efpsa journal of european psychology students

Poppels, T. and Swingley, D. (2014). Semi-supervised Phonetic Category Learning: Does Word-level Information Enhance the Efficacy of Distributional Learning? *Journal of European Psychology Students,* 5(3), 36-45, DOI: http://dx.doi.org/10.5334/jeps.ce

RESEARCH ARTICLE

Semi-supervised Phonetic Category Learning: Does Word-level Information Enhance the Efficacy of Distributional Learning?

Till Poppels* and Daniel Swingley†

To test whether word-level information facilitates the learning of phonetic categories, 40 adult native English speakers were exposed to a bimodal distribution of vowels embedded in non-words. Half of the subjects received phonetic categories aligned with lexical categories, while the other half received no such cue. It was hypothesized that the subjects exposed to lexically-informative training stimuli that were aligned with the target categories would outperform the control subjects on a perceptual categorization task after training. While the results revealed no such group differences, the data indicated that many subjects used the relevant dimension for categorization before having received any training. Implications regarding experimental design and suggestions for future research based on the results are discussed.

Keywords: phonetic category acquisition; distributional learning; semi-supervised learning

One of the first challenges infants face is learning about the speech sounds of the language that surrounds them. The speech sounds that are used contrastively to signal differences in meaning varies cross-linguistically. German, for example, distinguishes between /y/ and /i/ to mark semantic differences, as in 'South' / z y t / and 'see' / z i t /, while English has no such contrast. Likewise, Spanish does not use / i / and / ɪ / contrastively, while English requires a contrast between the two to distinguish words like 'beet' /bit/ and 'bit' /bɪt/.

There is ample evidence that infants know which speech sound contrasts are present in their native language by their first birthday. In a pioneering study, Werker and Tees (1984) compared adult speakers of Nthlakampx (a native language spoken in British Columbia) and adult English speakers on their ability to discriminate two speech sounds that were contrastive in Nthlakampx, but not in English. English adults performed significantly worse than Nthlakampx speakers. Notably, six-month-old infants who were learning English and had never been exposed to Nthlakampx were as good at discriminating the speech sounds as native Nthlakampx speakers. To determine the time course of the decline in this discriminatory ability, Kuhl, Williams, Lacerda, Stevens, and Lindblom (1992) tested six-month-old infants who were growing up in either English or Swedish-speaking environments on their ability to discriminate speech sounds that were contrastive in one language but not the other. They found that six-month-olds readily detected speech sound contrasts in their native language, but had lost the ability to discriminate between speech sounds that were not contrastive in their native language. It is now widely accepted that infants begin to learn to distinguish contrastive from non-contrastive differences between speech sounds before their first birthday (Bosch & Sebastián-Gallés, 2003; Polka & Werker, 1994), but exactly how that learning occurs remains an open question (Goudbeek, Cutler, & Smits, 2008; Goudbeek, Swingley, & Smits, 2009).

One possibility is that infants infer which speech sounds are contrastive through minimal pair analysis (Best, 1995; Lalonde & Werker, 1995; Werker & Pegg, 1992). Minimal pairs are word pairs that are identical in all but one speech sound. For example, the English words 'heat' /hit/ and 'hit' /hɪt/ form a minimal pair in English because they differ phonetically only in one segment - / i / vs. / ɪ / - which signals the semantic difference between the words. As minimal pairs isolate and highlight phonetic contrasts, it is possible, in principle, to infer the phonetic categories of a language given a sufficient number of minimal pairs. However, it does not seem plausible that this strategy is available to infants (Goudbeek et al., 2008). By the time phonetic categories are learned, the infant's lexicon is still relatively small (Bergelson & Swingley, 2013) and is likely to contain mostly phonetically dissimilar words (Caselli et al., 1995) that do not lend themselves to minimal pair analysis.

Another possibility, which is the predominant approach in the literature at present, is that infants discover phonetic categories by implicit analysis of the distributional

* University of Edinburgh, United Kingdom
s1036048@sms.ed.ac.uk

† University of Pennsylvania, United States
swingley@psych.upenn.edu

properties of speech sounds. Perceptual categories can be viewed as 'point clouds' that cluster in certain regions of a multidimensional acoustic space (e.g. Goudbeek et al., 2009) and are thus distributed non-uniformly along the dimensions of that space. According to distributional learning theories (e.g. Maye & Gerken, 2000), infants discover phonetic categories by tracking the frequency of speech sounds together with their acoustic properties. The notion of distributional learning is supported at a general level by research indicating that infants are capable of tracking and using statistical information regarding their linguistic input (e.g. Kuhl, 2000; Saffran, Aslin, & Newport, 1996).

Providing more specific support for the role of distributional learning in phonetic category learning, artificial language learning experiments suggest that bi-modally distributed speech sounds are learnable by infants through mere exposure to the distribution, while unimodal distributions do not afford category acquisition in the same way (Maye, Werker, & Gerken, 2002). In this pioneering research, Maye et al. exposed infants to one of two kinds of training, and then evaluated their learning using a discrimination task. During training, all participants listened to syllables that consisted of a variable consonant and an invariable vowel and formed a continuum from / d a / to / t a /. For infants in the experimental condition, the end points of that continuum occurred more frequently, forming a bimodal frequency distribution, while infants in the control condition listened to a unimodal distribution where the syllables in the center of the continuum occurred most often. Maye et al. hypothesized that infants exposed to a bimodal frequency distribution would form a two-category representation of the continuum, while those exposed to a unimodal distribution would treat all stimuli as belonging to the same category. This prediction was confirmed by infants' performance on a discrimination task after training, which showed that those with a bimodal training distribution more readily discriminated the end points of the continuum than those who had been exposed to a unimodal distribution. An analogous experiment, using a same/different category judgment task, found similar effects for adult learners (Maye & Gerken, 2000). In addition to its theoretical importance, this finding is valuable from a methodological perspective because it suggests that researchers may use adults as models for infant learners to study distributional learning.

