vogel & Raimy, we may consider that these different duration patterns constitute noise in both the natural and synthetic voices.

(INSERT FIGURE 7 HERE)

---

9 While the actual measurements of the two studies cannot be directly compared, it is nevertheless valuable to compare the studies since they used essentially the same experimental design and most of the same stimuli. Moreover, the general range of durations in the two studies is very similar.
Turning now to pitch (F0), the patterns for the natural and synthetic voices can be seen in Figure 8. A noticeable difference from Vogel & Raimy (2002) is the reduced pitch range for both voice types, with the synthetic voice ranges being quite minimal; in fact, a reduced range in the graph was required to make the patterns visible. Thus, the flatter prosody in the present study may be seen as constituting some degree of noise with respect to speech that makes use of a wider prosodic range.

Looking at the specific pitch patterns, it can be seen again that while Vogel & Raimy have clearly distinct patterns for compounds and phrases, a rise on W1 of compounds, and a fall elsewhere, both natural voices fail to show distinct patterns for compounds and phrases, another property that could be considered noise. While the patterns used by the two natural voices are different from each other, this seems not to be as important as the fact that each of these voices used the same pitch patterns for compounds and phrases, given that NAT1 and NAT2 showed essentially the same response accuracy. SYN2 similarly showed use of the same pattern for compounds and phrases, though this is difficult to detect in the very reduced pitch range. Interestingly, however, the pattern observed in what is considered the lowest quality voice, SYN2, is the same as that observed in Vogel & Raimy for the voice with the highest accuracy.

(INSERT FIGURE 8 HERE)

In sum, the examination of the duration and pitch patterns of the natural and synthetic voices and the comparison of these voices to those used in Vogel & Raimy suggest several acoustic differences among the voices that might be considered noise. This is by no means a
thorough acoustic analysis; however, it does begin to provide insight into the properties of the stimuli that may lead to more or less accuracy in distinguishing between compound and phrasal stress.

4. Discussion

4.1 Reaction time data and accuracy

The analysis of the reaction time data shows that subjects were significantly slower at correctly identifying a picture when they heard phrasal stress as opposed to compound stress on the A-N string in the carrier sentence “Show me where the A-N is.” Given that the reaction times of the responses were statistically independent of voice quality, identifying phrasal stress simply seems to require more processing time. This finding suggests that accessing a single item, the compound here, takes less time than accessing the two items of a phrase, and this access time does not interact with degree of noise in the voices. Frequency may also play a role since the frequency of the fixed two word sequence of a compound will be much greater than the frequency of the same two words found in the same sequence in phrasal constructions.

By contrast, the analysis of accuracy not only revealed a main effect of the stress pattern, it also revealed a main effect of voice type. That is, in addition to the general reduction in accuracy with phrasal stress, the accuracy was further significantly reduced with the synthetic voices, suggesting that noise played a negative role in the ability of subjects to correctly identify phrasal stress. The fact that the accuracy observed for compounds in the two natural and the two synthetic voices was almost the same, by contrast, shows that the same properties did not play a similar negative role in the compound responses. It would appear, however, that all four voices
included noise in comparison with Vogel & Raimy, where the accuracy with compounds (and phrases) was substantially higher.

4.2 Signal detection: discriminability and bias

The finding that phrasal stress was more poorly identified with the synthetic speech still begs the question of whether the synthetic and natural voices differed in terms of discriminability and/or bias. The Signal Detection Theory analysis offers insight into the relationship between discriminability and bias. Specifically, we observed that when the signal was embedded in more noise (i.e. with synthetic voices vs. natural voices), discriminability was poorer, and bias increased. In other words, the harder it is to perceive the stress patterns, the more the subjects rely on a pre-existent bias in responding. The natural voices exhibited better discrimination than the synthetic voices in terms of the $d'$ measurement. Moreover, the C score indicated that there was very little bias, as defined by SDT, observed with the natural voices.

