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

1. Introduction

- We propose a new learning system for underlying representations (UR’s).
- The system operates within mainstream phonology, using (the MaxEnt variant of)
  Optimality Theory.
- It inputs a labeled phonological paradigm, and outputs the UR’s plus the constraint
  weights of a MaxEnt phonology.

2. We aren’t the first to try to learn UR’s

- Learning of UR’s is an established challenge for phonological theory: Jarosz (2006),
  Nelson (2019).
- So let’s start with a broad overview, emphasizing what is new.

  BROAD OVERVIEW

3. Bringing together an ensemble of ideas

- There are three key ideas.
  - Expectation-Maximization
  - Early allomorph discovery
  - KK Hierarchy

4. Expectation-Maximization

- **Expectation-Maximization** (EM, Dempster et al. 1977) classical algorithm used for
  finding “hidden structures” such as URs.
- We offer a distinct and more effective version of EM.

5. Segmenting words into morphemes before figuring out phonology

- Contrary to popular belief, you do not have to learn both phonology and morpheme
  segmentation all at once.
- Parsing words into their morphemes stage-wise helps a lot with search space.
6. The KK hierarchy
   - This is a seminal taxonomy of how much surface and underlying forms may differ, laid out by Kenstowicz and Kisseberth (1977).
   - We use it as a way of generating contrained search spaces for possible UR’s.

BACKGROUND SECTIONS
BACKGROUND I: THE HIDDEN STRUCTURE PROBLEM

7. The hidden structure problem
   - **Hidden structure** = anything that is part of the linguistic representation which cannot be directly inferred from the phonetic data (Tesar and Smolinsky 2000)
   - Hidden structure is found when there are one-to-many relations between phonetic form and linguistic representation.
   - Examples:
     - Metrical feet: (bəˈnæ)nə vs. bə(ˈnænə)
     - All syntactic and morphological structure
     - Underlying representations: with phonological neutralization, a surface representation has multiple possible URs, which are hidden.

8. Why hidden structure makes learning hard (Tesar and Smolinsky 2000)
   - Learning the rule system interacts with learning the hidden representation.
   - E.g. the alternation [gris] ~ [griz-ə] ‘grey-masc./fem.’ in Catalan might be:
     - /gris/ with Intervocalic Obstruent Voicing
     - /griz/ with Final Devoicing
   - The challenge is to learn both phonology and UR’s at the same time.
   - In general, we need to avoid a bad outcome:
     - *Wrong decision made in Domain X shuts out the possibility of making the right decision in Domain Y.

9. For phonology, we have double hidden structure (Nelson 2019)
   - The words are divided into morphemes, whose segmental content is unknown prior to learning.
     - Possible morphemic affiliations for Catalan [grizə] ‘big-fem.’
       - $g_1\; r_1\; i_1\; z_1\; \ddot{a}_2$ big₁ fem₂ (correct)
       - $g_1\; r_1\; i_1\; z_2\; \ddot{a}_2$ big₁ fem₂ (-ə suffix)
       - $g_2\; r_1\; i_1\; z_1\; \ddot{a}_2$ big₁ fem₂ (g- -ə circumfix)
       - Etc.
So, we’re combining choices about morpheme division, UR’s, and phonology — this can be a huge search space.

10. **Our basic method**

- Expectation-Maximization ((4)) is designed to find hidden structure and can do a lot of the work.
- But we will carefully avoid overwhelming it, by limiting the size of the search space it must navigate.

**BACKGROUND II: ARCHITECTURE OF OUR MODEL**

11. **Preliminary: a small language to illustrate**

- We adopt a tiny subset of Catalan phonology (Wheeler 2005), looking ahead to a larger Catalan model later on.
- For now, two paradigms suffice, showing the classic pattern of Final Obstruent Devoicing:

  \[
  \begin{array}{l}
  \text{[gris]} \\
  \text{[griz-ə]} \\
  \text{[gros]} \\
  \text{[gros-ə]}
  \end{array}
  \begin{array}{l}
  \text{‘grey’} \\
  \text{‘grey-feminine’} \\
  \text{‘big’} \\
  \text{‘big-feminine’}
  \end{array}
  \]

  ➢ Catalan tolerates no final voiced obstruents in surface forms.
  ➢ The form [grosə] and others demonstrates that there is no intervocalic voicing.

- This is identical in all relevant respects pattern to the toy example used by Pater et al. (2012).

12. **The course of learning**

- Starting point: a batch of paradigms, labeled for what morphemes are present — but not for what segments embody them (cf. Nelson 2019):

  \[
  \begin{array}{l}
  \text{gros} \\
  \text{grosə} \\
  \text{gris} \\
  \text{grizə}
  \end{array}
  \begin{array}{l}
  \text{big} \\
  \text{big fem.} \\
  \text{grey} \\
  \text{grey fsg}
  \end{array}
  \]

- A constraint-based, MaxEnt system described in Appendix A and Hayes (2018) figures out what segments belong to what morphemes; this is indicated with coindexation:
Collating cosubscripted segments, we obtain **allomorph sets** for each morpheme:

- 'big' { gros }
- 'grey' { gris, griz }
- 'fem.' { ə }

We do two things with these allomorph sets.

- First, we use them to find **what segments alternate**.
  - This can be done **string-aligning** each pair of allomorphs, following phonetic similarity:
    
    | g | r | i | z |
    |---|---|---|---|
    | g | r | i | s |

  - There are well-understood ways to do this (Kruskal 1999, Bailey and Hahn 2001, Wilson and Obdeyn 2009, White 2017)
  - We learn here that in Catalan [z] alternates with [s].

