UCLA

Learning underlying representations: The role of Expectation-Maximization and the KK Hierarchy

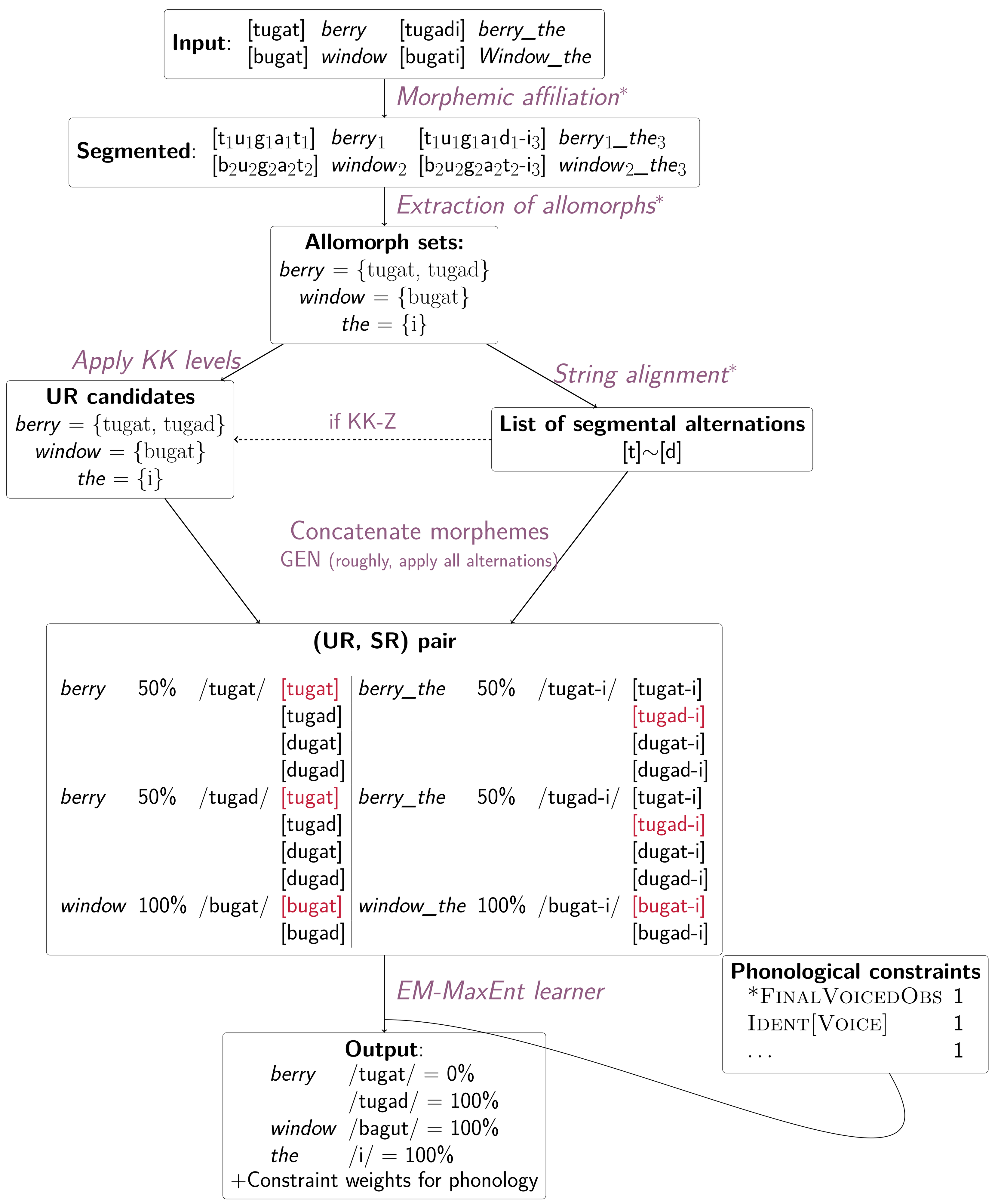
Yang Wang and Bruce Hayes
yangwangx@g.ucla.edu, bhayes@humnet.ucla.edu; Department of Linguistics, UCLA

1. Setting

- We return to the classical **UR-learning problem** (e.g. Tesar et al., 2003; Jarosz, 2006; Merchant, 2008; Pater et al., 2012; Tesar, 2014; Cotterell et al., 2015; Jarosz, 2015; Rasin and Katzir, 2016; Rasin et al., 2021; O'Hara, 2017; Nelson, 2019; Hua et al. 2021, Hua and Jardine 2021, Tan 2022, Ellis et al. 2022)
- Goals
  - **Scale up** learning simulations, to the size of first-year problem sets.
  - Readdress the **Abstractness Controversy** in computational terms

2. The proposed learning system: Architecture

Any procedure marked with * is omitted. For details, see the full paper. Feel free to ask!