It is important to note, however, that current artificial language learning experiments in support of the distributional learning account of phonetic categories assume idealized learning conditions. Most importantly, the distributions of naturally occurring speech sounds are rarely as clean as the idealized distributions used in these experiments (Goudbeek et al., 2009). If the learning of phonetic categories depends mainly on the detection of separate distributional clusters, it is not clear how such learning would occur in natural learning settings where categories overlap significantly (Swingley, 2009). An important question, therefore, is what factors can enhance the effect of distributional learning under conditions where acoustic cues are not sufficient to disambiguate phonetic categories.

One possibility has been explored in supervised learning experiments where learners receive feedback about their performance (e.g. Francis, Nusbaum, & Fenn, 2007), or are explicitly instructed as to what properties of the acoustic signal to focus on (e.g. Kondaurova & Francis, 2010). It has been found that corrective feedback and instruction do indeed enhance the efficacy of distributional learning (Goudbeek et al., 2009). This has important implications for adult learners, because supervision of that kind may be helpful for detecting phonetic contrasts in a foreign language that are not present in their native language (Kondaurova & Francis, 2010). Infants, however, do not receive explicit instructions and feedback is unavailable at least until they begin to produce words (Goudbeek et al., 2008). Supervised learning is, therefore, not a likely candidate for enhancing distributional learning in infants.

Another possibility comes from the fact that in natural settings, speech sounds rarely occur in isolation, and are typically embedded in larger linguistic structures such as syllables and words. In a review of the literature on the interaction between knowledge of the lexicon and language development, Swingley (2009) proposed that lexical knowledge may facilitate the acquisition of phonetic categories. Evidence that lexical knowledge can support phonological processing comes from Thiessen (2007), who addressed the question of why children are not as sensitive to phonetic contrasts during word learning as they are during discrimination tasks (Pater, Stager, & Werker, 2004; Stager & Werker, 1997). He found that contrastive vowels that were embedded in phonologically dissimilar lexical contexts were more readily treated as contrastive during word learning than vowels whose lexical contexts were phonetically similar. He concluded that the words in which speech sounds occur provide distributional information that encourages learners to pay attention to the phonetic differences between those sounds. Providing further support for the role of lexical knowledge in phonological processing, Eisner and McQueen (2005) demonstrated that the lexical status of words containing phonetically ambiguous segments affected adults' subsequent perception of the ambiguous speech sounds in isolation. In particular, the lexical contexts of the ambiguous speech sounds constituted real words for one phonetic interpretation, and non-words for the other. Participants' subsequent perception of the speech sounds in isolation was biased towards the interpretation that had been compatible with a real word. While this study was designed and interpreted in the context of talker normalization (aligning a given speaker's voice to one's own phonological system), rather than phonetic category learning, it illustrates the notion that lexical knowledge may affect the perception of speech segments.

Although the infants studied by Thiessen (2007) were between 14 and 17 months old, it is widely accepted that infants know the phonological forms of some words well before their first birthday (Jusczyk & Hohne, 1997; Jusczyk, 1999). A recent study has reported evidence that infants as young as six months have some knowledge of the semantic

properties of certain words (Bergelson & Swingley, 2012). It is, therefore, possible that infants' developing lexical knowledge supports the acquisition of phonetic categories by providing additional distributional cues (Feldman, Griffiths, & Morgan, 2009; Swingley, 2009).

To test this possibility, Feldman, Myers, White, Griffiths, and Morgan (2011; 2013 Experiment 1) used an artificial language learning paradigm (modeled after Maye and Gerken's experiment in 2000) and exposed adult learners to a uniform distribution of tokens from a vowel continuum. The end points of this continuum – / a / and / ɔ / – are treated in some dialects of English as a single phonetic category, and as two contrastive categories in others (Labov, 1991), so that both one- and two-category interpretations of the stimuli were plausible. While all subjects were exposed to the same distribution of sounds, they were embedded in lexical contexts that favored either a one-category (control condition) or a two-category interpretation (experimental condition) of the vowel continuum. Note that this design reflected a departure from supervised learning paradigms, because the participants did not receive explicit instructions or feedback during training. However, since participants in the experimental condition were exposed to lexical contexts that provided reliable distributional cues to the delineation of the target categories, the learning environment was not entirely unsupervised either - such designs are typically referred to as instances of semi-supervised learning (Chapelle, Schölkopf, & Zien, 2006). As hypothesized, the results indicated that subjects in the experimental condition treated the vowels as belonging to two categories that were aligned with the distribution of words, while control subjects treated the continuum as a single category. The authors concluded that word-level information is used to categorize speech sounds even if they are uniformly distributed and is, therefore, a plausible candidate for enhancing distributional phonetic category acquisition.

The current study

To better understand the effect of lexical context on phonetic category acquisition, the current study exposed participants to a range of vowels with varying formant structures, and subsequently tested them using a perceptual categorization task. During training, subjects were exposed to six monosyllabic non-words whose vowels followed a strictly bimodal frequency distribution in that they had either high or low second formants. For subjects in the experimental condition, these phonetic categories were aligned with lexical categories, such that vowels with a relatively low second formant (F2) were consistently embedded within three of the six non-words, and those with high F2 values were associated with the other three lexical contexts. In order to enhance a lexical interpretation of the training stimuli, they were presented together with referent pictures. After exposure, subjects were presented with a forced-choice categorization task using isolated vowels from the same formant range. It was predicted that subjects in the experimental condition would perform better at categorization than control subjects, because the lexical contexts of the training vowels were informative about phonetic categories in the experimental, but not the control, condition.

