

A learning algorithm for phonemicization in Stratal MaxEnt OT

Yang Wang¹ & Bruce Hayes²

¹ Dept. of Linguistics, University of Utah
yangx.wang@utah.edu

² Dept. of Linguistics, UCLA
bhayes@humnet.ucla.edu



DEPARTMENT OF LINGUISTICS
THE UNIVERSITY OF UTAH



Research question

Given a **list of words** (no meaning, no paradigms), how can a learner determine:

- ❖ which surface segments are allophones of the same phoneme
- ❖ the phonological grammar for allophony

We offer a **computational system for phonemicization**

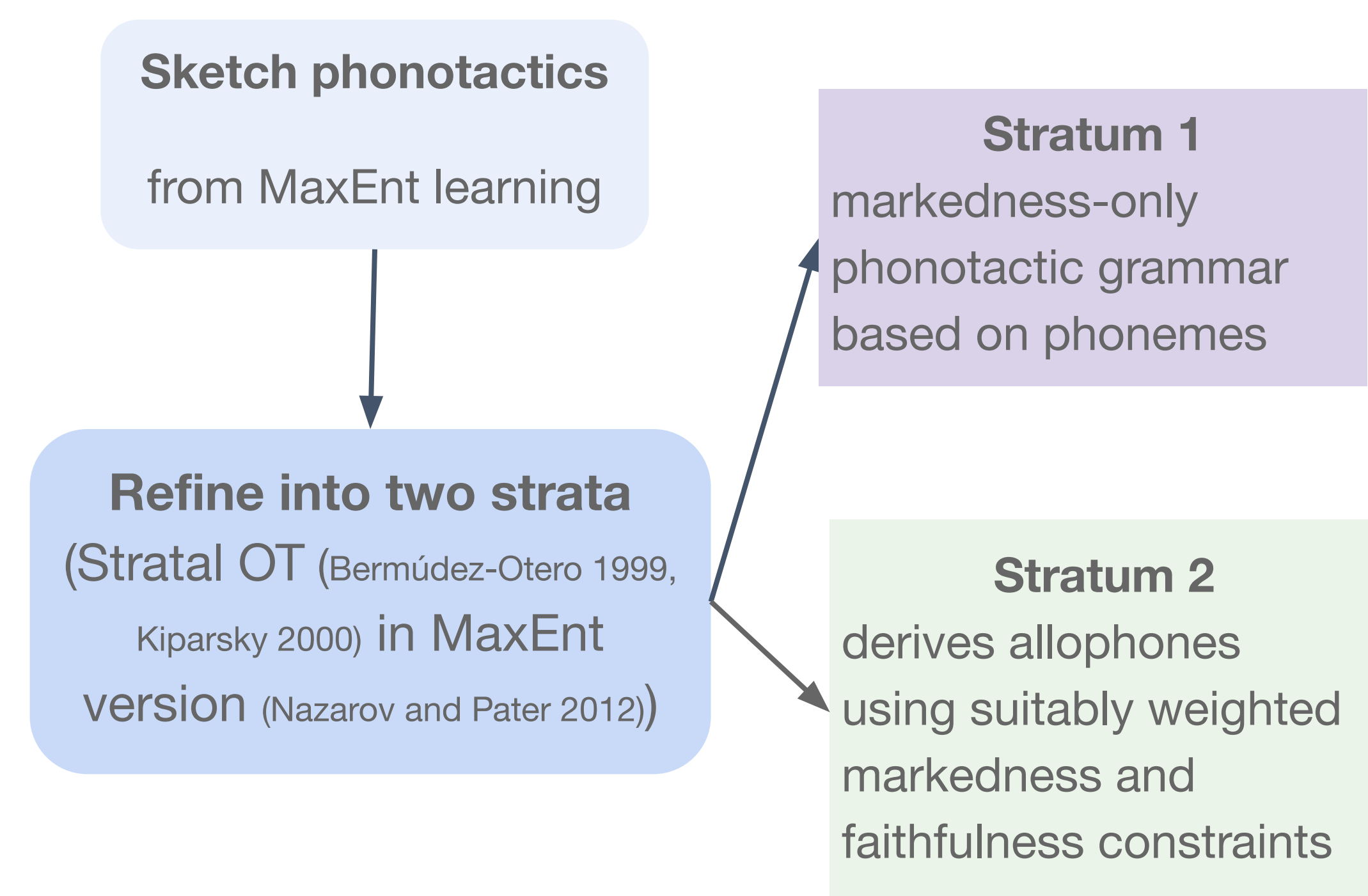
- inspired by earlier work: Peperkamp et al. (2006), Rasin & Katzir (2016), Wilson & Gallagher (2018), Gouskova (2024)