Download draft paper (still in progress) and references

References can be obtained by looking at the draft paper, downloadable by scanning the QR code.

Acknowledgement: Many thanks to Tim Hunter, Claire Moore-Cantwell, Colin Wilson, the audience of the UCLA Phonology Seminar and the anonymous reviewers for their comments, feedback and insights. Special thanks to Jennifer Kuo for providing the Seediq data.

3. Example Language: Tangale (Kidda 1985)

Target UR	*Noun*	*'the N'* /-i/	*'your N'* /-go/	*Gloss*
/tugad/	tugat	tugad-i	tugad-go	'berry'
/bugat/	bugat	bugat-i	bugat-ko	'window'
/wudo/	wudo	wud-i	wud-go	'tooth'
/lutu/	lutu	lut-i	lut-ko	'bag'

(+ 8 more stems, 2 more paradigm slots)

Word final obstruent devoicing	/tugad/ → [tugat]; cf. [tugad-i] Compare: [bugat] ∼ [bugat-i] with invariant [t]
Stem-final (short) vowel syncope before any suffix	/wudo-i/ → [wud-i] cf. [wudo]
Progressive voicing assimilation	/bugat-go/ → [bugat-ko] /lutu-go/ → lutgo → [lut-ko]

4. Constructing the UR candidates: the KK hierarchy

KK-hierarchy : Kenstowicz and Kisseberth (1977, ch 1)

- Key insight: the candidate URs can be projected from the set of surface allomorphs.
- Higher levels → a larger set of UR candidates

Level	*Meaning*	*e.g. Tangale 'berry'*
B"	Single surface base hypothesis Albright 2002	If base = isolation form,*/tugat/
C	UR is some surface allomorph	✓ /tugad/,*/tugat/
D & E	(omitted here)	
Z	Apply all attested alternations to all attested allomorphs	/tugad/, /tugat/, /tgad/, /dukat/ +28 others

- The learning task: select the right UR among the candidates, and learn phonology simultaneously.

5. Learning Tangale phonology: Modeling

- Data: 60 word forms, examplified as before
- Initialization: all UR candidates for each morpheme are **equiprobable**, all weights initialized as **1** for each weight optimization procedure
- Convergence criterion: $\Delta(\text{log likelihood}) < 10^{-5}$

6. Results

- Constraint set with weights fitted under both models:

Constraint	**KK-C**	**KK-Z**
*V]X	65.9	67.6
*FinalVoicedObs	35.9	17.9
Agree[Voice]	45.1	48.4
Ident[Voice]	4.4	5.26
Ident[Voice]$_{STEM}$	17.7	22.4
Max	24.8	27.31
Dep	36.7	0
...	...	...
Log-likelihood	-5.2×10^{-4}	-4.5

Successful learning under KK-C

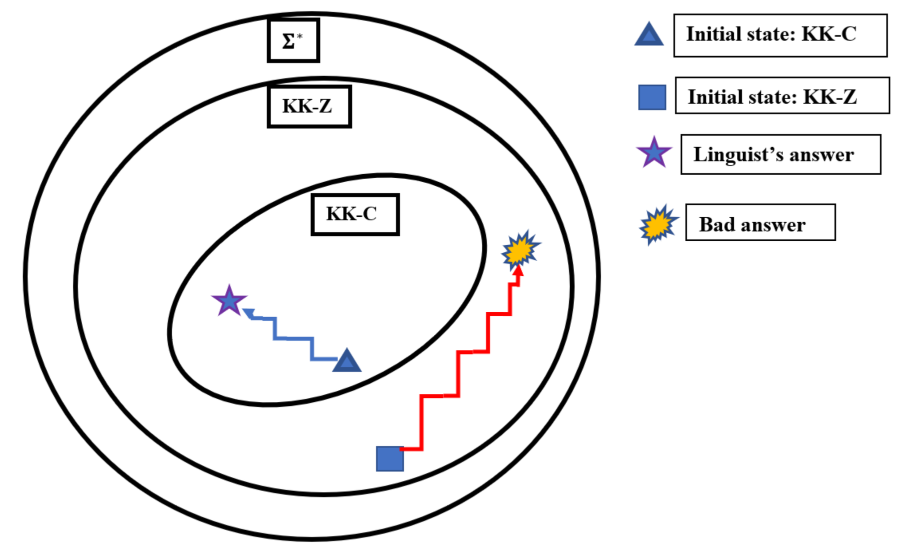
- the linguist's answer
- ... consensus UR's
- ... near-perfect output accuracy.