In line with Feldman et al.'s (2013, 2011) studies, this design allowed an investigation of whether the lexical context of speech sounds may enhance distributional learning. However, this study's design differs from theirs in three important aspects. First, the participants were trained on a strictly bimodal frequency distribution, whereas Feldman et al.'s participants listened to a uniform distribution. Consequently, all of the current participants received distributional acoustic cues and only those in the experimental condition received an additional lexical cue. Second, Feldman et al.'s design involved a different test of category learning. In particular, while the current participants heard one vowel on each trial and were asked to assign them to one of two categories, Feldman et al. presented participants with two vowels per trial and asked them to decide whether they belonged to the same or to different categories. Finally, to extend Feldman et al.'s design, two modifications were made: first, a categorization task was added prior to training (i.e., a 'pre-test', see below) to screen participants for idiosyncratic processing strategies and, second, the training stimuli were presented together with referent pictures to underscore the lexical nature of the non-words. All of those aspects are discussed in more detail in the following section.

Method

Design

A factorial design was adopted, wherein subjects were randomly assigned to one of two conditions (experimental or control) and exposed to one of two sets of training stimuli accordingly. Subsequently, all participants performed a forced-choice auditory categorization task. To explore the possibilities that individual differences may affect the results, participants performed a structurally-equivalent categorization task prior to training.

Participants

Forty undergraduate students between 19 and 22 years of age (21 female and 19 male) from the University of Pennsylvania participated for course credit. All were native English speakers and none were bilingual from birth. All participants were naïve to the purpose of the study and were not rewarded financially.

Materials

Several tokens of the non-word 'thoss' were produced by a female speaker and recorded. The vowels were extracted and their formants analyzed using the Praat software (Boersma, 2002). 'Thoss' was chosen to minimize effects of coarticulation on the quality of the vowel because /th/ and /s/ can be readily segmented from the vowel, which carries minimal transition information. The resulting vowel tokens were resynthesized to match one of twelve F1 levels (805; 816; 827; 839; 850; 862; 873; 885; 897; 909; 922; 934) and one of twelve F2 levels (1218; 1246; 1275; 1304; 1333; 1364; 1395; 1426; 1458; 1491; 1525; 1559), giving rise to a tightly-constrained vowel grid of

144 tokens defined in terms of their first and second formants (cf. **Figure 1a**).

Another set of vowels at the extreme ends of the F2 scale were created in the same way (cf. **Figure 1b**). First, two uniform distributions of random numbers were generated. These constants were added to or subtracted from the measured F1 and F2 values of an existing natural vowel token, creating resynthesized targets for the first and second formants of each token as shown in **Figure 1b**. These resynthesized vowels were then re-inserted into six consonant contexts, creating 16 tokens each of 6 non-words: 'dop', 'thoss', 'vot', 'skod', 'pok', and 'gogz'. For the pre-test a subset of 32 of the test-stimuli was used, which preserved the range of F1 and F2 (cf. **Figure 1c**).

This procedure provided good control over vowel formants and ensured sufficient similarity between training and test stimuli. Six pictures of referents were paired with the six non-words in order to underline the referential nature of the lexical context, in which the target speech sounds were embedded. The pictures, shown in **Figure 2**, were photographs of unusual objects or conjunctions of objects that were constructed in Anne Fernald's lab in the early 1990s. They were selected for being readily namable in the absence of an existing conventional label.

Procedure

The experiment was conducted using PsychoPy software (Peirce, 2007) running on a MacBook Pro with a 13 inch screen. In addition to written, on-screen instructions, the experimental procedure was explained to the participants verbally and any questions resolved prior to testing. Participants were given a language background questionnaire, which served to confirm that they were monolingual, native English speakers. After signing the consent form and completing the language background questionnaire, participants sequentially completed three experimental phases: a pre-test; a training session; and a post-test. Each phase was preceded by detailed, written, on-screen instructions; none of the subjects reported any uncertainty or confusion in relation to the tasks.

Training phase. In 288 trials, participants were exposed to three blocks of 96 training stimuli. As described above, these stimuli consisted of 16 vowels with varying first and second formants (see **Figure 1b**), embedded in six lexical contexts, forming the non-words 'thoss', 'vot', 'skod', 'pok', 'dop', and 'gogz'. A cover task directed participants' attention towards the acoustic properties of the stimuli, without revealing the true purpose of the training phase. During the cover task, subjects were asked to rate each of the auditory stimuli in terms of its 'perceived naturalness' on a 5-point Likert scale, where '1' signified 'very synthetic' and '5' stood for 'very natural'. None of the subjects reported any suspicion that the naturalness rating was a cover task, and all remained naïve to the purpose of the training phase.

To add plausibility to a lexical interpretation of the auditory stimuli as words, each was paired and presented together with one of six referent pictures (see **Figure 2**). Thus, for a given participant, the word 'vot' was always