- We also the string-aligned allomorph sets and project from them a set of possible UR’s.
  - For this, we need the **KK hierarchy**.

**BACKGROUND III: THE KK HIERARCHY**

13. **Source**

- Chapter 1 of Kenstowicz and Kisseberth (1977) *Topics in Phonological Theory*

14. **What it is**

- A series of formal criteria, ranging from very restrictive to completely unrestrictive, about how much UR’s can diverge from surface forms.

15. **What we need from the KK hierarchy**

- The original hierarchy is very broad (ultra-concrete to ultra-abstract) and we use only the middle region.
- In this region, we construe the KK hierarchy as a means of projecting **possible UR’s from the string-aligned allomorph sets**.
- In what follows, we cover a subset of the hierarchy in ascending order (ever more ambitious UR’s).
16. KK-B’’: Single-Surface-Base Hypothesis

- The UR of the stems of a language is always chosen from the same slot in the paradigm.
- Albright (2005, 2008, 2010) has produced a substantial body of work aimed at establishing this as a tenable empirical hypothesis.
  - One of his examples: the slot chosen for UR’s of Yiddish present-tense verbs (Albright 2010) is the 1st sg.
- In Albright’s rendition, the slot that is chosen is the most informative.
  - Catalan case ((11)): clearly, we must use the feminine (/griz/-ə), since it preserves the underlying voicing distinction.
- If the Single-Surface-Base Hypothesis is correct, then search space issues for UR become trivial — try the contexts one by one and see what works best.
  - This doesn’t need EM.
  - But sometimes our EM model, which values informativeness, does find the Albrightian solution.
- The Single-Surface-Base Hypothesis, while intriguing, is not a consensus position in the field — it raises hard questions about how to derive paradigms from such a restricted starting point.

17. Next level up: KK-C, “Pick and Choose”

- Source: KK (1977:32)
- The UR is always identical to some paradigm slot, but not necessarily the same slot across paradigms.
- Why? For different stems, a different slot in the paradigm might be most informative.

18. An example that needs “Pick and Choose”: Conservative Boston English

- For /r/-final and V-final verb stems, you must choose the pre-vocalic-suffix allomorph, since R-Dropping can neutralize the distinction in the isolation form:

\[
\begin{array}{c|c|c}
\text{/sɔə/} & \text{soar} & [ˈsɔ] \\
\text{/sɔr/} & \text{saw} & [ˈsɔ] \\
\end{array}
\]

Caution: in Innovating Boston English, this form is pronounced [sɔɹɪŋ], suggesting the UR saw = /sɔ/ with a process of [ɹ] Epenthesis. To make the argument valid, we need to find speakers (elderly, highly educated? Beacon Hill?) who speak like the British linguist John Wells (1982:page xxx) with respect to /ɹ/ Dropping but have Tapping in their idiolects like other North Americans.
• For /t/-final and /d/-final verb stems, you must choose the isolation allomorph for the UR, since Tapping can neutralize the distinction in the affixed forms:

```
/bʌt/  butt  [ˈbʌt]  butting  [ˈbʌrɪŋ]
/bʌd/  bud   [ˈbʌd]  budding [ˈbʌrɪŋ]
```

19. KK-D: “cobbling”

• Source: KK (1977:32)
• Every segment of a UR is identical to a segment of some allomorph in the paradigm, but the segments of the UR need not ever occur all together in any particular allomorph.
• Thus we “cobble the UR together” from various allomorphs.

20. An example of cobbling: Seediq (Kuo 2020)

• Austronesian, Taiwan
• Phonology:
  ➢ Penultimate stress
  ➢ Pretonic vowels all reduce to [u].
  ➢ Postonic mid vowels (/e/, /o/) also reduce to [u].
• So you need to have both a suffixed and unsuffixed allomorph to obtain the classical UR.

```
```

• Kuo actually argues that cobbling is wrong, and pursues an analysis with the single-surface-base hypothesis.
• (Apology: all our Seediq “data” are schematic and faked, for now. We will fix.)

21. Pick-and-choose will not suffice to obtain classical UR’s for Seediq

<table>
<thead>
<tr>
<th>Stem</th>
<th>Pick-and-choose candidates</th>
<th>What is needed</th>
</tr>
</thead>
<tbody>
<tr>
<td>‘smoke’</td>
<td>/hegut/, /huget/</td>
<td>/heget/</td>
</tr>
</tbody>
</table>

22. Projecting the set of possible UR’s in Seediq using Cobbling

• smoke, with both allomorphs and string alignment:

\[2\text{ Caution: data simplified; we left out the alternation } t \rightarrow \text{̊s} / \text{ɪs} / \_ \text{word} \]
Mix and match: choose all possible combinations, based on the alternations.

/ hegut/
/ heget/ This one ultimately proves to be correct.
/ hugut/
/ huget/

23. KK-D': “hy-per-cobbling”

This will be presented below.

24. Summarizing the KK hierarchy

For us, it projects a candidate set for UR’s from a string-aligned allomorph set.

Higher levels on the KK hierarchy project larger candidate sets.