This leads to the expectation that there should be a significant correlation between $d'$ and C scores for the synthetic voices but not for the natural voices. Specifically, it is expected that there will be a systematic relationship whereby the lowering of $d'$ leads to an increase in bias observed in the synthetic but not the natural voices. A correlation analysis confirmed this. The correlation between $d'$ and C in the response patterns with both natural voices was not significant ($r=.05$), but the correlation between $d'$ and C in the responses to both synthetic voices was highly significant ($r = -.4$, p < 0.05).
Interestingly, SYN1 showed better discriminability than the higher quality synthetic voice, SYN2, in terms of the $d'$ measurement; it also showed a greater presence of bias as indicated with the C score. An account of this is offered below in terms of the acoustic analysis.

### 4.3 Acoustic properties and “noise”

The picture that emerges thus far is that there are consistently slower response times and lower accuracy scores associated with phrasal stress as opposed to compound stress in both the synthetic and natural voices, as well as in Vogel & Raimy (only for accuracy since RT was not measured). The response times were essentially the same for the synthetic and natural voices, indicating that the time required to make a decision was not affected by the noise introduced by certain acoustic properties of these voices. Thus, the poorer performance with phrasal stress must be attributed to some inherent challenge posed by this pattern. One possibility, as suggested above, is the fact that a compound interpretation requires recognition of a single lexical item, while a phrasal interpretation requires recognition of two items. An additional challenge may be created by the relatively greater frequency of compound sequences compared to phrasal sequences.

In other respects, the synthetic voices showed evidence of noise in comparison with the natural voices, and both showed evidence of noise in comparison with Vogel & Raimy. Several aspects of the acoustic properties of the different voices provide insight into the nature of this noise. As noted above, all of the voices in the present study present a reduced pitch range with respect to Vogel & Raimy. This may be due to the fact that the stimuli for the latter were recorded for use with children, and thus may have exhibited somewhat exaggerated prosody
typical of child directed speech, even though the recordings were not made specifically in that speech style. Given that the pitch range in the synthetic voices is by far the most reduced, we can establish a noise ranking based on pitch range (from most to least “noise”): SYN1, 2 > NAT1, 2 > Vogel & Raimy. This ranking coincides exactly with the overall accuracy observed with these three groups.10

The acoustic measurements also provide insight into two additional patterns that would not have been expected based on our original prediction that NAT2 would yield better results than NAT1, and that SYN2 would yield better results than SYN1. Instead, the two natural voices resulted in essentially the same responses. Both the duration and pitch patterns may be responsible for this. For both measures, Vogel & Raimy had distinct patterns corresponding to the two stress patterns, while NAT1 and NAT2 each used the same patterns for compound and phrasal stress. Thus, it seems that the two natural voices introduced the same degree of noise (i.e. were essentially identical) when compared with Vogel & Raimy, despite any other differences that might exist between them.

With regard to the synthetic voices, although both showed poor accuracy rates, SYN1 had a higher $d’$ score than SYN2, indicating that it had a greater degree of discriminability. This is at first glance surprising, but in fact, it appears to be a direct effect of the type of synthesis

10 Another difference in the studies that may have affected the overall accuracy rates which is not directly linguistic is the fact that Vogel & Raimy, having less technology available, tested the participants with actual pictures, and required that they physically mark their responses by placing a post-it on the correct picture. This involvement may have led subjects to attend more closely to the task. By contrast, responding with a computer key press can be expected to introduce at least a minimal amount of error due to simply pressing the wrong key on occasion. Moreover, the less interactive experience may have led to somewhat less engagement in the task. Thus, if noise is taken to be anything that reduces accuracy, the presentation method itself may be seen as adding some degree of noise to all four voices in the present study, in comparison to that in Vogel & Raimy. As the same method was used for all the voices in our study, the only place such methodological noise might be relevant is between the overall accuracy associated with the two studies, not among the voices in the present study.
used to create the voices. Recall that for both voices the basic prosody is encoded in the synthesis process. This may, however, be modified by the prosodic properties of the segments of the corpus that are concatenated in forming the words. Since SYN2 draws its segments from a much larger corpus, the basic prosody will be subject to greater modification. Contrary to the expectation that an increased choice of segments for use in a concatenative synthesis process will result in a “better” voice, in this case, the modifications appear to have introduced a disadvantageous distortion of the more basic, albeit stereotyped, prosody. Thus, while the segmental properties of SYN2 might be better than those of SYN1, the prosody, at least as far as compound and phrasal stress are concerned, is less optimal, and as such may be seen as introducing additional noise into SYN2.

