Supplementary analysis 1: extending categories beyond the frequent-words data used to derive them

Another way to evaluate the quality of the types-based and tokens-based categories derived using data from frequent words is to see how well each set of categories supports classification of vowel tokens that were not used to derive those categories, namely the vowels that appeared in relatively infrequent words. Implementing this analysis requires a procedure for assigning new tokens to each category. One procedure would be to consider each token's distance from the center of each derived category, and assign the each token to the nearest category. The problem with this procedure is that it does not reflect the boundaries of each category; a large one and a small one would be treated the same.

Instead, we implemented an algorithm that assigns each new token to the category of the nearest labeled token. For example, the word *derecho* only occurred once in the corpus, and was therefore not used to derive the types-based categories (and thus not the tokens-based categories either). The first [e] in *derecho* was closest (in z-scored F1, F2 space) to the average formant values of the first syllable of the frequent word *verdad*, and was therefore assigned to that syllable's category. The same was done for the tokens analysis: the first [e] in *derecho* was assigned to the closest token from the tokens analysis (in this case, an instance of the word *y*). In this way, every token in the corpus was classified. To be sure, this simple scheme is probably not how phoneme categorization actually works, but it has the virtue of acknowledging the overall shapes of the category splits determined by each analysis.

The categorization results are given in Fig. S1, which includes category assignments for the 1064 tokens used to generate the categories, together with the 1127 tokens assimilated into these categories using the nearest-neighbor procedure described above. The types-based categories were derived from orthographic types (as in Analysis 1); Fig. S2 shows results from the same procedure applied to the phonological, vowel-ignorant types of Analysis 2.

By comparison with Fig. 3, the orthographic types analysis includes more off-diagonal elements, but the main trends are the same. The extended types and tokens analyses place many [o] tokens

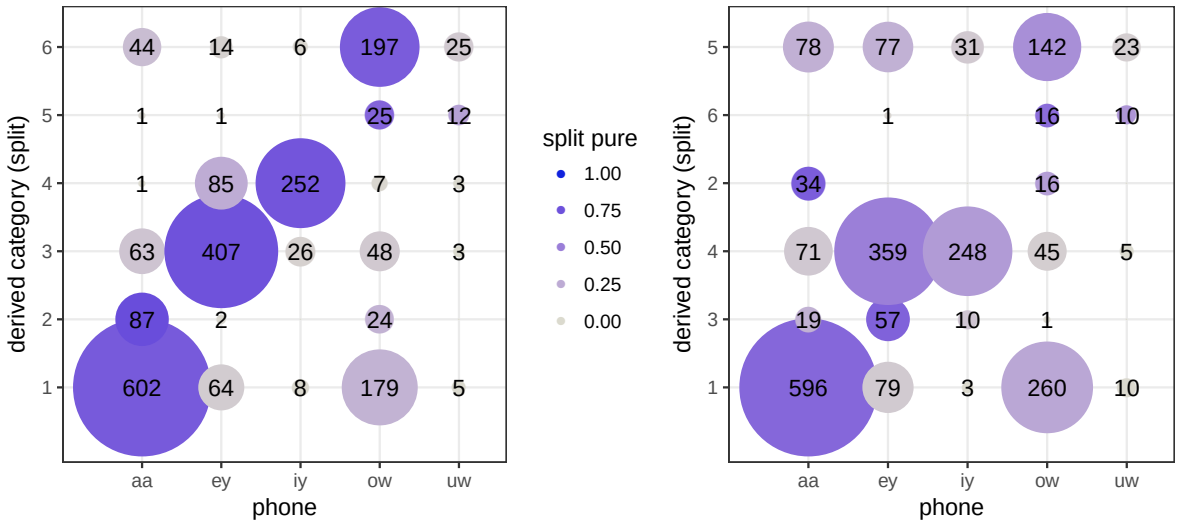

Figure S1. Assignment of vowels to categories based on orthographic words (left panel) and based on tokens (right panel). Conventions are as in Fig 3; this plot includes extension to tokens that were not in frequent words.

