Appendix B: Complete Analyses of Accuracy

Performance in the training tasks was primarily evaluated in terms of the accuracy at test. While performance was very good (generally above 90% correct) in both alternation conditions, in both experiments, there were differences as a function of condition. We present complete analyses of both experiments here.

B.1. Experiment 1

In Experiment 1, accuracy was analyzed separately for Experiment 1a (the voicing alternation condition vs. control) and Experiment 1B (manner alternation vs. controls)

B.1.1. Experiment 1a: Accuracy

All groups were highly accurate at test (M= 94.2%). Table B1 shows accuracy broken down by language group for each type of stimulus. Preliminary analyses for both accuracy and fixation data showed no effect of o-condition (singular or plural), so we do not include it here or in subsequent analyses. Stimuli are broken down by prefix (/an/ or /o/) and underlying initial consonant (/d/, /t/, or /z/). The additional columns labeled “filler” refer to the trials in which the prefix alone was sufficient to determine the target picture. In these trials, it is assumed that if the participants have learned the meanings of each of the two prefixes, they can select the correct output, even if they are unsure of the noun labels.

To determine whether accuracy differed across conditions or item types, we conducted a repeated-measures ANOVA on empirical logit transformed data, with prefix type (/o/ vs. /an/) and item type (d, t, z, and filler) as the within-subjects factors and rule group (voicing vs. control) as the between-subjects factor. Results revealed a significant effect of prefix type, F(1, 32) = 8.33, $p = 0.007$, such that participants were a bit more accurate on the /o/-prefixed words than the /an/-prefixed words. There was also an effect of item type, F(3, 96) = 3.69, p = 0.014. There was no effect of rule group (F < 1). None of the interactions were significant.

Table B1. Average accuracy (standard deviation) by group for Experiment 1

	/an/					/o/				
	d	**t**	**z**	**filler**	**total**	**d**	**t**	**Z**	**filler**	**total**
control group	.946 (.039)	.943 (.088)	.954 (.052)	.957 (.031)	.952 (.054)	.944 (.051)	.954 (.063)	.946 (.065)	.954 (.037)	.951 (.058)
voicing group	.933 (.074)	.867 (.116)	.955 (.047)	.938 (.054)	.928 (.078)	.939 (.066)	.951 (.052)	.960 (.035)	.934 (.058)	.942 (.060)
manner group	.928 (.073)	.948 (.065)	.817 (.121)	.937 (.068)	.918 (.096)	.937 (.048)	.944 (.071)	.974 (.033)	.950 (.039)	.951 (.056)

Follow-up analyses exploring the effect of item type revealed that within the o-prefix, there were no effects of item type or group (all F < 1). Within the an-prefix, though, there was a significant effect of item type, F(3, 96) = 6.30, p = 0.001, and an interaction between item type and group, F(3, 96) = 3.18, p = 0.028. Posthoc pairwise comparisons using a Bonferroni adjustment for multiple comparisons revealed significant differences in accuracy between the /an+t/ items and the /an+z/ items (p = 0.008): Participants were somewhat less accurate for (alternating) /an+t/ stimuli, driven primarily by lower accuracy in the voicing group, for whom the /an+t/ stimuli were phonetically [an+d]. Interestingly, this lower accuracy did not extend to the /an+d/ items, which were also phonetically ambiguous.

B.1.2. Experiment 1b: Accuracy

Experiment 1b compared the manner alternation group with the control group. Again, both groups were highly accurate in the test trials, with an overall average accuracy of 94.3% (compare to 94.2% in Experiment 1a). Accuracy broken down by language-group and prefix are in Table B1.

A repeated-measures ANOVA, with prefix type (/o/ vs /an/) and item type (t, d, z, and filler) as the within-subjects factors and rule group (manner vs. control) as the between subjects factor was conducted on empirical logit transformed data. Results revealed significant effects of prefix type, F(1, 33) = 23.50, $p < 0.001$, with higher accuracy for /o/-initial trials. There was no main effect of item type, F(3, 99) = 1.90, $p = 0.136$, or of rule group, F < 1. There were significant two-way interactions between prefix type and rule group, F(1, 33) = 9.46, $p = 0.004$, between prefix type and item type, F(3, 99) = 5.81, $p = 0.001$, and a three-way interaction between prefix type, item type, and rule group, F(3, 99) = 9.82, $p < 0.001$.