4.4. Conclusion

The widely made observation that compound stress appears to be more accurately identified than phrasal stress raises questions as to the source of this differential behavior. Voices with different degrees of “noise” were compared to assess the effect of degradation of the signal on the response patterns. Indeed, more noise led to less overall accuracy in all responses, but it more strongly affected the phrasal stress responses than the compound stress responses. By contrast, while response time also showed better (faster) performance with compound stress than with phrasal stress, noise in the signal did not affect the results.

Using Signal Detection Theory, we found that noise in the signal significantly reduced discriminability of the compound and phrasal stress patterns, and this in turn resulted in a bias favoring the selection of the compound interpretation. Since the challenge of a noisy signal led
to a compound interpretation preference, and since response time was not affected by noise, but still favored (faster) compound responses, this indicates that there must be some basic cognitive strategies underlying the compound response strategy.

The existence of a non-perceptual bias is a likely possibility for two reasons. First, listeners could be biased towards interpreting an A-N sequence as a single lexical item, as this would provide a faster way to analyze a sentence (i.e. lexical access of a single word being faster than lexical access and phrase building of two words). That is, if listeners first try to interpret an A-N sequence as a single word and find a corresponding item in their lexicon, they would then be biased to interpret the string as a compound, even if they had no perceptual difficulties with phrasal stress. Indeed, the reaction time data show that A-N strings with compound stress are more quickly identified than those with phrasal stress. This conclusion is furthermore supported by the findings in McCauley et al. (2012), where significant ERP responses showed perceptual discrimination of sound-picture incongruities both for phrasal stress and compound stress, while the behavioral data in that study nevertheless showed the compound stress bias.

Secondly, since compound words are more frequent in a language than the sequence of the same two words in phrases, this could also bias subjects towards choosing the more frequent, compound, interpretation of the string. The SDT analysis revealed that the synthetic voices resulted in low sensitivity and a corresponding strong bias towards the compound stress. By contrast, the natural voices exhibited high sensitivity and no significant bias was found in the SDT analysis. Nevertheless, subjects in the natural voice groups still made more errors with phrasal stress than with compound stress. Since some part of the response preference for the compound interpretation is independent of noise and bias as determined by the SDT analysis, the
relative frequency of the compounds themselves, as opposed to the phrases, may be seen as contributing to the observed preference.
Figure captions

Figure 1: Screen displays for single trials. Subject were presented with either “Show me where the greenhouse is” with compound stress, or “Show me where the green house is” with phrasal stress, and required to select the correct picture with a button press.

Figure 2: Mean reaction time per voice type and stress type, correct judgments only. Error bars represent +/- 1 standard error of the mean.

Figure 3: Interaction between VOICETYPE and accuracy of STRESS identification.

Figure 4: Mean C scores (bias) and $d'$ scores (sensitivity) per VOICETYPE. High numbers for $d'$ indicate that subjects were good at discriminating compound vs. phrasal stress; low numbers for C indicate a strong bias away from selecting a phrasal stress response.

Figure 5. Duration Patterns: Compounds vs. Phrases (Vogel & Raimy, 2002).

Figure 6. Pitch Patterns: Change in F0 for Compounds vs. Phrases (Vogel & Raimy, 2002).

Figure 7. Left panel: Duration Patterns: Compounds vs. Phrases (Natural Voices). Right panel: Duration Patterns: Compounds vs. Phrases (Synthetic Voices).

Figure 8. Pitch Patterns. Left panel: Compounds vs. Phrases (Natural Voices). Right panel: Compounds vs. Phrases (Synthetic Voices).
Acknowledgements

We would like to thank Dr. Tim Bunnell (Director, Speech Research Lab at A.I. DuPont Hospital for Children in Wilmington, DE) for extensive assistance with the use of ModelTalker for this project, as well as for help with the acoustic analysis.
References


Figure

The graph illustrates the percent correct scores for different voice types: synthetic speech and natural speech. The data points show that natural speech achieves higher percent correct scores compared to synthetic speech for both compound stress and phrasal stress conditions.