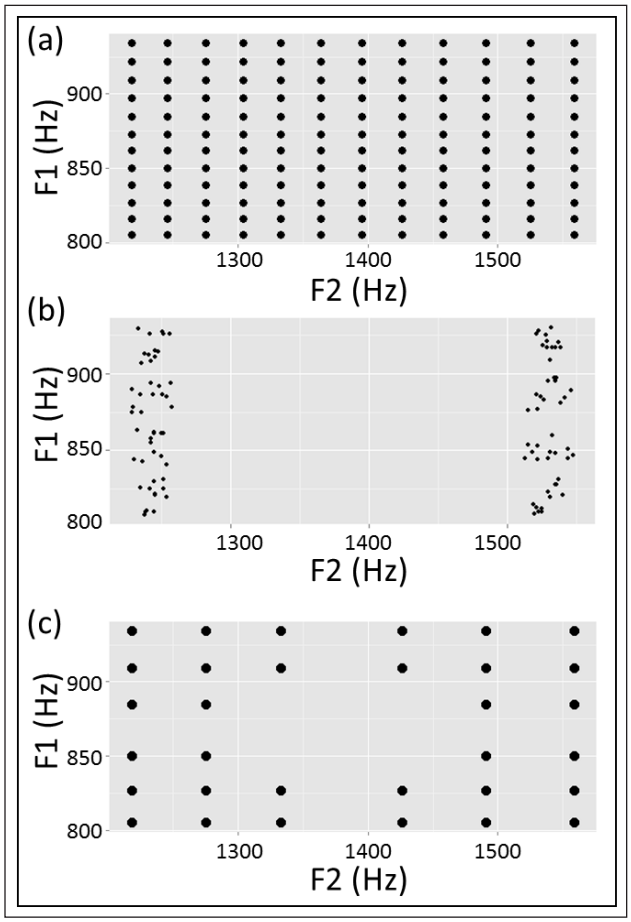

Figure 1: Vowels by first and second formant presented during the post-test (a), the training (b), and the pre-test (c).

accompanied by a particular object's image; the word 'dop' was always accompanied by a different particular image and so forth. It was expected that this would underline the referential nature of the non-words and that they would consequently be treated more like lexical items. Subjects in the experimental condition and the control condition were exposed to the same set of vowels, but their training differed with respect to the association between words and phonetic categories. In the experimental ('aligned') condition, three of the six words contained low-F2 vowels only, and the other three words contained only vowels with high second formants. These word categories were thus aligned with the phonetic categories the subjects were supposed to learn. In the control condition however, each word was associated with both high- and low-F2 vowels, such that the lexical context was uninformative about the to-be-learned phonetic categories.

To counteract the possibility that spurious correlations between acoustic properties of the stimuli were systematically affecting how subjects approached the stimuli, they were presented in a pseudo-random order that was constrained to avoid excessive repetitions and clusters with regard to F1, F2, or the lexical context. The first 10 trials were controlled more rigorously to avoid accidental contrasts (e.g. high vs. low F1) that could bias subjects early on towards particular hypotheses about the target categories.

Figure 2: Referent pictures paired with the non-words that were presented during training.

As the perception of vowels may be affected by their immediate phonetic context (e.g. Viswanathan, Magnuson, & Fowler, 2010), words that were associated with each phonetic category in the aligned condition were counter-balanced, such that words 1, 2, and 3 contained low-F2 vowels for half the subjects and high-F2 vowels for the other half.

Test phase. After exposure to the training stimuli, subjects were presented with 288 vowel tokens in a forced-choice categorization task. During each trial, subjects listened to one of 144 vowels that differed from each other with regard to their first and second formants (cf. **Figure 1a**) and were presented twice in separate blocks. They were then asked to sort the vowel into one of two categories. To avoid confusion about how to use the categories, subjects were explicitly instructed to assign similar sounding vowels to the same category and different sounding vowels to separate categories.

Pre-test. The pre-test was structurally identical to the post-test, but used only a subset of the stimuli employed in the latter (see **Figure 1c**). The purpose of the pre-test was to explore the possibility that participants bring idiosyncratic processing biases to the task. Individual differences in the way the speech sounds in question were approached could potentially obscure the effect of distributional learning if there was one. However, since extensive exposure to the relevant phonetic information could also induce the formation of processing strategies which could interfere with any learning during training, only a subset of the post-test stimuli were chosen for the pre-test. The intention was that that pre-test be sufficient to reveal, but not to form, idiosyncratic processing strategies. In 32 trials, participants were thus presented with vowels reflecting the full formant range of the post-test stimuli and faced an identical forced-choice categorization task.

Categorization Response	Phonetic Category	
	Low F2	High F2
Left	6	105
Right	138	39

Table 1: Exemplary post-test response pattern.

Analysis Strategy

The experimental design provided two sets of binary responses for each subject: 32 from the pre-test; and 288 from the post-test. The primary dependent measure was subjects' performance on the post-test categorization task, which was defined as the proportion of correct vs. incorrect responses. Instead of using the center category boundary, however, all possible vertical category boundaries were considered and the one that best described the subject's response pattern was used as a standard against which her or his performance was measured. Based on this normative standard, the percentage of correct responses reflected the number of correctly identified category members divided by the total number of categorization judgments. For illustration, the example response pattern in **Table 1** can be considered:

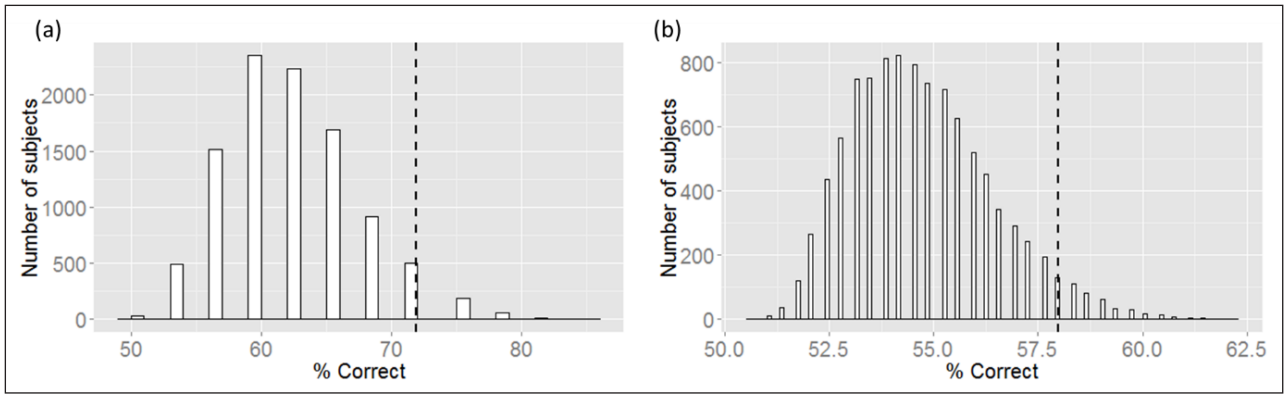

Figure 3: Chance distribution of percentage of correct responses to pre-test (a) and post-test stimuli (b). Dashed lines indicate top 5% threshold.