Goals

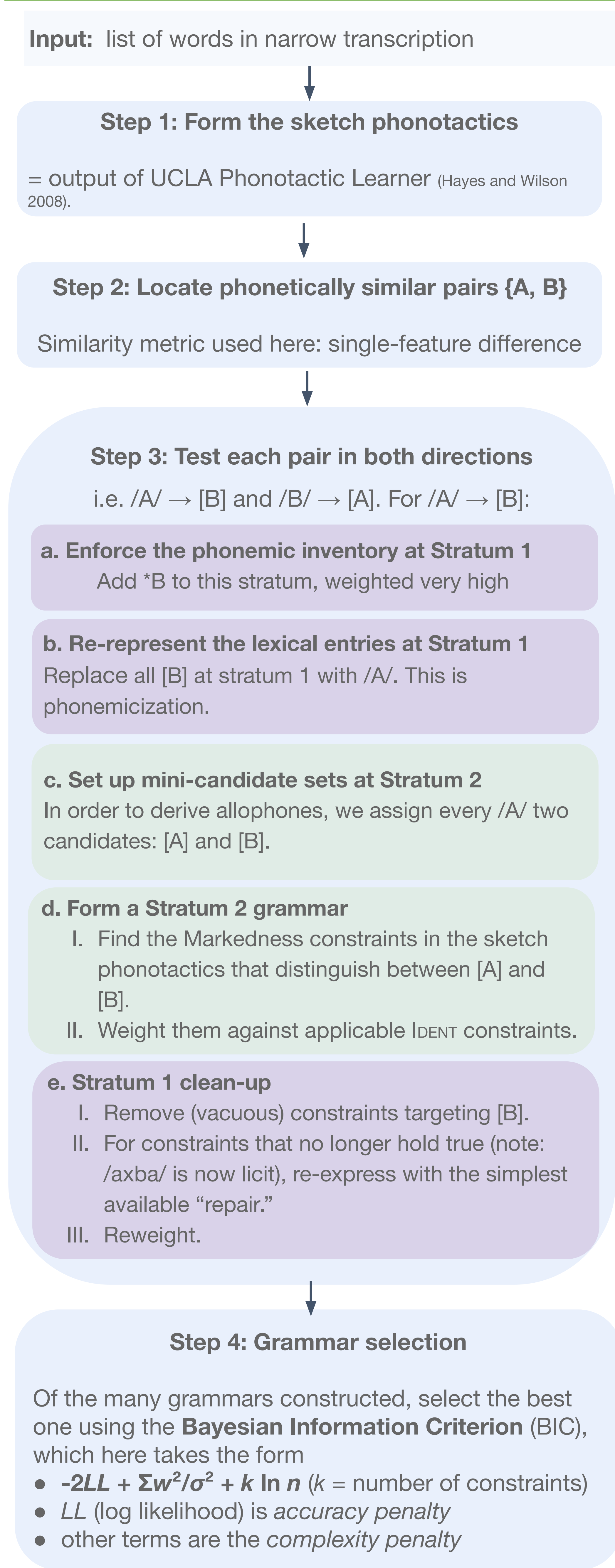
Pattern	Criterion of success
True allophony Pseudo-Russian /x/ → ɣ / __ [-son,+voice]	1. No /ɣ/ in UR 2. Find key markedness constraint: *[-voice][-son,+voice] Outranks IDENT(voice)
Wrong direction Pseudo-Russian with /ɣ/ as the underlying phoneme.	Reject this
Pairs that are phonemic Pseudo Russian /p/, /b/	Maintain as separate phonemes, rendered faithfully
Parallel allophones defined on natural classes; Pseudo Quechua /i, u/ → [e, o] /adjacent to uvulars	Find the general analysis