- **Compound Stress**: 
  - Synthetic speech: 45%
  - Natural speech: 69%
  - 78% maximum score

- **Phrasal Stress**: 
  - Synthetic speech: 71%
  - Natural speech: 78%
  - 78% maximum score
Supplementary materials

A: Practice stimuli

Table A1: Practice sentences

<table>
<thead>
<tr>
<th>StimCode</th>
<th>Compound stress sentence</th>
<th>Phrasal stress sentence</th>
<th>Compound picture</th>
<th>Phrasal picture</th>
</tr>
</thead>
<tbody>
<tr>
<td>03</td>
<td>Show me where the blacktop is.</td>
<td>Show me where the black top is.</td>
<td>blacktop</td>
<td>black top</td>
</tr>
<tr>
<td>08</td>
<td>Show me where the ladybug is.</td>
<td>Show me where the lady bug is.</td>
<td>ladybug</td>
<td>lady bug</td>
</tr>
<tr>
<td>13</td>
<td>Show me where the toystore is.</td>
<td>Show me where the toy store is.</td>
<td>toystore</td>
<td>toy store</td>
</tr>
<tr>
<td>14</td>
<td>Show me where the whitehouse is.</td>
<td>Show me where the white house is.</td>
<td>whitehouse</td>
<td>white house</td>
</tr>
<tr>
<td>PF1</td>
<td>Show me where the horseshoe is.</td>
<td>Show me where the brown shoe is.</td>
<td>horseshoe</td>
<td>brown shoe</td>
</tr>
<tr>
<td>PF2</td>
<td>Show me where the hairbrush is.</td>
<td>Show me where the curly hair is.</td>
<td>hairbrush</td>
<td>curly hair</td>
</tr>
<tr>
<td>PF3</td>
<td>Show me where the rainbow is.</td>
<td>Show me where the heavy rain is.</td>
<td>rainbow</td>
<td>heavy rain</td>
</tr>
<tr>
<td>PF4</td>
<td>Show me where the candycane is.</td>
<td>Show me where the long candy is.</td>
<td>candycane</td>
<td>long candy</td>
</tr>
<tr>
<td>PF5</td>
<td>Show me where the cellphone is.</td>
<td>Show me where the old phone is.</td>
<td>cellphone</td>
<td>old phone</td>
</tr>
<tr>
<td>PF6</td>
<td>Show me where the teddybear is.</td>
<td>Show me where the sleeping bear is.</td>
<td>teddybear</td>
<td>sleeping bear</td>
</tr>
<tr>
<td>PF7</td>
<td>Show me where the sunset is.</td>
<td>Show me where the smiling sun is.</td>
<td>sunset</td>
<td>smiling sun</td>
</tr>
<tr>
<td>PF8</td>
<td>Show me where the pencilcase is.</td>
<td>Show me where the yellow pencil is.</td>
<td>pencilcase</td>
<td>yellow pencil</td>
</tr>
</tbody>
</table>
Figure A1: Matching practice pictures

03-C-blacktop
03-P-black top
08-C-ladybug
08-P-lady bug

13-C-toystore
13-P-toy store
14-C-whitehouse
14-P-white house

PF1-C-horseshoe
PF1-P-brown shoe
PF2-C-hairbrush
PF2-P-curly hair

PF3-C-rainbow
PF3-P-heavy rain
PF4-C-candycane
PF4-P-long candy

PF5-C-cellphone
PF5-P-old phone
PF6-C-teddybear
PF6-P-sleeping bear

PF7-C-sunset
PF7-P-smiling sun
PF8-C-pencilcase
PF8-P-yellow pencil
B: Experimental stimuli