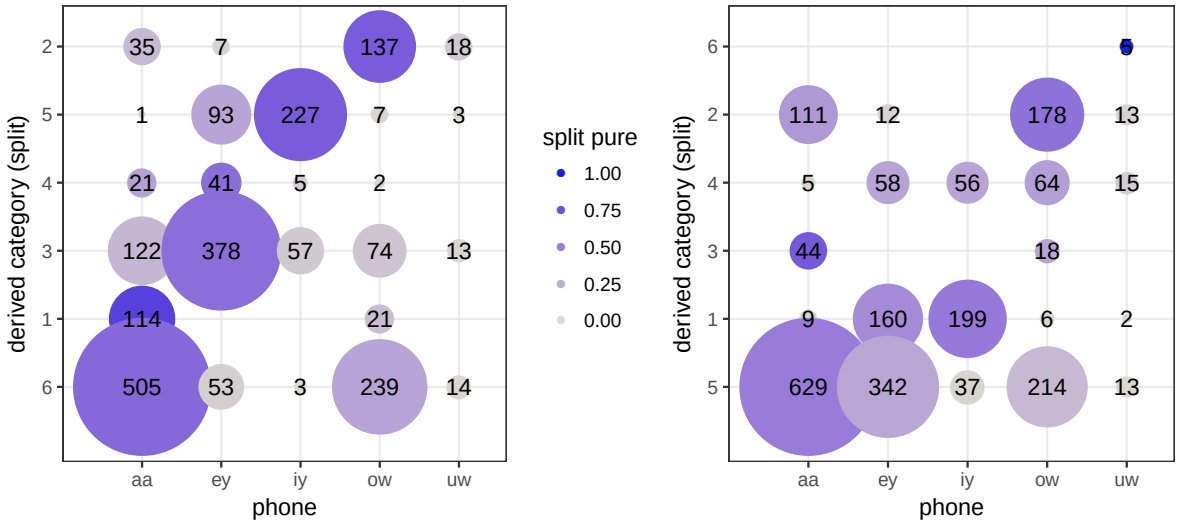

Figure S2. Assignment of vowels to categories that were based on transcribed, vowel-neutralized words (left panel) and the linked tokens-based categories (right panel). Conventions are as in Fig 3; this plot includes extension to tokens that were not in frequent words.

in a category principally devoted to [a], though this is more severe in the tokens analysis (260/480 versus 179/480). The types analysis succeeds in differentiating [i] and [e], where the tokens analysis conspicuously fails.

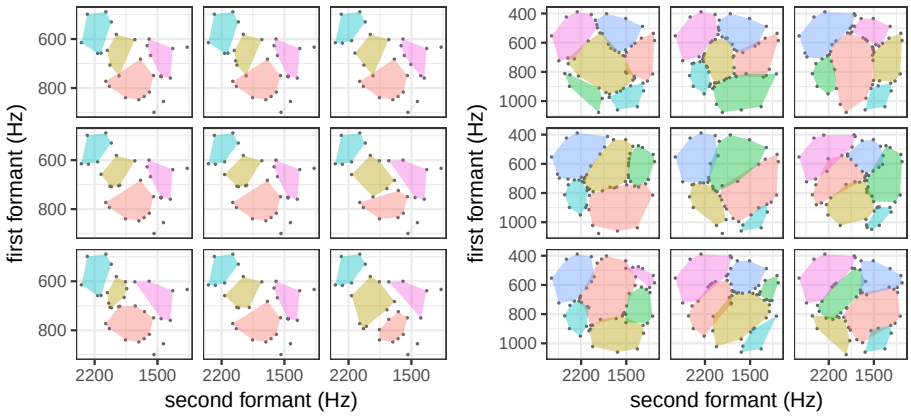

Figure S3. Outlines of categories found in resamplings of the orthographic-types dataset from Analysis 1. The left panel shows the types analysis, the right shows the tokens analysis. Color assignments (online) are arbitrary. Minimum word frequency was set to 5, and number of categories 6.

Inspection of this plot reveals the manifestly inferior match between categories and Spanish phones in the tokens-based analyses than in the types-based analyses. Lexical contexts estimated from consonantal frames yield categories that provide a better basis for extention than categories derived from un-aggregated frequent-word tokens.

Supplementary analysis 2: instability of the tokens-based categories

The similarity spaces derived by unsupervised cluster analysis are never completely nonsensical (each cluster always contains elements that are near one another) but category boundaries and overall morphology tend to be arbitrary when the analysis is seeking clusters over a space that is essentially uniformly populated. The arbitrary nature of the clusters found in the tokens analysis may be seen by repeatedly resampling portions of the dataset (here, leaving out a random half-percent of the data). Nine examples of such resamplings are shown in Figure S3.

These examples show that while the types analysis underwent slight perturbations as a result of tinkering with the precise composition of the dataset, even 0.5% random exclusions exercised a massive influence on the analysis by tokens. This is consistent with the conclusion that the Spanish vowels' formants, as measured over all the instances in the dataset, do not exhibit clearly

separable modes.

Supplementary analysis 3: randomizing formants across words, within vowel

Vowels' phonetic characteristics are expected to vary to some degree according to which word they appear in. Some of this variation is caused by phonetic context effects (e.g., the /a/ in *hot* will normally be less nasalized than the /a/ in *mom*). Additional variation is caused by the fact that word types are not randomly distributed over sentential environments, speakers' emotional states, prosodic contexts, and so on. If such effects are very strong, a given vowel as it is typically realized in a particular word might not resemble that vowel as it is typically realized in another word. In the main text, we tested how effective vowel clustering over word types might be despite this lexical variation. But how much do such context effects matter?