BACKGROUND V: PHONOLOGY

25. We use MaxEnt Optimality Theory

MaxEnt: Goldwater and Johnson (2003)
Optimality Theory: Prince and Smolensky (1993)

26. An important issue in computational OT

Is your GEN big enough to avoid error?
See, e.g., Karttunen (2006), Riggle (2009)

27. Can allomorph-alignment help?

We have so far twice performed allomorph-alignment (see (12), (22)):

Catalan:

<table>
<thead>
<tr>
<th>g</th>
<th>r</th>
<th>i</th>
<th>z</th>
</tr>
</thead>
<tbody>
<tr>
<td>g</td>
<td>r</td>
<td>i</td>
<td>s</td>
</tr>
</tbody>
</table>

Seediq:

<table>
<thead>
<tr>
<th>h</th>
<th>e</th>
<th>g</th>
<th>u</th>
<th>t</th>
</tr>
</thead>
<tbody>
<tr>
<td>h</td>
<td>u</td>
<td>g</td>
<td>e</td>
<td>t</td>
</tr>
</tbody>
</table>
28. **Hypothesis**

- Language learners *do not ponder candidate forms that would involve alternations not already seen.*
- This cannot be entirely true, but it may be close enough.
- If true, we may usefully consider an economical approach to GEN:

29. **The Method of Free Substitution**

- Idea from Eisenstat (2009); we simplify is slightly.
- *To find all possible surface candidates for a given UR, apply all possible alternations to all its segments, in free combination.*

- Example for Mini-Catalan:
  - Since [z] alternates with [s], the surface candidates for the UR /griz/ are [griz] and [gris].

- Example for Seediq:
  - Since [a] alternates only with [u], and [u] alternates with every vowel, the surface candidates for UR /namut/ are:
    - [namit], [namet], [namat], [namot], [namut],
    - [numit], [numet], [numat], [numot], [numut]

- In versions of the learner not reported here we move beyond this limited form of Free Substitution, motivated by empirical cases where native speakers really do introduce novel alternations.

30. **Constraints**

- For now, we assuming all constraints from UG (Prince and Smolensky 1993:2); in practice, we hand-craft them.
- See below for how we might learn the constraints.

31. **We have the ingredients we need**

- **Morpheme parsing** ((12)) yields the *allomorph set.*
- **String alignment** ((12)) yields the *set of segmental alternations.*
- The **KK hierarchy** ((13)-(24)), suitably deployed, projects a *set of possible UR’s* for each morpheme, depending on the particular level of the hierarchy we are using.
  - … and when a word has more than one morpheme, we form every combination of UR’s for all of its morphemes.
- For each UR, we can use the **Method of Free Substitution** ((28)) to create the *full surface candidate set.*
- The **real world** provides an estimate of the *probability of each surface candidate* (often zero)
- UG provides the constraint set, and each constraint must receive a proper weight.
32. **What EM will do**

- Assign a probability distribution over the candidate UR set of each morpheme (often: 1 for the right answer, 0 for the others)
- Assign a weight to each phonological constraint.
- If successful, the result will roughly match the empirically observed frequencies of surface forms.

**EXPECTATION-MAXIMIZATION**

33. **Our problem is a standard job for EM**

- It must learn both **UR probabilities** and **constraint weights**, in the face of the complex interaction of the two.

34. **The Mini-Catalan example: we start with the UR candidates**

- Here are the data again for reference:

<table>
<thead>
<tr>
<th>Morpheme</th>
<th>Surface Form</th>
</tr>
</thead>
<tbody>
<tr>
<td>gros</td>
<td>Big</td>
</tr>
<tr>
<td>grosə</td>
<td>Big fsg</td>
</tr>
<tr>
<td>gris</td>
<td>Grey</td>
</tr>
<tr>
<td>grisə</td>
<td>Grey fsg</td>
</tr>
</tbody>
</table>

- If we duly parse out the stems and the suffix, Pick-and-Choose (KK-C) will give us the following set of UR candidates:

  `‘big’:  { /gros/ }`
  `‘grey’: { /griz/, /gris/ }`
  `‘fem.’: { /-ə/ }`

- We wish to assign probabilities to these UR candidates.
- We adopt a noncommittal starting point, with equal probabilities when there are multiple candidates:

  `‘grey’:  { /griz/ 0.5, /gris/ 0.5}`
  `‘big’:  { /gros/ 1 }`
  `‘fem.’: { /-ə/ 1 }

35. **The Mini-Catalan continued: the UG constraints**

- We adopt these unchanged from Pater et al. (2012).

*FINAL VOICED OBSTRUENT = *Z]*
*INTERVOCALIC VOICELESS OBSTRUENT = *VTV
IDENT(voice)
• Intuition:
  - \(*_{\text{FINAL VOICED OBSTRUENT}}\) must be very strong, to force Final Devoicing despite \(\text{IDENT(voice)}\).
  - \(\text{IDENT(voice)}\) must be strong, to keep /s/ from voicing intervocally.
  - \(*_{\text{VTV}}\) must be much weaker than \(\text{IDENT(voice)}\).

36. Tableau for the target answer (from an Excel spreadsheet)

<table>
<thead>
<tr>
<th></th>
<th>(\text{*Z})</th>
<th>(\text{*VTV})</th>
<th>(\text{IDENT(voice)})</th>
<th>(p)</th>
</tr>
</thead>
<tbody>
<tr>
<td>/griz/</td>
<td>20</td>
<td>0</td>
<td>1</td>
<td>0.99995</td>
</tr>
<tr>
<td></td>
<td>gris</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>griz</td>
<td>1</td>
<td></td>
<td>0.00005</td>
</tr>
<tr>
<td>/griz-ə/</td>
<td>20</td>
<td>0</td>
<td>1</td>
<td>0.99995</td>
</tr>
<tr>
<td></td>
<td>grizə</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>/gros/</td>
<td>10</td>
<td>0</td>
<td>1</td>
<td>1.00000</td>
</tr>
<tr>
<td></td>
<td>grisə</td>
<td>1</td>
<td>1</td>
<td>0.00000</td>
</tr>
<tr>
<td>/gros-ə/</td>
<td>20</td>
<td>0</td>
<td>1</td>
<td>0.99995</td>
</tr>
<tr>
<td></td>
<td>grizə</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