Table B1: Experimental sentences

<table>
<thead>
<tr>
<th>StimCode</th>
<th>Compound stress sentence</th>
<th>Phrasal stress sentence</th>
<th>Compound picture</th>
<th>Phrasal picture</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Show me where the blackbelt is.</td>
<td>Show me where the black belt is.</td>
<td>blackbelt</td>
<td>black belt</td>
</tr>
<tr>
<td>2</td>
<td>Show me where the blackboard is.</td>
<td>Show me where the black board is.</td>
<td>blackboard</td>
<td>black board</td>
</tr>
<tr>
<td>4</td>
<td>Show me where the bluejay is.</td>
<td>Show me where the blue jay is.</td>
<td>bluejay</td>
<td>blue jay</td>
</tr>
<tr>
<td>5</td>
<td>Show me where the greenhouse is.</td>
<td>Show me where the green house is.</td>
<td>greenhouse</td>
<td>green house</td>
</tr>
<tr>
<td>6</td>
<td>Show me where the highchair is.</td>
<td>Show me where the high chair is.</td>
<td>highchair</td>
<td>high chair</td>
</tr>
<tr>
<td>7</td>
<td>Show me where the hotdog is.</td>
<td>Show me where the hot dog is.</td>
<td>hotdog</td>
<td>hot dog</td>
</tr>
<tr>
<td>9</td>
<td>Show me where the orangetree is.</td>
<td>Show me where the orange tree is.</td>
<td>orangetree</td>
<td>orange tree</td>
</tr>
<tr>
<td>10</td>
<td>Show me where the paperboys are.</td>
<td>Show me where the paper boys are.</td>
<td>paperboys</td>
<td>paper boys</td>
</tr>
<tr>
<td>11</td>
<td>Show me where the redhead is.</td>
<td>Show me where the red head is.</td>
<td>redhead</td>
<td>red head</td>
</tr>
<tr>
<td>12</td>
<td>Show me where the softball is.</td>
<td>Show me where the soft ball is.</td>
<td>softball</td>
<td>soft ball</td>
</tr>
<tr>
<td>15</td>
<td>Show me where the yellowjacket is.</td>
<td>Show me where the yellow jacket is.</td>
<td>yellowjacket</td>
<td>yellow jacket</td>
</tr>
<tr>
<td>16</td>
<td>Show me where the bigbirds are.</td>
<td>Show me where the big birds are.</td>
<td>bigbirds</td>
<td>big birds</td>
</tr>
</tbody>
</table>
Figure B1: Matching experimental pictures

01-C-blackbelt
01-P-black belt
02-C-blackboard
02-P-black board

04-C-bluejay
04-P-blue jay
05-C-greenhouse
05-P-green house

06-C-highchair
06-P-high chair
07-C-hotdog
07-P-hot dog

09-C-orangetree
09-P-orange tree
10-C-paperboys
10-P-paper boys

11-C-redhead
11-P-red head
12-C-softball
12-P-soft ball

15-C-yellowjacket
15-P-yellow jacket
16-C-bigbirds
16-P-big birds
C: Filler stimuli

