Model	Dev		Test	
	Exp	1-Best	Exp	1-Best
Basic	0.325	0.379	0.344	0.411
User-Adapted	0.327	0.427	0.344	0.439

Table 2: Quality correlations: basic and user-adapted models.

Feature Removed	QC	
	Expected	1-Best
None	0.522	0.425
Cognate	0.516	0.366*
Context	0.510	0.366*
History	0.499*	0.259*

Table 3: Impact on quality correlation (QC) of removing features from the model. Ablated QC values marked with asterisk* differ significantly from the full-model QC values in the first row ($p < 0.05$, using the test of Preacher (2002)).

5.1 Feature Ablation

To test the usefulness of different features, we trained our model with various feature categories disabled. To speed up experimentation, we sampled 1000 instances from the training set and trained our model on those. The resulting QC values on dev data are shown in Table 3. We see that removing history-based features has the most significant impact on model performance: both QC measures drop relative to the full model. For cognate and context features, we see no significant impact on the expected QC, but a significant drop in the 1-best QC, especially for context features.

5.2 Analysis of User Adaptation

Table 2 shows that the user-specific features significantly improve the *1-best* QC of our model, although the much smaller improvement in *expected* QC is insignificant.

User adaptation allows us to discern different styles of incidental comprehension. A user-adapted model makes fine-grained predictions that could help to construct better macaronic sentences for a given user. Each user who completed at least 10 HITs has their user-specific weight vector shown as a row in Figure 6. Recall that the user-specific weights are not used in isolation, but are *added* to backoff weights shared by all users.

These user-specific weight vectors cluster into four groups. Furthermore, the average points per HIT differ by cluster (significantly between each cluster pair), reflecting the success of different strategies.[8] Users in group (a) employ a generalist strategy for incidental comprehension. They pay typical or greater-than-typical attention to all features of the current HIT, but many of them have diminished memory for vocabulary learned during past HITs (the hist+ feature). Users in group (b) seem to use the opposite strategy, deriving their success from retaining common vocabulary across HITs (hist+) and falling back on orthography for new words. Group (c) users, who earned the most points per HIT, appear to make heavy use of context and pronunciation features *together* with hist+. We also see that pronunciation similarity seems to be a stronger feature for group (c) users, in contrast to the more superficial orthographic similarity. Group (d), which earned the fewest points per HIT, appears to be an "extreme" version of group (b): these users pay unusually little attention to any model features other than orthographic similarity and hist+. (More precisely, the model finds group (d)'s guesses harder to predict on the basis of the available features, and so gives a more uniform distribution over V^e.)

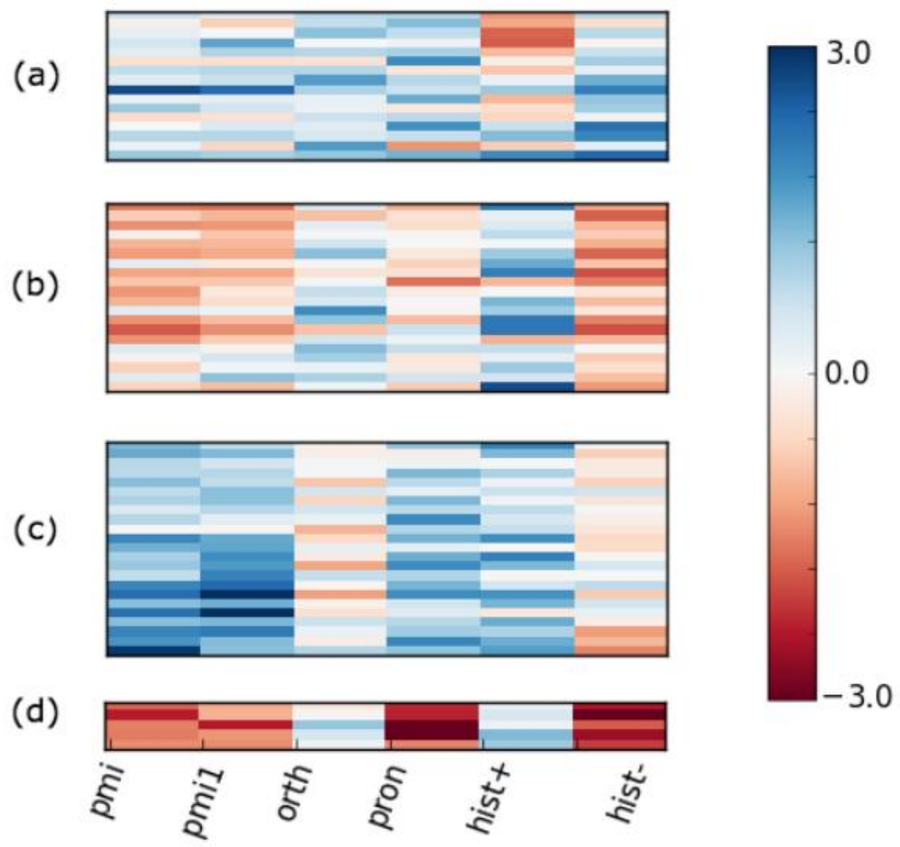

Figure 6: The user-specific weight vectors, clustered into groups. Average points per HIT for the HITs completed by each group: (a) 45, (b) 48, (c) 50 and (d) 42.

6 Future Improvements to the Model

Our model's feature set (section 4.1) could clearly be refined and extended. Indeed, in a separate paper (Knowles et al., 2016), we use a more tightly controlled experimental design to explore some simple feature variants. A cheap way to vet features would be to test whether they help on the task of modeling reference translations, which are

[8] Recall that in our data collection process, we award points for each HIT (section 3.4). While the points were designed more as a reward than as an evaluation of learner success, a higher score does reflect more guesses that were correct or close, while a lower score indicates that some words were never guessed before the system revealed them as clues.