How our model works

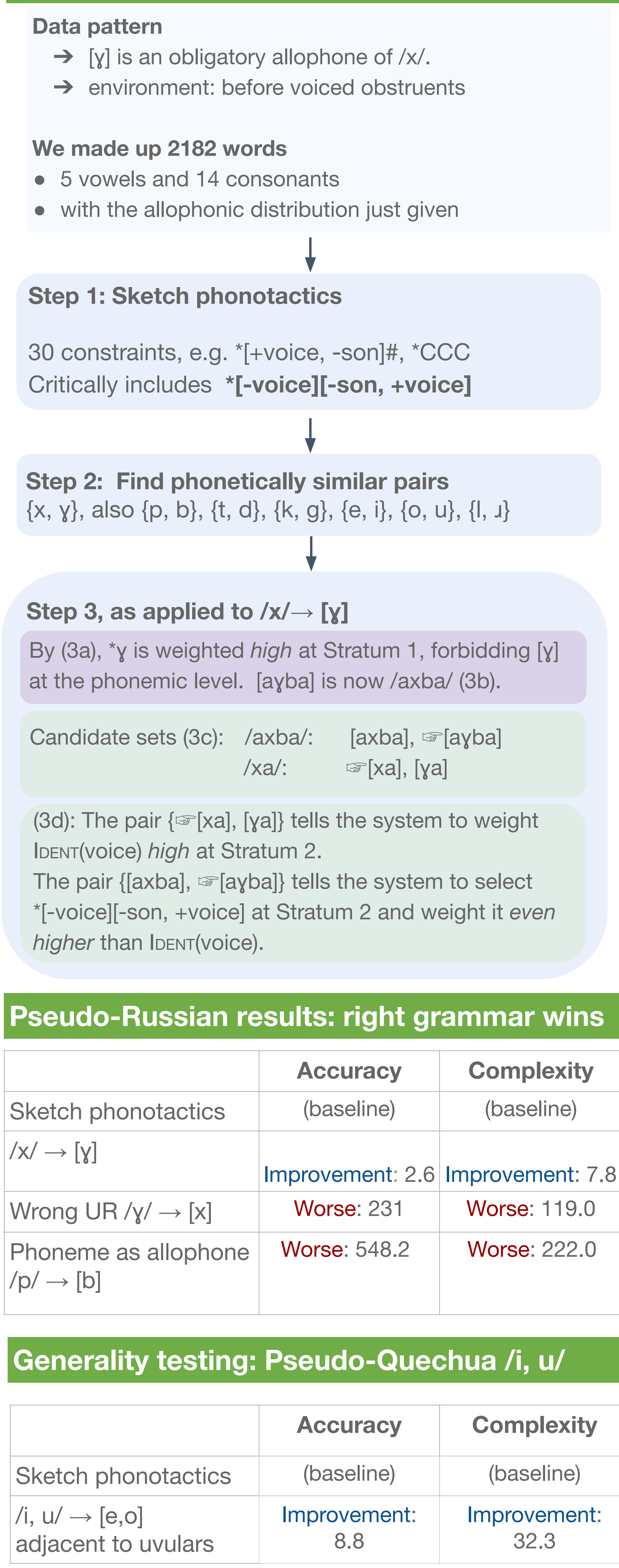
Phonemicization: *refinement of phonotactic learning*
(e.g. Hayes and Wilson, 2008; Wilson and Gallagher, 2018)



The model in more detail



The model applied to Pseudo-Russian



Pseudo-Russian results: right grammar wins

	Accuracy	Complexity
Sketch phonotactics	(baseline)	(baseline)
/x/ → [ɣ]	Improvement: 2.6	Improvement: 7.8
Wrong UR /ɣ/ → [x]	Worse: 231	Worse: 119.0
Phoneme as allophone /p/ → [b]	Worse: 548.2	Worse: 222.0

Generality testing: Pseudo-Quechua /i, u/

	Accuracy	Complexity
Sketch phonotactics	(baseline)	(baseline)
/i, u/ → [e, o] adjacent to uvulars	Improvement: 8.8	Improvement: 32.3

Simplicity is essential

Experiment: try to prove that English [ð] is allophone of /ʒ/.
Method:
(1) Remove pairs like *bathe* ~ *beige* from the training set.
(2) Allow for additional constraint discovery at Stratum 2, e.g.

*ʒ / aɪ ____ # ([ð] in *lithe*, *blithe*, *writhe*, *scythe*, *tithe*, no [ʒ])

✓ With enough ad hoc constraints, the grammar becomes **more accurate** than the correct two-phoneme analysis.
ΔLL = +85.8 improvement
✗ But **simplicity penalty** is huge: worse by 227.3

Thus, as in earlier work, phonemic analysis depends on a blend of accuracy and simplicity.

Broader context: phonemicization solves problems in OT

- ✓ Wilson and Gallagher's "patchwork problem"
- ✓ Gouskova (2024): the Russian [ɣ] allophone cannot be derived in single-stratum OT.
- ✓ Trims an otherwise-vast search space for learning alternations (Wang and Hayes, in press)
- ✓ Matches experimental work suggesting *phonemes* are cognitive categories (e.g., Gudschinsky et al. 1970, Jaeger 1980, Derwing et al., 1986; see review in Kazanina et al., 2018).

These are problems for OT because its classical version, with Rich Base and Lexicon Optimization, rejects phonemicization.

Where to go from here?

- Optional allophones** necessitate more powerful learning systems
- allophony is learned in tandem with the lexicon
 - Key challenge: what distinguishes optional allophony (*cat* [kæɪ] ~ [kæt]) from lexical doublets (*pecan* [pə'kən] ~ [pə'kæn])?
- Dual-source allophones.** Bloch (1941): [r] is derived
- from /ɹ / after [θ] (*three* [θri])
 - from /t/ intervocalically (*butter* ['bʌrə]).

Thanks to Gillian Gallagher for providing the Quechua data, and to the anonymous reviewers for feedback. Code, data and model outputs will be released with the AMP proceedings paper.