Table C1: Filler sentences

<table>
<thead>
<tr>
<th>StimCode</th>
<th>Compound stress sentence</th>
<th>Phrasal stress sentence</th>
<th>Compound picture</th>
<th>Phrasal picture</th>
</tr>
</thead>
<tbody>
<tr>
<td>F1</td>
<td>Show me where the bathtub is.</td>
<td>Show me where the full tub is.</td>
<td>bathtub</td>
<td>full tub</td>
</tr>
<tr>
<td>F2</td>
<td>Show me where the classroom is.</td>
<td>Show me where the empty room is.</td>
<td>classroom</td>
<td>empty room</td>
</tr>
<tr>
<td>F3</td>
<td>Show me where the fishbowl is.</td>
<td>Show me where the big fish is.</td>
<td>fishbowl</td>
<td>big fish</td>
</tr>
<tr>
<td>F4</td>
<td>Show me where the birdhouse is.</td>
<td>Show me where the purple bird is.</td>
<td>birdhouse</td>
<td>purple bird</td>
</tr>
<tr>
<td>F5</td>
<td>Show me where the jumprope is.</td>
<td>Show me where the blue rope is.</td>
<td>jumprope</td>
<td>blue rope</td>
</tr>
<tr>
<td>F6</td>
<td>Show me where the bookcase is.</td>
<td>Show me where the thick book is.</td>
<td>bookcase</td>
<td>thick book</td>
</tr>
<tr>
<td>F7</td>
<td>Show me where the carseat is.</td>
<td>Show me where the old car is.</td>
<td>carseat</td>
<td>old car</td>
</tr>
<tr>
<td>F8</td>
<td>Show me where the cupcake is.</td>
<td>Show me where the tall cake is.</td>
<td>cupcake</td>
<td>tall cake</td>
</tr>
<tr>
<td>F9</td>
<td>Show me where the toothbrush is.</td>
<td>Show me where the dark tooth is.</td>
<td>toothbrush</td>
<td>dark tooth</td>
</tr>
<tr>
<td>F10</td>
<td>Show me where the snowflake is.</td>
<td>Show me where the deep snow is.</td>
<td>snowflake</td>
<td>deep snow</td>
</tr>
<tr>
<td>F11</td>
<td>Show me where the doorknob is.</td>
<td>Show me where the round door is.</td>
<td>doorknob</td>
<td>round door</td>
</tr>
<tr>
<td>F12</td>
<td>Show me where the teapot is.</td>
<td>Show me where the shiny pot is.</td>
<td>teapot</td>
<td>shiny pot</td>
</tr>
<tr>
<td>F13</td>
<td>Show me where the footprint is.</td>
<td>Show me where the wet foot is.</td>
<td>footprint</td>
<td>wet foot</td>
</tr>
<tr>
<td>F14</td>
<td>Show me where the raincoat is.</td>
<td>Show me where the green coat is.</td>
<td>raincoat</td>
<td>green coat</td>
</tr>
<tr>
<td>F15</td>
<td>Show me where the sailboat is.</td>
<td>Show me where the large boat is.</td>
<td>sailboat</td>
<td>large boat</td>
</tr>
<tr>
<td>F16</td>
<td>Show me where the schoolbus is.</td>
<td>Show me where the green bus is.</td>
<td>schoolbus</td>
<td>green bus</td>
</tr>
<tr>
<td>F17</td>
<td>Show me where the stopsign is.</td>
<td>Show me where the striped sign is.</td>
<td>stopsign</td>
<td>striped sign</td>
</tr>
<tr>
<td>F18</td>
<td>Show me where the tablecloth is.</td>
<td>Show me where the square table is.</td>
<td>tablecloth</td>
<td>square table</td>
</tr>
<tr>
<td>F19</td>
<td>Show me where the catfood is.</td>
<td>Show me where the fat cat is.</td>
<td>catfood</td>
<td>fat cat</td>
</tr>
<tr>
<td>F20</td>
<td>Show me where the mousehole is.</td>
<td>Show me where the pink mouse is.</td>
<td>mousehole</td>
<td>pink mouse</td>
</tr>
<tr>
<td>F21</td>
<td>Show me where the snowman is.</td>
<td>Show me where the tall man is.</td>
<td>snowman</td>
<td>tall man</td>
</tr>
<tr>
<td>F22</td>
<td>Show me where the lampshade is.</td>
<td>Show me where the thin lamp is.</td>
<td>lampshade</td>
<td>thin lamp</td>
</tr>
<tr>
<td>F23</td>
<td>Show me where the bedroom is.</td>
<td>Show me where the big bed is.</td>
<td>bedroom</td>
<td>big bed</td>
</tr>
<tr>
<td>F24</td>
<td>Show me where the flowerpot is.</td>
<td>Show me where the red flower is.</td>
<td>flowerpot</td>
<td>red flower</td>
</tr>
</tbody>
</table>
Figure C1: Matching filler pictures

F1-C-bathtub
F1-P-full tub
F2-C-classroom
F2-P-empty room
F3-C-fishbowl
F3-P-big fish
F4-C-birdhouse
F4-P-purple bird
F5-C-jumprope
F5-P-blue rope
F6-C-bookcase
F6-P-thick book
F7-C-carseat
F7-P-old car
F8-C-cupcake
F8-P-tall cake
F9-C-toothbrush
F9-P-dark tooth
F10-C-snowflake
F10-P-deep snow
F11-C-doorknob
F11-P-round door
F12-C-teapot
F12-P-shiny pot