37. Summary: goal of learning for Mini-Catalan

• Probabilities of UR’s should come out:
  - ‘grey’: \{ /griz/ 1, /gris/ 0 \}
  - ‘big’: \{ /gros/ 1 \}
  - ‘fem.’: \{ /-ə/ 1 \}

• Weights of the constraints should come out something like:

\[
\begin{align*}
*_{\text{FINAL VOICED OBSTRUENT}} &= \text{*Z} \\
*_{\text{INTERVOCALIC VOICELESS OBSTRUENT}} &= \text{*VSV} \\
\text{IDENT(voice)} &= 10
\end{align*}
\]

38. Nuts and bolts I: the search is guided by maximizing Likelihood

• Likelihood = predicted probability of the entire observed dataset (product over all data)
• This is perhaps familiar: it is the standard way to set the constraint weights in MaxEnt.

39. Nuts and bolts II: predicted probability for individual data

• This is also normal in MaxEnt, but here there is a wrinkle, for we don’t yet fully know the UR probabilities.
• Pending improvement, we compute the predicted probability of a surface form as the weighted sum of its probability across all the UR’s from which it could be derived.
40. **The core action of our system**

- In brief, we go round and round a square spiral staircase, with steadily rising Likelihood.
- Here are the four stages:
  - Use the most recent UR probabilities and constraint weights to make a prediction about the frequencies of UR-SR pairs.
  - Optimize the phonological weights using these predicted frequencies.
  - Use the new phonological weights and the previous UR probabilities to make a prediction about frequencies of UR-SR pairs.
  - Use these predicted frequencies to optimize the UR probabilities.
  - Iterate until stable.

41. **Starting point**

- The square is entered at upper left with
  - For each morpheme, all UR candidates are equiprobable.
  - All constraint weights at 1 (not zero!); this avoids starting out in a local maximum.

42. **Convergence**

- Dempster et al. showed you always converge on a stable result (see Wang (ms.) for proofs pertaining to our own version.)
- This often is the global maximum (= most accurate, highest likelihood of data), but sometimes the algorithm gets stuck in a local maximum — see later on.

43. **Four-stage optimization is nonstandard, but apparently essential**

- We tried the more standard two-stage optimization — it fails badly even on mini-Catalan.
44. Would you like to see it happen in detail?

- We have posted an Excel spreadsheet (https://linguistics.ucla.edu/people/hayes/EM/LearningMiniCatalanWithEM.xlsx) that shows every single step for the Catalan mini-example.

**TESTING THE MODEL: A TOUR OF EMPIRICAL CASES, IN INCREASING COMPLEXITY**

I. CONTINUING THE LEARNING MINI-CATALAN EXAMPLE

45. It seems to work nicely

- Weights and UR probabilities are learned at their correct values in just a few iterations.

46. History of the phonological constraint weights

- *Final Voiced Obs.* is surface-true and does some work, so rockets to the top.
- IDENT(voice) is violated in a winner (unfaithful [gris] from /griz/), so rises slightly more slowly to the middle.
- *VTV temporarily rises, foolishly imagining that it is being helpful for [griza], but in fact it is harmful for [grosə] and soon sinks back down in embarrassment, until it is safely well below IDENT(voice).
47. History of the UR probabilities

- The only nontrivial one (more than one candidate) is /griz/. It quickly migrates to probability 1 as it should.

48. Final results

- UR probabilities
  - ‘grey’ /griz/ 0.9999996
  - /gris/ 0.0000004
  - ‘big’ /gros/ 1

- Final result II: Constraint weights and tableau

<table>
<thead>
<tr>
<th></th>
<th>*FINAL</th>
<th>VOICE</th>
<th>OBS</th>
<th>*VTV</th>
<th>IDENT (voice)</th>
<th>p</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>24.15</td>
<td>1.13</td>
<td>12.42</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Grey</td>
<td>[griz]</td>
<td>[gris]</td>
<td>1</td>
<td></td>
<td>0.999992</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>0.000008</td>
<td></td>
</tr>
<tr>
<td>Grey_fsg</td>
<td>/griz-ə/</td>
<td>[gris-ə]</td>
<td>1</td>
<td></td>
<td>0.999999</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>0.000001</td>
<td></td>
</tr>
<tr>
<td>Big</td>
<td>/gros/</td>
<td>[gros]</td>
<td>1</td>
<td></td>
<td></td>
<td>1.31E-16</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>Big_fsg</td>
<td>/gros-ə/</td>
<td>[gros-ə]</td>
<td>1</td>
<td></td>
<td>0.999988</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>0.000012</td>
<td></td>
</tr>
</tbody>
</table>

49. We are doing well, we think, in comparison to Pater et al. (2012)

- Our Mini-Catalan example is identical (names aside) to their main example.
For Pater et al., everything is a constraint, e.g. UR_OF_GREY_IS_/GRIZ/.

They use a standard MaxEnt optimization on the whole constraint set, thus defined.

Result for this case: the UR probabilities can be different in different parts of the paradigm.

UR of ‘grey’ in [gris].

- /gris/ 0.92
- /griz/ 0.08

UR of ‘grey’ in [griz-ə]

- /griz/ 1.00

We’re not sure what this means —

- Isn’t the UR supposed to embody the contribution of a morpheme to pronunciation across its paradigm?
- What stops the language user from wrongly employing /gris/ in the feminine, yielding *[grisa]?