This particular subject sorted 138 of the 144 low-F2 vowels into one category (labeled 'Left') but failed to recognize 6 vowels that belonged to the same category. As for the high-F2 vowels, 105 of them were correctly categorized and 39 were falsely judged as belonging to the low-F2 category. The percentage of correct responses for this subject is therefore 84.74%.

As a result of this procedure, it was mathematically impossible to receive a score lower than 50% and the chance-level performance was thus necessarily higher than 50%. While the primary test of the hypothesis involved a comparison of mean performance across conditions, the chance level for each task was also calculated. Ten thousand hypothetical subjects were modeled who randomly categorized each of the pre-test and post-test stimuli and calculated the percentage of correct responses according to the above formula. As for the experimental subjects, all possible vertical category boundaries for each hypothetical subject were tested and the one that was the best fit of the subject's responses was selected. This boundary served as the subject-specific standard for evaluating individual responses as correct or incorrect. **Figure 3** shows the distribution of individual performance scores of the 10,000 random subjects. The best 5% of this random population categorized at least 57.99 % of the post-test stimuli and at least 71.88 % of pre-test stimuli correctly. These values were used as the chance level against which the performance of the experimental subjects was compared.

In addition to the proportion of correct responses, subjects' performance was evaluated using logistic regression analyses with F1 and F2 of the stimuli as continuous predictors and categorization choice as the binary dependent variable. The predictors were centered by subtracting them from the mean F1 and F2 values of all stimuli.

All group comparisons used the Welch two-sample t-test to account for the possibility of unequal variances and non-parametric distribution at the population level.

Results

Case Exclusion and Data Preparation

In total, four subjects were excluded from the analysis: two because they judged more than 90% of the stimuli in the post-test as belonging to the same category and their data were uninformative for the purpose of the present study; and two because of experimenter error. Among the remaining participants, there was no difference in post-test performance between counter-balanced trial orders, $t\,(33) = 0.76$, $p = 0.45$, and the data were collapsed for all subsequent analyses.

Naturalness ratings

While the five-point Likert scale ratings of the naturalness of the training stimuli were collected only to focus the subjects' attention on the acoustic properties of the stimuli and were not of theoretical interest, they also served the purpose of screening for inattentive participants. All subjects used at least three levels of the Likert-scale and most reported spontaneously after the experiment that they had taken the rating task seriously.

Group Results

As the primary test of the hypothesis that the kind of training subjects received would affect the way they categorized the test vowels, the mean percentage of correct responses during post-test categorization across conditions was compared (cf. **Figure 4a**). Subjects in the aligned condition (N = 18) responded on average correctly on 79.32% of post-test trials (SD = 8.83) and did slightly better than subjects in the control condition (N = 18) whose mean performance was 76.81% (SD = 10.46). A t-test revealed that the difference was not significant, however (cf. **Figure 4a**), $t\,(33) = 0.78$, $p = 0.44$.

To be thorough, another comparison across conditions using logistic regression was performed. F1 and F2 were entered as predictors with categorization response as a binary dependent variable to find the best-fit model for each subject. This alternative performance measure afforded a complementary group comparison of mean betas, which reflect the contribution of F2 variance in the stimulus set to subjects' categorization choices. As expected, Spearman's rank correlation revealed that the resulting beta values were positively correlated with the percentage of correct responses, $\varrho = .4$, $p = 0.016$. The comparison of mean betas across conditions revealed that they did not differ significantly between the aligned (N = 18) and the control (N = 18) conditions ($M = 0.017$; $SD = 0.01$), $t\,(33) = 0.002$, $p > 0.99$.

Finally, each subject's individual post-test performance was evaluated by comparing it to the chance distribution shown in **Figure 3b**. **Table 2** summarizes the number of subjects across conditions who scored as good as or better than the best 5% of the chance population. The number of subjects who did or did not differ from chance was identical across conditions, which was formally confirmed by a chi-square test, χ^2 (1, N = 36) = 0, p = 1.

These analyses are compatible with the null hypothesis and disconfirm the original prediction that subjects in the aligned condition would outperform control subjects on the post-test categorization task.

Pre-test response patterns

As a number of subjects were performing close to the overall maximum score in both conditions, it is possible that the group comparisons were affected by an overall ceiling effect that was independent of training. To test this hypothesis, subjects' pre-test performances were examined. It was found that their responses followed one of three patterns. Some subjects (N = 16) categorized the stimuli systematically based on F2. Other subjects (N = 9) used both F1 and F2 for categorization in an unsystematic way and performed at chance level (i.e., they categorized less 57.99% of the pre-test stimuli correctly). A third set of subjects (N = 11) responded uniformly or near-uniformly by sorting more than 80% of the pre-test stimuli into one category and less than 20% into the other category. **Figure 5** illustrates the nature of the three response patterns.

Note that the subjects who used F2 systematically in the pre-test were already paying attention to the correct dimension prior to training and were, therefore, less trainable than those who performed at chance or uniformly. As shown in **Table 3**, all three response patterns were found to similar extents in both conditions, which made it possible to re-run the group comparisons of subjects based on their trainability as reflected in their pre-test response pattern.

Group comparison reconsidered

Accounting for subjects who did not use F2 systematically for categorization during the pre-test, the difference between conditions was found to be slightly larger, but still not significant (cf. **Figure 4b**). The remaining subjects in the aligned condition (N = 10) performed at 77.47% (SD = 11.02) while those in the control condition (N = 10) scored at 72.43% (SD = 11.8), t (18) = 0.99, p = 0.34. In accordance with this result, comparing the results of the logistic regression analyses revealed that categorization choices of subjects in the aligned (N = 10) and the control conditions (N = 10) were predicted by F2 variance equally well (M = 0.01; SD = 0.01), t (18) = 0.76, p = 0.46.