One way to evaluate this is to randomly assign vowel tokens' measurements to words, and then cluster over the means of these (shuffled-data) words. If phonetic context effects are important, this shuffled analysis would be expected to be superior to the true lexical analysis. Here, using the neutralized-vowel dataset of Analysis 2, we performed such a shuffling analysis. We first filtered the dataset to include only vowel tokens of words occurring five times or more. Then we separated the tokens into groups according to their vowel label (all the a's, all the e's, . . .). Within each group, assignments of formant pairs to lexical contexts were randomly shuffled. Finally, mean f1 and f2 values were computed for each word, and cluster analysis was done over these pseudo-word-prototypes. Results of nine randomizations are showin in Supplementary Figures S4 and S5.

These simulations show that without lexical effects on vowel phonetics, but with the grouping benefits provided by tagging vowels with their lexical context, categorization is much more successful. Notably, most cases of confusion between [a] and [o], and between [a] and [ɛ], are absent. This suggests that the tendency of words to present their component sounds in characteristic ways reduces the efficiency of type-based clustering by decreasing the overlap among words that contain (nominally) the same vowel.

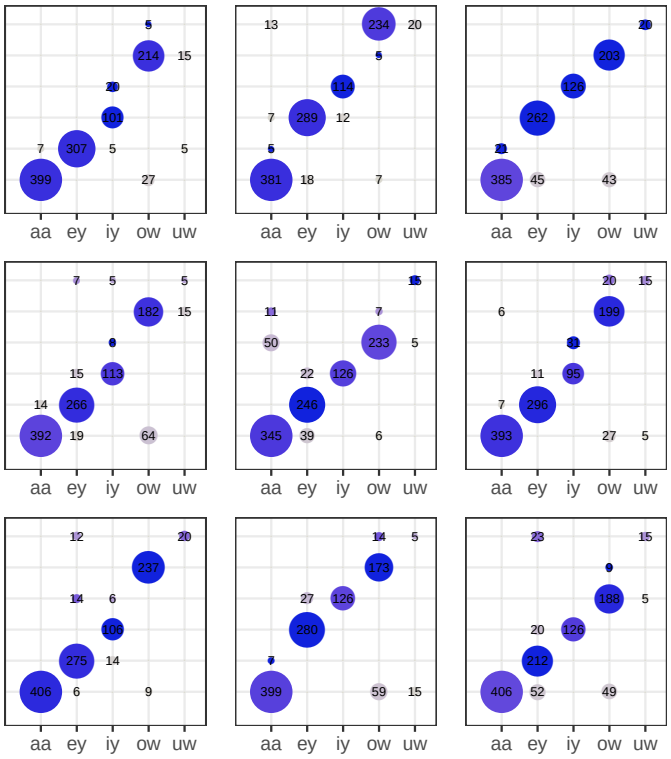

Figure S4. Derived-category uniformity of nine random assignments of formant values to words, within vowel category. Ideal performance would place all tokens of a given vowel (column) into the same derived category (row). These plots may be compared to the left panel of main text Figure 6. Lexical categories were the vowel-neutralized words of Analysis 2. Minimum word frequency was set to 5, and number of categories 6.

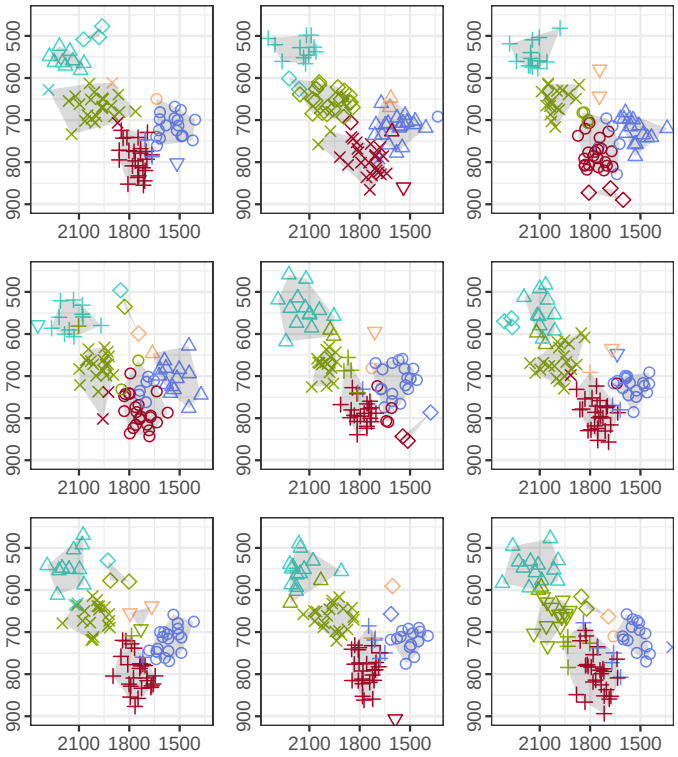

Figure S5. Assignment of vowels to categories based on word types, where formants were randomly assigned to words. Gold-standard categories are given by point color, and derived categories are shown by gray polygons and point shape. Nine randomizations are shown; these are the same randomizations that are displayed in the previous Figure. This set of results may be compared with the left panel of Figure 7 in the main text. Minimum word frequency was set to 5, and number of categories 6.