Our own system, at each stage, calculates UR probability for a stem across its whole paradigm, and obtains a single, correct value.

II. MINI-CATALAN WITH LEXICAL EXCEPTIONS

50. Adding in lexical exceptions

- Let the process of Final Devoicing have a 2% lexical exception rate — how does the system respond?

Weights:

*FINALVOICEDOBS 22.69
*VTV 0.09
IDENT(voice) 18.80

- This time, *FINALVOICEDOBS is only about 3 units higher than IDENT(voice), not 12 as before — this enforces 98% Final Devoicing.

UR probabilities:

- Invariant [s]: 1.00 /s/
- Invariant [z] (rare): 1.00 /z/
- Alternating [z ~ s] 0.9999997 /z/

51. Discussion: constraint weights

- The weights impose Zurovian frequency matching (Zuraw 2000, 2010): the predicted probability of final devoicing in a novel form matches the prevalence of final devoicing in the existing lexicon.

3 We have substituted our own data examples for clarity.
We think the literature broadly supports the existence of Zurovian frequency matching, modulated by various UG effects not modeled here. See e.g. Zuraw and Hayes (2017).

So we think our model is behaving in just the right way.

- Side note: In Zuraw’s theory, words not predictable in the grammar are lexically stored. We suggest doing this for forms whose generated probability is below some high threshold.

52. UR probabilities when there are lexical exceptions

- The winning UR’s are essentially invariant (very close to one).
- Why?
  - The learning system maximizes likelihood.
  - Forms with [-ə] manifest the UR unaltered, and are invariant.
  - There is no payoff in having a UR probability other than one.
- General pattern: grammar creates variation, matching input data; lexicon stays invariant if that is the best option for the data.

53. What about free variation?

- I.e. a single UR gives rise systematically to multiple outputs.
- We also tried this.
- The training data are not the same; but again we match the empirical probabilities while obtaining invariant UR’s.

III. ON TO COBBLING: OUR SEEDIQ RESULTS

54. Preparing the data

- For diagnostic clarity, we addressed Pseudo-Seediq, a language with exactly 25 stems, embodying all possible combinations of the five Seediq vowels.
- We obtained our surface forms (75 words) thus:
  - 25 bare stems (stressed on penult of stem, like [hégut] for /heget/)
  - 25 with a monosyllabic suffix (penultimate word stress → stress final syllable of stem: /heget-an/ → [hugétan])
  - 25 with a disyllabic suffix sequence (penultimate word stress → both stem vowels stressless) /heget-an-i/ → [hugutáni])
- We applied phonology to all of these, as just shown:
  - Pretonic vowels all reduce to [u].
  - Posttonic /e/ and /o/ reduce to [u].
- Here are representative pseudo-data:

```plaintext
b í r i h     stemRare
b u r í h a n  stemRare loc
b u r u h á n i  stemRare locimp
d í m u ts     stemForEating
```
55. Phonological constraints assumed

<table>
<thead>
<tr>
<th>PENULT STRESS</th>
<th>Cover constraint; in a full analysis there would be a battery of metrical constraints.</th>
</tr>
</thead>
<tbody>
<tr>
<td>PRETONIC REDUCTION</td>
<td>*any vowel except [u] when pretonic</td>
</tr>
<tr>
<td>POSTONIC REDUCTION</td>
<td>*[e], [o] when posttonic</td>
</tr>
<tr>
<td>IDENT(low)</td>
<td></td>
</tr>
<tr>
<td>IDENT(back)</td>
<td></td>
</tr>
<tr>
<td>IDENT(high)</td>
<td></td>
</tr>
<tr>
<td>IDENT(stress)</td>
<td>This is at the bottom of the grammar, for stress is not phonemic.</td>
</tr>
<tr>
<td>*MAP(e, i)</td>
<td>Surprisingly, stressless /e/ does not reduce to the closest legal vowel ([i]), but saltates to [u]. We follow Zuraw’s *MAP system, as implemented for saltation in Hayes and White (2015).</td>
</tr>
<tr>
<td>*MAP(e, u)</td>
<td>Included as a Faithfulness constraint violated in winners (when /e/ reduces to [u]).</td>
</tr>
</tbody>
</table>

56. Sanity check: Pick-and-Choose fails

- Seediq fails badly when we try to do it with KK-C, Pick-and-Choose
  - Why? Pick-and-Choose fails to include appropriate UR candidates in the search space.

<table>
<thead>
<tr>
<th>Stem</th>
<th>Pick-and-choose candidates</th>
<th>Cobbled candidates</th>
<th>Correct candidate</th>
</tr>
</thead>
<tbody>
<tr>
<td>‘smoke’</td>
<td>/hegut/, /huget/</td>
<td>/hugut/, /huget/, /heget/</td>
<td>/heget/</td>
</tr>
<tr>
<td>Surface</td>
<td>[hegut], [huget]</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

- The system does the best it can, guessing 0.5 probability for both /hegut/ and /huget/.
57. The model that worked

- With Cobbling in effect, all correct UR’s rose to essentially 100% probability.
- Here is the time course of learning for /netek/ ‘peck’:

![Time course graph]

- */nutuk/ falls the fastest, because it yields the most errors.