Finally, the same test on a third subset of subjects (solely those who performed at chance but not uniformly during the pre-test) returned a larger difference across conditions, but also failed to reach the significance threshold (cf. **Figure 4c**). The remaining subjects in the aligned condition (N = 5) performed at 81.6% (SD = 6.94) in the post-test, compared to those in the control condition (N = 4) who scored at 68.66% (SD = 14), t (4) = 1.69, p = 0.16. This result, too, was mirrored in the comparison of betas from the logistic regression analyses. F2 was predictive of categorization choice for subjects in the aligned condition (N = 5; M = 0.02; SD = 0.01) as well as for those in the control condition (N = 4; M = 0.01; SD = 0.01), t (4) = 0.949, p = 0.39.

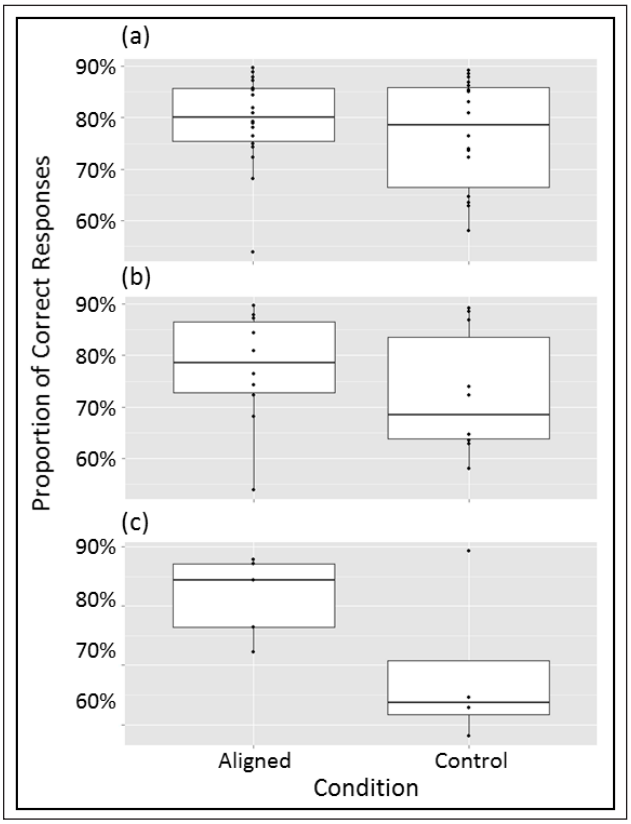

Figure 4: Post-test performance in percentage correct across condition based (a) all subjects (N = 36); (b) subjects that did not use F2 systematically on the pre-test (N = 20); and (c) subjects that performed at chance, but not uniformly, during the pre-test (N = 9).

Difference from Chance	Condition	
	Aligned	Control
Yes	17	17
No	1	1

Table 2: Number of subjects in each condition who performed better than chance in the post-test.

Discussion

All group comparisons revealed that there was no significant difference in post-test performance between conditions, which disconfirms the research hypothesis. However, in both conditions a number of participants performed close to the overall maximum, which suggests that a ceiling effect may have concealed an effect of training. To explore this possibility, the pre-test data were analyzed and it was discovered that subjects displayed different response patterns. Most notably, eight subjects in each condition were already using the second formant

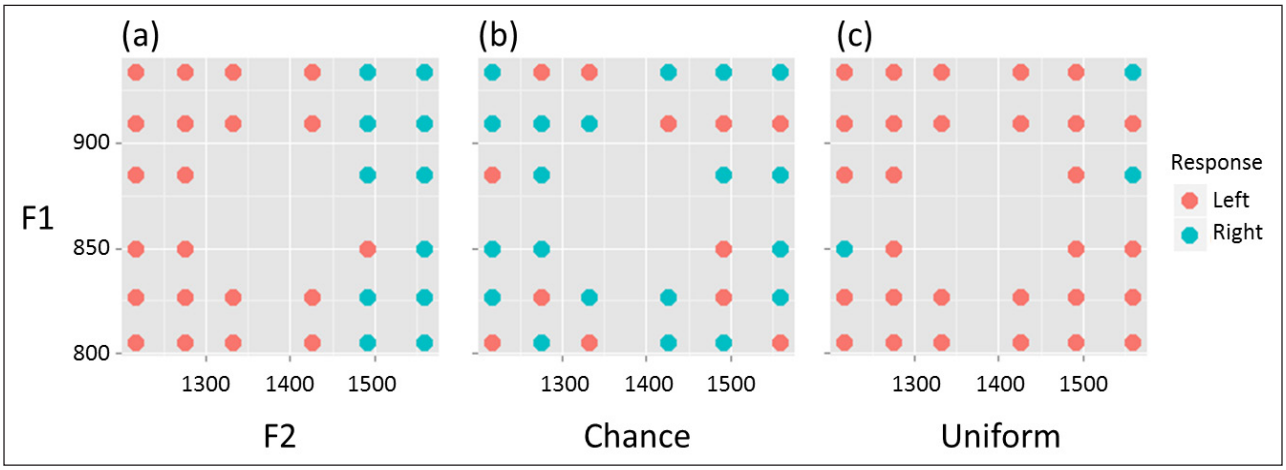

Figure 5: Pre-test response patterns of examplary subjects who were classified as (a) "F2", (b) "Chance", or (c) "Uniform".