Failed learning with KK-Z

- ... in URs: learnt erroneous */tgat/, */tugat/
- ... thus in outputs:
  P('your berry', *[tugatko])=25% ([tugadgo])
  P('berry', *[tugado])=75% ([tugat])

7. Diagnosis of KK-Z failure

- The system gets trapped in a **local maximum**



- Partial diagnosis: the KK-Z UR candidate /tgat/ → [tugat] forces the system to consider an epenthesis grammar → low weights to Dep → terminal confusion
- This doesn't happen when the initial state has a more reasonable candidate set.

8. Summary & Conclusions

- **The need to scale up**: We worked on scaling up to problem-set size.
  Other languages: Catalan, Lamba, Seediq, English plurals, Indonesian
- **Exploring abstractness**
  - The abstract debate long ago (1970's): whether abstract phonology is learnable.
  - Computational phonology can address such claims concretely: try the same learning system with different degrees of abstractness permitted.

  Too-abstract UR → large search space → fatal local maxima

  - All of this is tentative, pending additional research :
  1. in learning algorithms
  2. in psycholinguistics: what URs are actually learnt

Appendix 1. the Objective

The log-likelihood of the observed data (cf. O'Hara 2017)

$$ln(P(D\,|\,\theta, W)) = \sum_{(s,\omega)\in D} f(s,\omega)\, ln\left(\sum_{u_\omega} P(s \mid u_\omega, W) P(u_\omega\,|\,\omega,\theta)\right) \tag{1}$$

where

$$p(s\,|\,u_\omega, W) = \frac{exp(-\sum_i W_i C_i(u_\omega\,,\,s))}{\sum_{s'\in\text{Gen}(u_\omega)} exp(-\sum_i W_i C_i(u_\omega\,,\,s'))} \tag{2}$$

$$P(u_\omega\,|\,\omega,\theta) = \prod_{i=1}^{n} \theta_{(\mu_i,u_i)} \text{ if } u_\omega = u_1u_2u_3\ldots u_n \text{ and } \omega = \mu_1\mu_2\mu_3\ldots\mu_n \tag{3}$$

Appendix 2. Searching with Expectation Maximization

Given D = $\{(s,\omega)\}$ (SR, Word) pairs, the learner conducts an iterative search, maximizing the likelihood.

Input: D, maxIteration, θ^0, W^0
Output: θ^f, W^f

Function *EM-MaxEnt Learner*:
- iter = 0
- **Repeat until converge**
  - E = E-step(W^t, θ^t)
  - W^{t+1} = M-step$_W$(E)
  - E' = E-step(W^{t+1}, θ^t)
  - θ^{t+1} = M-step$_\theta$(E')
  - *iter +=1*
- **while** *iter ≤maxIteration*

Function *M-step$_W$ (E)*:
- W = $\underset{W}{\text{argmax}} \sum_{(s,\,\omega)} \sum_{u_\omega} E(u_\omega, s,\,\omega)\, ln(p(s|u_\omega, W))$
- **return** W

Function *E-step(W, θ)*:
- **for** $(s,\omega) \in D$ **do**
  - **for** $u_\omega \in U_\omega$ **do**
    - $P_{u_\omega} = \dfrac{P(s \mid u_\omega; W)\; P(u_\omega|\,\omega,\theta)}{\sum_{u'\in U_\omega} P(s \mid u'; W)\; P(u'|\,\omega,\theta)}$
    - E(u_ω , s, ω) = $P_{u_\omega} \cdot f(s,\omega)$
- **return** E

Function *M-step$_\theta$ (E)*:
- $M \leftarrow$ set of word forms containing μ
- $\theta_{(\mu,u)} = \dfrac{\sum_{(s,\,\omega)\in M}\sum_{u_\omega} E(u_\omega,s,\omega)}{\sum_{u'}\sum_{(s,\,\omega)\in M}\sum_{u'_\omega} E(u'_\omega\,,s,\omega)}$
- **return** θ