- Here is the time course arriving at a viable set of constraint weights:

![Constraint weights graph]

58. Something to watch out for

- Weights of Markedness constraints are sometimes surprisingly low.
  - Observe that PostonicReduction — an inviolable principle of Seediq — has a rather weak weight, about 2.4.
  - Why? Because of our free-substitution GEN, which includes no candidates (such as [nétek]) that need this constraint to rule them out — stressless [e] is not in the symbol alphabet!
  - We suspect that a full weighting process would have to make use of both phonotactics and alternations.
IV. MOVING UP TO FULL PROBLEM-SET SCALE: CATALAN

59. Catalan phonology

- Romance. Catalonia
- There is excellent descriptive and analytic work, including Mascaró (1976) and Wheeler (2005).
- We started with the popular problem set from Kenstowicz and Kisseberth (1979), amplifying it with further study.
- We open with a survey of phenomena and the constraints that can handle them.

60. The chain of word final truncations

**Homorganic cluster simplification:**

\[
\begin{align*}
/\eta k/ & \rightarrow [\eta] / \_\_\_ (s) / \_\_\_ (s) / \text{word} & /\text{bla}n/ & \rightarrow [\text{bla}n] \text{‘white-msg.’} \quad (\text{cf. [bla}n-\text{a] fsg.}) \\
/nt/ & \rightarrow [n] / \_\_\_ (s) / \text{word} & /\text{sant}/ & \rightarrow [\text{san}] \text{‘holy-msg’} \quad (\text{cf. [san-\text{a] fsg.)}) \\
/rt/ & \rightarrow [r] / \_\_\_ (s) / \text{word} & /\text{fort}/ & \rightarrow [\text{for}] \text{‘strong-msg.’} \quad (\text{cf. [fort-\text{a] fsg.)})
\end{align*}
\]

**Single-consonant dropping — opaque, and exceptionful**

\[
\begin{align*}
/n/ & \rightarrow \emptyset / \_\_\_ (s) / \text{word} & /\text{bon}/ & \rightarrow [\text{bo}] \text{‘good-msg.’} \quad (\text{cf. [bon-\text{a] fsg.)}) \\
& & \quad \text{but there are many lexical exceptions, e.g.} \\
/prəgon/ & \rightarrow [prəgon] \text{‘proclamation’} \quad (\text{cf. [prəgon-\text{a] fsg.)})
\end{align*}
\]

\[
\begin{align*}
/r/ & \rightarrow \emptyset / \_\_\_ (s) / \text{word} & /\text{dur}/ & \rightarrow [\text{du}] \text{‘hard-msg.’} \quad (\text{cf. [dur-\text{a] fsg.)}) \\
& & \quad \text{but there are many lexical exceptions, e.g.} \\
/dur-s/ & \rightarrow [\text{dus}] \text{‘mpl.’} \quad (\text{cf. [dur-\text{a] fsg.)}) \\
/pur/ & \rightarrow [\text{pur}] \text{‘pure-msg.’} \quad (\text{cf. [pur-\text{a] fsg.)}) \\
/pur-s/ & \rightarrow [\text{purs}] \text{‘mpl.’}
\end{align*}
\]

**Key constraints and weighting relationships**

- \text{*FINAL HOMORGANIC CLUSTER >> MAX} \quad \text{forcing simplification}
- \text{*FINAL [n] > MAX} \quad \text{stochastic weighting, forcing (sometime) deletion}
- \text{*CODA [r] > MAX} \quad \text{stochastic weighting, forcing (sometime) deletion}
- \text{MAX(seg+seg) >> *FINAL [n], *CODA[r]} \quad \text{ensuring opacity: double deletion forbidden}

I-\text{CONTIG} \text{ is needed to block */fort/ \rightarrow */fot].}

61. Epenthesis

- Like English: final sibilant clusters created by plural */-s/ are split by Epenthesis.

\[
/griz-s/ \rightarrow [\text{grizus}] \text{‘grey-pl.’}
\]
• Ranking needed: *SIBILANTCLASH >> DEP

62. Final Devoicing

• See (11) above for Final Devoicing, which also affects /b, d, g, dʒ, z/.
  ➢ Needed weight: *FINAL VOICED OBSTRUENT >> IDENT(voice)
• We also include AGREE(voice), so that the voiceless /-s/ suffix can require devoicing of the preceding stem-final segment.

63. A quirk of Catalan Final Devoicing

• Oddly, final /ʒ/ surfaces as [tʃ], not [ʃ] (which is both legal and phonetically closer).
  ➢ This again is Saltation (Lubowicz 2002, Ito and Mester 2003, Hayes and White 2015):
  \[
  \begin{array}{c}
  \text{ʃ} \\
  \text{ʒ} \\
  \end{array}
  \begin{array}{c}
  \text{ʃ} \\
  \text{ʃ} \\
  \text{ʃ} \\
  \text{ʒ} \\
  \text{tʃ} \\
  \text{ʒ} \\
  \text{ʒ} \\
  \text{ʒ} \\
  \text{ʒ} \\
  \text{ʒ} \\
  \end{array}
  \]
• As before, we follow the “recipe” for Saltation given in Hayes and White (2015):
  ➢ Ban the unwanted “short” journeys with *MAP constraints.
  ➢ Here, the key constraint is *MAP(ʒ ~ ʃ), which forces the victory of [tʃ].
  ➢ To be thorough, we included all six *MAP constraints definable with the four segments above.
  ➢ Catalan is an “unnatural language”, per Zuraw (2007, 2013), since a *MAP constraint that bans a short journey (*MAP(ʒ, ʃ)) must be stronger than a *MAP constraint that bans a longer journey (*MAP(ʒ, tʃ)).
  ➢ There is evidence, not covered here, that Catalan speakers are “repairing” this pattern.