Experimental Condition	Pre-test Response Pattern		
	'F'	'Chance'	'Uniform'
Aligned	8	5	5
Control	8	4	6

Table 3: Crosstabulation of Condition and Pre-test Response Pattern.

systematically for categorization prior to training. In a sense, these subjects were not trainable because they were already using the strategy intended for them to 'discover' during training. Unsurprisingly, those subjects scored highly on the post-test, but their performance could not be attributed to the effect of training. Subset analyses revealed that the difference between conditions increased as less trainable subjects were excluded, but it failed to reach the significance threshold. As the size of the subsets was relatively small, however, the statistical comparisons are not conclusive. In what follows, three potential modifications to the current experimental design are described that may yield more informative results.

First, following Feldman et al. (2013, 2011) in using a same/different task instead of sorting vowels into categories may yield more informative pre-test results. For example, 11 of the current subjects responded uniformly or near-uniformly on the pre-test, and some of them reported after the experiment that they had been expecting more striking differences. That is, they were aware that each stimulus belonged to one of two categories, but were uncertain about the formant range of the phonetic space in question. Thus, their response pattern did not conclusively indicate how they approached the two acoustic dimensions. A same/different task may focus their attention on the discriminability of individual vowels instead of estimating the size of the total vowel space, and capture more directly their perceptual appraisal of the vowels. As a result, some of the subjects who performed uniformly in the current pre-test may pay attention systematically to either F1 or F2, or perform at chance. Thus, this procedure may reflect more faithfully the individual differences subjects bring to the task, and yield more conclusive group comparisons when this information is used for subsetting.

A second modification could involve increasing the spread of F1 to make it more compelling for categorization. None of the subjects used F1 systematically prior to training whereas a remarkable 44% used F2, which suggests that the vowel space may have biased subjects across conditions to use F2. Increasing the spread of F1 may make the variation along this dimension more salient, and as a result some subjects may initially hypothesize that F1 is the relevant dimension for categorization. Such subjects would be maximally trainable, and it would be interesting to see whether the training stimuli presented in the aligned condition would be more effective in changing their strategy than those presented in the control condition.

A third modification is related to the idea of using subjects' initial hypotheses to test their propensity to change their categorization strategy based on the training they receive. Employing the same test stimuli used here, but changing the training distribution to target F1 rather than F2, could make use of the fact that many subjects used F2 prior to training for testing whether they were more likely to change their strategy after being exposed to the lexically informative training distribution, as compared to the lexically uninformative distribution. Ideally, subjects' pre-test response pattern would be evaluated online and dynamically determine which acoustic dimension was targeted during training. One of the main results reported here is that subjects used different strategies for categorizing the test vowels prior to training. Finding individual response strategies prior to training would enable researchers to provide a training distribution that would encourage all subjects to change their strategy, which would be ideal for testing the effect of word-level information on the acquisition of phonetic categories.

References

Bergelson, E., & **Swingley, D.** (2012). At 6–9 months, human infants know the meanings of many common nouns. *Proceedings of the National Academy of Sciences*, *109*(9), 3253–3258. DOI: http://dx.doi.org/10.1073/pnas.1113380109

Bergelson, E., & **Swingley, D.** (2013). The acquisition of abstract words by young infants. *Cognition*, *127*(3), 391–397. DOI: http://dx.doi.org/10.1016/j.cognition.2013.02.011

Best, C. T. (1995). Learning to perceive the sound pattern of English. *Advances in Infancy Research*, *9*, 217–217.

Boersma, P. (2002). Praat, a system for doing phonetics by computer. *Glot International*, *5*(9/10), 341–345.

Bosch, L., & **Sebastián-Gallés, N.** (2003). Simultaneous Bilingualism and the Perception of a Language-Specific Vowel Contrast in the First Year of Life. *Language and Speech*, *46*(2–3), 217–243. DOI: http://dx.doi.org/10.1177/00238309030460020801

Caselli, M. C., Bates, E., Casadio, P., Fenson, J., Fenson, L., Sanderl, L., & **Weir, J.** (1995). A cross-linguistic study of early lexical development. *Cognitive Development*, *10*(2), 159–199. DOI: http://dx.doi.org/10.1016/0885-2014(95)90008-X

Chapelle, O., Schölkopf, B., & **Zien, A.** (2006). *Semi-supervised learning* (Vol. 2). Cambridge, MA: MIT press. DOI: http://dx.doi.org/10.7551/mitpress/9780262033589.001.0001

Eisner, F., & **McQueen, J. M.** (2005). The specificity of perceptual learning in speech processing. *Perception & Psychophysics*, *67*(2), 224–238. DOI: http://dx.doi.org/10.3758/BF03206487

Feldman, N. H., Griffiths, T. L., & **Morgan, J. L.** (2009). Learning phonetic categories by learning a lexicon. In *Proceedings of the 31st Annual Conference of the Cognitive Science Society* (pp. 2208–2213). Amsterdam. DOI: http://dx.doi.org/10.1037/a0017196

Feldman, N. H., Myers, E., White, K., Griffiths, T., & **Morgan, J. L.** (2011). Learners use word-level statistics in phonetic category acquisition. In *Proceedings of the 35th Annual Boston University Conference on Language Development* (pp. 197–209). Somerville, MA. Retrieved from http://www.cog.brown.edu/research/infant_lab/Feldman%20Learners%20Use%20Word%20Level.pdf

Feldman, N. H., Myers, E. B., White, K. S., Griffiths, T. L., & **Morgan, J. L.** (2013). Word-level information influences phonetic learning in adults and infants. *Cognition*, *127*(3), 427–438. DOI: http://dx.doi.org/10.1016/j.cognition.2013.02.007

Francis, A. L., Nusbaum, H. C., & **Fenn, K.** (2007). Effects of training on the acoustic phonetic representation of synthetic speech. *Journal of Speech, Language, and Hearing Research*, *50*(6), 1445. DOI: http://dx.doi.org/10.1044/1092-4388(2007/100)

