

Directory to Supplementary Materials

for Yang Wang and Bruce Hayes, “Learning phonological underlying representations: the role of abstractness”

1. Appendices to text

- 1.1: Description of the morpheme parser
- 1.2: Formulae for EM learning

2. Illustration of EM with Pseudo-German (§5.7)

- You can run the iterative algorithm yourself if you have access to the Excel Solver.
- Excel spreadsheet: EMDemo_PseudoGerman.xlsx

3. Material specific to the language simulations (§6)

3.1 Code

UR learner	em.py
UR generator and surface GEN	gen.py
Interpolator for phonetically intermediate sounds	interpolation.py
Store allomorph sets and features for a language	lang.py
A demo with Pseudo-German	EMDemo_Pseudogerman.ipynb
Relevant files for the demo with Pseudo-German	./PseudoGermanDemo

3.2 Training data

Pseudo-German
Catalan
Tangale
Seediq
Paka20

3.3 Morpheme parses

- Program output traces showing how the segmentation of words into morphemes was searched for.

3.4 Lists of winning URs and tableaux

For each language X:

- XFullInputTableau.csv: input to the EM learner
- X_goldURs.csv: a list of linguist-preferred URs for each morpheme present

- UR probability lists:
 - XFinalURProbabilities.csv.
 - These list all URs with the probabilities assigned to them by the model.
- Final output tableaux: XTableauWinningURsOnly.csv.
 - For brevity, these files omit all URs with probability less than 0.01. However, all surface candidates for each included UR are present.
 - The tableaux also give the final values of the constraint weights.
 - For all candidates, we give both the UR and the SR. In the model of grammar assumed, a form is generated first by selecting a UR according to the UR probabilities, then deriving an SR from this UR using the MaxEnt phonology. Hence there are two relevant probability values, both listed in the tableaux:
 - $P(\text{SR}|\text{UR})$: the probability that from a particular UR, the SR given will be derived within the phonology.
 - $P(\text{SR}, \text{UR})$: the probability that a particular UR-SR pair will be output by the system as a whole

Since in most cases the model learns a unique UR, $P(\text{SR}|\text{UR})$ and $P(\text{SR}, \text{UR})$ usually come out the same.
- Simulations included:
 - Pseudo-German with KK-C
 - Catalan with KK-C
 - Catalan with KK-E
 - Catalan with KK-Z (fails)
 - Catalan with token variation
 - Tangale with KK-C
 - Tangale with KK-Z (fails)
 - Seediq with KK-C (fails)
 - Seediq with KK-D
 - Seediq Restructuring with KK-C
 - Seediq Restructuring with KK-D (fails to learn the vowel matching trend)
 - Paka-20 with KK-C (fails)
 - Paka-20 with KK-E

3.5 *Graphs for learning trajectories*

- Provided only for a few representative cases, these show how the following values change during the course of learning:
 - probabilities for representative morphemes
 - constraint weights
 - log-likelihood

3.6 *Wug test*

- Provided the learner with a new paradigm ([prime, prime-s, primer-ə, primer-ə-s]) Catalan as wug testing.

- Program output traces for the parsing
- CatalanWug_KK-C_FullInputTableau.csv: input to the EM learner
- CatalanWug_goldURs.csv
- CatalanWug_KK-C_Tableau_WinningURsOnly.csv
- CatalanWug_KK-C_FinalURProbabilities.csv
- The graph for the learning trajectory of the wug paradigm

4. Learning highly abstract URs in Klamath (hand-provided UR candidates)

This addresses work by O’Hara (2017). We show that if we hand-include URs with the abstract vowel /e/ in the set of UR candidates, and deploy constraints along the lines proposed by O’Hara,¹ then the correct URs are selected, along with appropriate constraint weights.

We give files with the same format as above, for:

- A failed learning simulation using KK-C (O’Hara’s proposed UR is not in the hypothesis space; the same would hold for KK-D and KK-E).
- A successful learning simulation, in which a single abstract UR candidate /ce:we/, is hand-included in the UR candidate set. All other computations are as described in our paper.

¹ We fleshed out O’Hara’s system thus: (a) MAX and DEP include versions specific to [i], [a] and [e], to guarantee that it is [a] that is inserted but /e/ that is deleted; (b) we render O’Hara’s cover constraint PHONOTACTIC with specific Markedness constraints; i.e. *[-sonorant]CC and *[+syl][+syl]. Note that we did not include the abstract vowel /ɪ/ as a candidate UR, thus, omitting relevant constraints (IDENT[ATR], *ɪ).