64. Overall, it’s a decent-sized problem set

• The Hasse diagram below gives the full set of classical OT rankings, calculated by FReD (Fusional Reduction Algorithm, Brasoveanu and Prince 2011) in OTSoft 2.6 (Hayes, Tesar and Zuraw 2022)
  ➢ In order to be able to use FReD, we pretended that R Drop and N Drop are exceptionless.
65. Our procedure as applied to Catalan

- Input file of 33 four-member paradigms, with masculine/feminine and singular/plural

  
  \[
  \begin{align*}
  [\text{ultim}] & \quad \text{Last masc sing} \\
  [\text{ultims}] & \quad \text{Last masc plur} \\
  [\text{ulta}] & \quad \text{Last fem sing} \\
  [\text{ultimas}] & \quad \text{Last fem plur} \\
  [\text{bo}] & \quad \text{Good masc sing} \\
  [\text{bons}] & \quad \text{Good masc plur} \\
  [\text{bona}] & \quad \text{Good fem sing} \\
  [\text{bones}] & \quad \text{Good fem plural} \\
  [\text{ple}] & \quad \text{Full masc sing} \\
  [\text{plens}] & \quad \text{Full masc plur} \\
  [\text{plena}] & \quad \text{Full fem sing} \\
  [\text{plenas}] & \quad \text{Full fem plur} \\
  [\text{sa}] & \quad \text{Sane masc sing} \\
  [\text{sans}] & \quad \text{Sane masc plur} \\
  [\text{sana}] & \quad \text{Sane fem sing} \\
  [\text{sanas}] & \quad \text{Sane fem plur} \\
  [\text{pragon}] & \quad \text{Proclamation masc sing} \\
  [\text{pragons}] & \quad \text{Proclamation masc plur} \\
  [\text{pragona}] & \quad \text{Proclamation fem sing} \\
  [\text{pragonas}] & \quad \text{Proclamation fem plur}
  \end{align*}
  \]

- Special treatment for /n/-stems and /r/-stems:
  - For each class we included 3 deletors and 1 non-deletor.
  - The goal was to see if the system would predict .75/.25 frequency matching.

66. What happened with these data

- Our morpheme-divider located the suffixes [-s] and [-ə], yielding e.g.

  \[
  \begin{align*}
  [\text{bo}] & \quad \text{Last} \\
  [\text{bon-s}] & \quad \text{Last plur} \\
  [\text{bon-a}] & \quad \text{Last fem} \\
  [\text{bon-a-s}] & \quad \text{Last fem plur}
  \end{align*}
  \]

- All allomorph sets were collected and aligned, e.g.
From this, we obtained the alternation list, which for Catalan is fairly extensive:

> { n ~ ∅, r ~ ∅, d ~ ∅, t ~ ∅, k ~ ∅, p ~ ∅, o ~ ∅,
  d ~ t, g ~ k, b ~ p, z ~ s, dʒ ~ tʃ, ʒ ~ tʃ }

We opted for Pick-and-Choose on the KK hierarchy, yielding a set of candidate UR’s, e.g.:

> { bon, bo }

There were either 1 or 2 candidate URs per stem.

Surface candidates (GEN), averaging 85 per word, were then created using the alternation list.

Constraints employed: see (57)-(60) above.

We ran our version of EM using these candidate URs, surface candidates, constraints, and violations.

It took 10 minutes to do 20 iterations, which brought the system very close to the maximum Likelihood value.

67. Results

- For all 33 stems, the correct UR was assigned high probability, no lower than 0.9998 in any case.
  - This includes UR’s for forms with /r/ and /n/, with lexical variation: the /r/ or /n/ always appears in these URs.
- For all stems with an exceptionless pattern, the correct winner was derived with high probability, never lower than 0.99999.
- For the /n/ and /r/ stems, with lexical variation, the system frequency-matched: .75/.25 in each case.
- So we are happy with this as a preliminary result.

V. AN EVEN HIGHER LEVEL ON THE KK HIERARCHY, AND HOW IT LEADS TO FAILURE

68. Purpose

- Show that EM is not a panacea for UR learning; you need to regulate the hypothesis space as well.

69. A new KK level, invented by us: KK-D’, “hypercobbling”

- Apply the Method of Free Substitution (29) to obtain UR’s (we’re already doing this to obtain SR’s).
- This finds all the UR’s found by KK-D (Cobbling) and, often, more.
70. What happened when we re-ran Catalan with hypercobbling

- The system got stuck in a local maximum.
- This local maximum attributes variation in /r/-stems to the underlying representation, not the grammar.
  - So [du] ~ [dur-ə] is 0.5 /dur/ and 0.5 /du/.
  - The grammar has MAX far above *CODA /r/ — no deletion of coda /r/ in the phonology.
  - This frequency-matches the masculine forms, though badly (should be .75/.25, not .5/.5).
  - It fails outright in the feminine: the bad output *[du-ə] is generated with substantial frequency.

71. The wisdom of Jarosz

- Jarosz (2006, Hopkins dissertation) is the seminal first application of EM to phonological learning.
- Jarosz repeatedly emphasizes that EM is no panacea, but depends on the size/nature of the search space.
  - “The initialization of the parameters is an important factor in determining how good the final grammar is ….. when the model is too free, initialization drastically affects performance.”
  - “In related work … Charniak (1993) reports that in experiments running EM … with 300 different randomly generated initial parameters, 300 different local maxima were learned.”
- Our system is meant to use knowledge from other domains (pre-phonological morpheme parsing, KK-hierarchy-based UR limitations) to ease the burden on EM to the point that it can succeed.
- Hypercobbling, as far as we can tell, serves no useful purpose (i.e. in including correct hypotheses not found by simpler methods); it is here simply to illustrate Jarosz’s point in context.