Follow-up analyses revealed that there were no effects of prefix type, F(1, 15) = 1.39, $p = 0.257$ or item type (F < 1) within the control group. However, within the manner group, there was a main effect of prefix type, F(1, 18) = 35.61, $p < 0.001$, and item type, F(3, 54) = 3.13, p = 0.033, as well as an interaction between the two, F(3, 54) = 14.44, $p < 0.001$. Posthoc pairwise comparisons using a Bonferroni adjustment for multiple comparisons revealed differences in accuracy when the stimulus was /an + z/ compared with each of the other /an/ stimuli ($p < 0.05$ for each of the three comparisons), and when the stimulus was /o + z/ compared with /o + d/ or the /o/ filler item (both $p < 0.05$). As in Experiment 1a, participants in the manner group were less accurate with the alternating (/an + z/) stimuli (which were phonetically [an + d], resulting in ambiguity between /an + z/ and /an + d/ stimuli) than with the non-alternating (/an + t/) stimuli. Again, this lower accuracy did not extend to the /an + d/ items, which were also phonetically ambiguous.

B.2. Experiment 2

Participants in Experiment 2 (Table B.2) were less accurate at test than those in Experiment 1, though they achieved close to 90% accuracy averaged across the various trained items.

Table B.2. Average accuracy (standard deviation) by condition for Experiment 2. Note as this was accuracy only on the test trials, it only reflects the high frequency trained items.

		/an/					/o/				
		d	**t**	**z**	**filler**	**total**	**d**	**t**	**z**	**filler**	**total**
Voicing	**trained**	.909 (.084)	.764 (.200)	.961 (.044)	.950 (.074)	.896 (.069)	.908 (.097)	.903 (.112)	.947 (.066)	.946 (.046)	.926 (.062)
	gen.	.865 (.121)	.714 (.190)	.850 (.161)	.912 (.096)	.835 (.114)	.859 (.160)	.824 (.185)	.870 (.136)	.946 (.081)	.875 (.114)
Manner	**trained**	.915 (.110)	.939 (.070)	.768 (.178)	.922 (.114)	.886 (.094)	.944 (.086)	.952 (.056)	.917 (.113)	.925 (.102)	.934 (.080)
	gen.	.832 (.149)	.850 (.164)	.608 (.233)	.912 (.105)	.800 (.130)	.883 .131)	.886 (.121)	.861 (.170)	.928 (.101)	.890 (.104)

We again conducted a repeated-measures ANOVA on log-odds transformed data, with training type (trained vs. generalization), prefix type (/o/ vs. /an/), and item type (d, t, z, and filler) as within-subjects factors. These were conducted separately for the voicing and manner alternation groups.

B.2.1. Voicing alternation

Results revealed main effects of training type, F(1, 21) = 10.91, p = 0.003, such that participants were more accurate on trained items than generalization items. There was a marginal effect of prefix type, F(1, 21) = 3.16, p = 0.09), such that participants were more accurate on the /o/-prefixed words. Finally, item type was also significant, F(3, 63) = 12.93, p < 0.001. There was a significant two-way interaction between prefix-type and item type, F(3,

63) = 6.43, $p = 0.001$; all other interactions were non-significant. This interaction was due to differences in accuracy on the /t/-item by prefix: participants were more accurate when the stimulus began with [o + t] than [an + t] in both trained items, t(21) = 3.71, $p < 0.001$, and generalization items, t(21) = 3.28, $p = 0.004$; there were no differences due to prefix in the other item types. In sum, participants had greater accuracy on trained items than on generalization items, and they had the poorest overall accuracy on /an + t/ items, which were phonetically [an + d] and therefore ambiguous with the /an + d/ items.

B.2.2. Manner alternation

We again conducted a repeated-measures ANOVA on empirical-logit transformed data, with training-type (trained vs. generalization), prefix-type (/o/ vs. /an/), and item-type (d, t, z, and filler) as within-subjects factors. Results revealed main effects of training type, F(1, 21) = 13.05, $p = 0.002$, such that participants were more accurate on trained items than generalization items. There was also a main effect of prefix type, F(1, 21) = 19.52, $p < 0.001$, such that participants were more accurate on the /o/-prefixed words. And there was a main effect of item type, F(3, 63) = 18.57, $p < 0.001$. Posthoc pairwise comparisons using a Bonferroni adjustment for multiple comparisons revealed significant differences in accuracy when the item type was /z/ vs. each other item type (all $p < 0.001$).

There were significant two-way interactions between training-type and item-type, F(3, 63) = 3.28, $p = 0.027$ and between prefix-type and item-type, F(3, 63) = 7.83, $p < 0.001$. Posthoc comparisons found differences between the prefixes for /z/-items (training: t(21) = 4.84, $p < 0.001$; generalization: t(21) = 6.9, $p < 0.001$) and for filler items in the generalization words, t(21) = 2.24, $p = 0.036$, as well as differences due to training type for /z/-items (an-prefix: t(21) = 3.57, $p = 0.002$; o-prefix: t(21) = 2.59, $p = 0.017$) and d-items (an-prefix: t(21)

= 2.68, $p = 0.014$; o-prefix: t(21) = 3.47, $p = 0.002$).

In sum, participants were more accurate on trained items than on generalization items, and they had the poorest overall accuracy on /an + z/ items, which were phonetically [an + d] and therefore ambiguous with the /an + d/ items.