Goudbeek, M., Cutler, A., & **Smits, R.** (2008). Supervised and unsupervised learning of multidimensionally varying non-native speech categories. *Speech Communication*, *50*(2), 109–125. DOI: http://dx.doi.org/10.1016/j.specom.2007.07.003

Goudbeek, M., Swingley, D., & **Smits, R.** (2009). Supervised and unsupervised learning of multidimensional acoustic categories. *Journal of Experimental Psychology: Human Perception and Performance*, *35*(6), 1913–1933. DOI: http://dx.doi.org/10.1037/a0015781

Jusczyk, P. W. (1999). How infants begin to extract words from speech. *Trends in Cognitive Sciences*, *3*(9), 323–328. DOI: http://dx.doi.org/10.1016/S1364-6613(99)01363-7

Jusczyk, P. W., & **Hohne, E. A.** (1997). Infants' Memory for Spoken Words. *Science*, *277*(5334), 1984–1986. DOI: http://dx.doi.org/10.1126/science.277.5334.1984

Kondaurova, M. V., & **Francis, A. L.** (2010). The role of selective attention in the acquisition of English tense and lax vowels by native Spanish listeners: Comparison of three training methods. *Journal of Phonetics*, *38*(4), 569–587. DOI: http://dx.doi.org/10.1016/j.wocn.2010.08.003

Kuhl, P. K. (2000). A new view of language acquisition. *Proceedings of the National Academy of Sciences*, *97*(22), 11850–11857. DOI: http://dx.doi.org/10.1073/pnas.97.22.11850

Kuhl, P. K., Williams, K. A., Lacerda, F., Stevens, K. N., & **Lindblom, B.** (1992). Linguistic experience alters phonetic perception in infants by 6 months of age. *Science*, *255*(5044), 606–608. DOI: http://dx.doi.org/10.1126/science.1736364

Labov, W. (1991). The three dialects of English. *New Ways of Analyzing Sound Change*, *5*, 1–44.

Lalonde, C. E., & **Werker, J. F.** (1995). Cognitive influences on cross-language speech perception in infancy. *Infant Behavior and Development*, *18*(4), 459–475. DOI: http://dx.doi.org/10.1016/0163-6383(95)90035-7

Maye, J., & **Gerken, L. A.** (2000). Learning phonemes without minimal pairs. In *Proceedings of the 24th Annual Boston University Conference on Language Development* (Vol. 2, pp. 522–533). Somerville, MA: Cascadilla Press. Retrieved from http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.188.5940&rep=rep1&type=pdf

Maye, J., Werker, J. F., & **Gerken, L.** (2002). Infant sensitivity to distributional information can affect phonetic discrimination. *Cognition*, *82*(3), B101–B111. DOI: http://dx.doi.org/10.1016/S0010-0277(01)00157-3

Pater, J., Stager, C., & **Werker, J.** (2004). The perceptual acquisition of phonological contrasts. *Language*, 384–402. DOI: http://dx.doi.org/10.1353/lan.2004.0141

Peirce, J. W. (2007). PsychoPy—Psychophysics software in Python. *Journal of Neuroscience Methods*, *162*(1–2), 8–13. DOI: http://dx.doi.org/10.1016/j.jneumeth.2006.11.017

Polka, L., & **Werker, J. F.** (1994). Developmental changes in perception of nonnative vowel contrasts. *Journal of Experimental Psychology-Human Perception and Performance*, *20*(2), 421–435. DOI: http://dx.doi.org/10.1037/0096-1523.20.2.421

Saffran, J. R., Aslin, R. N., & **Newport, E. L.** (1996). Statistical learning by 8-month-old infants. *Science*, *274*(5294), 1926–1928. DOI: http://dx.doi.org/10.1126/science.274.5294.1926

Stager, C. L., & **Werker, J. F.** (1997). Infants listen for more phonetic detail in speech perception than in word-learning tasks. *Nature*, *388*(6640), 381–382. DOI: http://dx.doi.org/10.1038/41102

Swingley, D. (2009). Contributions of infant word learning to language development. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *364*(1536), 3617–3632. DOI: http://dx.doi.org/10.1098/rstb.2009.0107

Thiessen, E. D. (2007). The effect of distributional information on children's use of phonemic contrasts. *Journal of Memory and Language*, *56*(1), 16–34. DOI: http://dx.doi.org/10.1016/j.jml.2006.07.002

Viswanathan, N., Magnuson, J. S., & **Fowler, C. A.** (2010). Compensation for coarticulation: Disentangling auditory and gestural theories of perception of coarticulatory effects in speech. *Journal of Experimental Psychology: Human Perception and Performance*, *36*(4), 1005. DOI: http://dx.doi.org/10.1037/a0018391

Werker, J. F., & **Pegg, J. E.** (1992). Infant speech perception and phonological acquisition. *Phonological Development: Models, Research, Implications*, 285–311.

Werker, J. F., & **Tees, R. C.** (1984). Cross-language speech perception: Evidence for perceptual reorganization during the first year of life. *Infant Behavior and Development*, *7*(1), 49–63. DOI: http://dx.doi.org/10.1016/S0163-6383(84)80022-3

How to cite this article: Poppels, T. and Swingley, D. (2014). Semi-supervised Phonetic Category Learning: Does Word-level Information Enhance the Efficacy of Distributional Learning? *Journal of European Psychology Students,* 5(3), 36-45, DOI: http://dx.doi.org/10.5334/jeps.ce

Published: 22 August 2014

Copyright: © 2014 The Author(s). This is an open-access article distributed under the terms of the Creative Commons Attribution 3.0 Unported License (CC-BY 3.0), which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited. See http://creativecommons.org/licenses/by/3.0/.

]u[*Journal of European Psychology Students* is a peer-reviewed open access journal published by Ubiquity Press.

OPEN ACCESS