SUMMARY

72. One-sentence version

- Our version of EM can serve as the basis for learning correct URs from data on the scale of a problem set, provided it is given a search set contrained by (a) pre-phonological morpheme parsing; (b) some level of the KK hierarchy; (c) a universal phonological constraint set.
  - By “scale of a problem set”, we mean at least a few dozen forms, with multiple alternation patterns, including the possibility of Cobbling.
73. **We suggest this is progress**

- The earlier work on UR learning that has inspired/helped us has worked only with very small problems, often with hand-invented data.

74. **Moreover**

- Our system is happy to learn free variation.
- Faced with lexical variation, it engages (correctly, we think) in frequency-matching.
- But even if the data vary, our system has been able to avoid inappropriate variation in URs.

75. **Reflections on joint learning**

- Goldwater (2018) makes an eloquent case for **joint learning**: you learn better if you learn everything at once, in one single algorithm (you can apply all relevant information at every stage).
  - Computational work that illustrates her point: Elsner et al. (2013), Feldman et al. (2013).
- Our work is *not* in conformity with Goldwater’s principle — for us, morpheme division is learned at a separate, earlier stage.
- Yet joint learning surely has its limits — do humans learn allophones and island constraints at the same time?
- We conjecture that our deviation from Goldwater’s principle may be a necessary one, though we are alert to ways we might increase the “jointness” of our system.

**FUTURE WORK**

76. **Scaling up from problem-set size to whole-language size**

- Simply feeding our system thousands of stems leads it to run out of memory.
- We have a bunch of ideas to try out, hoping to improve the capacity of the system.
- One possibility is to let the initial stage of data fix the phonology in place, so learning future UR’s is less work.
  - This is plausibly what happens when adults expand their vocabulary (Arbuckle et al. 1989).

77. **Learning the constraints**

- We are not alone in being skeptical of the Universal Constraint Set. Can’t we *learn* the constraints instead? Some ideas:
- For Faithfulness:
  - The full alternation set is a rich source of information.
The key will be to extract from it a set of Faithfulness constraints that are sufficiently general to project to novel cases, but sufficiently detailed to cover the data.

- For Markedness:
  - An obvious move is to appeal to prior *phonotactic learning*
  - It is a hallmark of OT that phonotactics guides the morphophonemic alternation pattern.
  - Hale (xxx find ref.), Hayes and Wilson (2008), Chong (2019) suggest that the connection comes from learning: phonotactic learning tells you the strong Markedness constraints in advance.

  - We might usefully ask: what phonological properties do “unfaithful” UR’s have that the surface forms lack?

78. Wug-testing?

- A full theory of phonological learning must be able to pass a wug test.
- Albright and Hayes (2003) were able to do this at a primitive level, assuming KK B”, single-surface-base, and an ad hoc substitute for OT.
- Theories with more elaborate UR’s need to incorporate principles that perform inverse probabilistic mapping: given an SRs, assign a probability distribution over UR candidates, then apply the phonology in the forward direction.
- Moreover, they must permit the constraint set to be rich enough to capture islands of reliability (Zuraw 2000, Albright 2000, 2002, Ernestus and Baayen 2003); native speaker knowledge tends to be much more fine-grained than the problem-set level.

79. Thank you for your attention and feedback :=)

---

4 Meaning: A UR candidate that, if you select it as correct, surface forms would unfaithful to it.
APPENDIX: FINDING ALLOMORPHS IN SURFACE DATA

80. Candidates for a MaxEnt grammar

- Over the entire data, we are given surface forms and the morphemes present.
- Let each morpheme have an index.
- GEN: all possible coindexations of the segments with their morpheme.
- Correct coindexations for Mini-Catalan:

  \[
  \begin{align*}
  g_1 \ r_1 \ o_1 \ s_1 & \quad \text{big}_1 \\
  g_1 \ r_1 \ o_1 \ s_1 - \partial_2 & \quad \text{big-feminine}_2 \\
  g_3 \ r_3 \ i_3 \ s_3 & \quad \text{grey}_3 \\
  g_3 \ r_3 \ i_3 \ z_3 - \partial_2 & \quad \text{grey}_3\text{-feminine}_2
  \end{align*}
  \]

- Coindexation is more realistic than boundary insertion (see e.g. Nelson 2019) because it allows non-contiguous surface allomorphs (e.g., stems split by infixation or metathesis).

81. Constraints for a MaxEnt “grammar” of morpheme parsing

- **BE SIMILAR**
  - penalizes difference between allomorphs of same morpheme
  - uses the metric of Bailey and Hahn (2001), with more recent advances in phonological similarity by Wilson and Obdeyn (2009) and White (2017)
  - relates to existing research: Pressure toward allomorph-similarity is known elsewhere in phonology (historical change, Kiparsky (1982); elicitation from children, Jo (2017), Do (in press); artificial grammar learning studies (Wilson 2006; White 2013, 2014, Chong and Sundara).
  - Expressed in a strong version for stem allomorphs, a weak one for affixes, per typology
- **BE VARIEGATED**
  - Don’t have a great number of stems that start or end with the same segment.
  - Helps prevent affix segments that have assimilated to the stem from being misinterpreted as part of the stem.
- **BE CONTIGUOUS**
  - Comes in different strengths, based on the length (1, 2, 3, 4, max) of the interruption.

82. This works reasonably well

- It can find the allomorphs in the cases studied here.
- Where the search procedure fails to find the right answer, often the right answer is more harmonic than the found answer (and its neighbors = local maximum)
- Future work: output a set of parses, including probability, so that later processing can take up the slack, avoiding uncorrectable error.
References


Hale, Mark, on learning connection xxx find this if it exists


Nazarov? xxx

●