RUTGERS
THE STATE UNIVERSITY
OF NEW JERSEY

Learning Phonological Grammars for Output-Driven Maps

Bruce Tesar
Linguistics Dept. / Center for Cognitive Science
Rutgers University, New Brunswick

DGfS, Osnabruck. March 4, 2009.

Acknowledgements

- Crystal Akers (Rutgers University)

Outline

- Computational issues in the learning of phonologies
  - learning rankings and underlying forms
  - further structure required for plausibly efficient learning
- Output-Driven Maps
- Exploiting ODM structure in learning

Learning Phonologies

- Must simultaneously learn the ranking and the lexicon of underlying forms (Tesar & Smolensky 1996; Hale & Reiss 1997).
- Exhaustively evaluating all possible lexicon-ranking combinations (Hale & Reiss 1997) is hopelessly intractable.

Prior Work

- Jarosz (2006): likelihood maximization.
  - Separately evaluates each possible UF for each morpheme (as well as each possible ranking).
- Apoussidou (2007): lexical constraints against possible underlying forms (also Boersma 2001).
  - Separately evaluates each possible UF for each morpheme.

Contrast Pair and Ranking

- Evaluate local lexica for a small morpheme set (Merchant & Tesar 2005/2008, Merchant 2008).
  - Local lexicon: possible assignment of feature values to **unset** underlying features.
  - Better than all possible UFs, but still exponential in the number of unset features.
- Both underlying feature setting and ranking information extraction are dependent upon evaluating all local lexica.

Need Additional Structure

- These techniques are still implausibly slow.
  - Processing all UFs for even a modest number of morphemes gets expensive very quickly.
- Faster learning will require additional posited structure in the space of possible grammars.
- Proposal: Output-Driven Maps

Terminology

- A **candidate** is an input, an output, and a correspondence relation between them.
  - An **input** for a word is constructed from the underlying forms (UFs) for the morphemes of the word.
  - $/p_1a_2k_3a\!:_4/ \rightarrow [p_1\acute{a}_2k_3a_4]$
- A candidate has a set of **disparities**.
  - Differences between input-output correspondents.
- A **mapping** is an optimal candidate.
- A phonological **map** is the set of optimal candidates.

A System for Illustration

- Words: root + suffix
  - Both roots and suffixes are monosyllabic.
- Each vowel has two features:
  - Vowel length: long (+) or short (–)
  - Main stress: stressed (+) or unstressed (–)
- Example surface words:
  - *páka pá:ka paká páka: pa:ká: pa:ká*
  - Each word has two morphemes
  - Each word has exactly one main stress in the output.

The Constraints

- Six Constraints

MainLeft	main stress on the initial syllable
MainRight	main stress on the final syllable
*V:	no long vowels
WSP	long vowels are stressed
FaithStress	correspondents have equal stress value
FaithLength	correspondents have equal length value

(McCarthy & Prince 1993, 1995; Prince 1990; Rosenthall 1994)

Language A

r1=/*pa*/	r2=/*pa:*/	r3=/*pá*/	r4=/*pá:*/	
páka	*pá:ka*	*páka*	*pá:ka*	s1=/-*ka*/
páka	*pá:ka*	*páka*	*pá:ka*	s2=/-*ka:*/
paká	*paká*	*páka*	*pá:ka*	s3=/-*ká*/
paká:	*paká:*	*páka*	*pá:ka*	s4=/-*ká:*/

Ranking: WSP ≫ FS ≫ ML ≫ MR ≫ FL ≫ *V:

Lexical stress (default initial), long vowels shorten in unstressed position.

Note: s1 /-*ka*/ and s2 /-*ka:*/ neutralize in all environments.

Output Restrictions

- **Theoretical claim**: most phonological requirements enforce output restrictions.
- Said another way, phonological disparities are driven by restrictions on the output.

- How can this be formally expressed?

Output-Driven Maps

(Tesar 2008)

- A map is output-driven if:
  - for every grammatical candidate A→X of the map:
  - if candidate B →X (same output) has greater similarity than A→X,
  - then B→X is also grammatical.

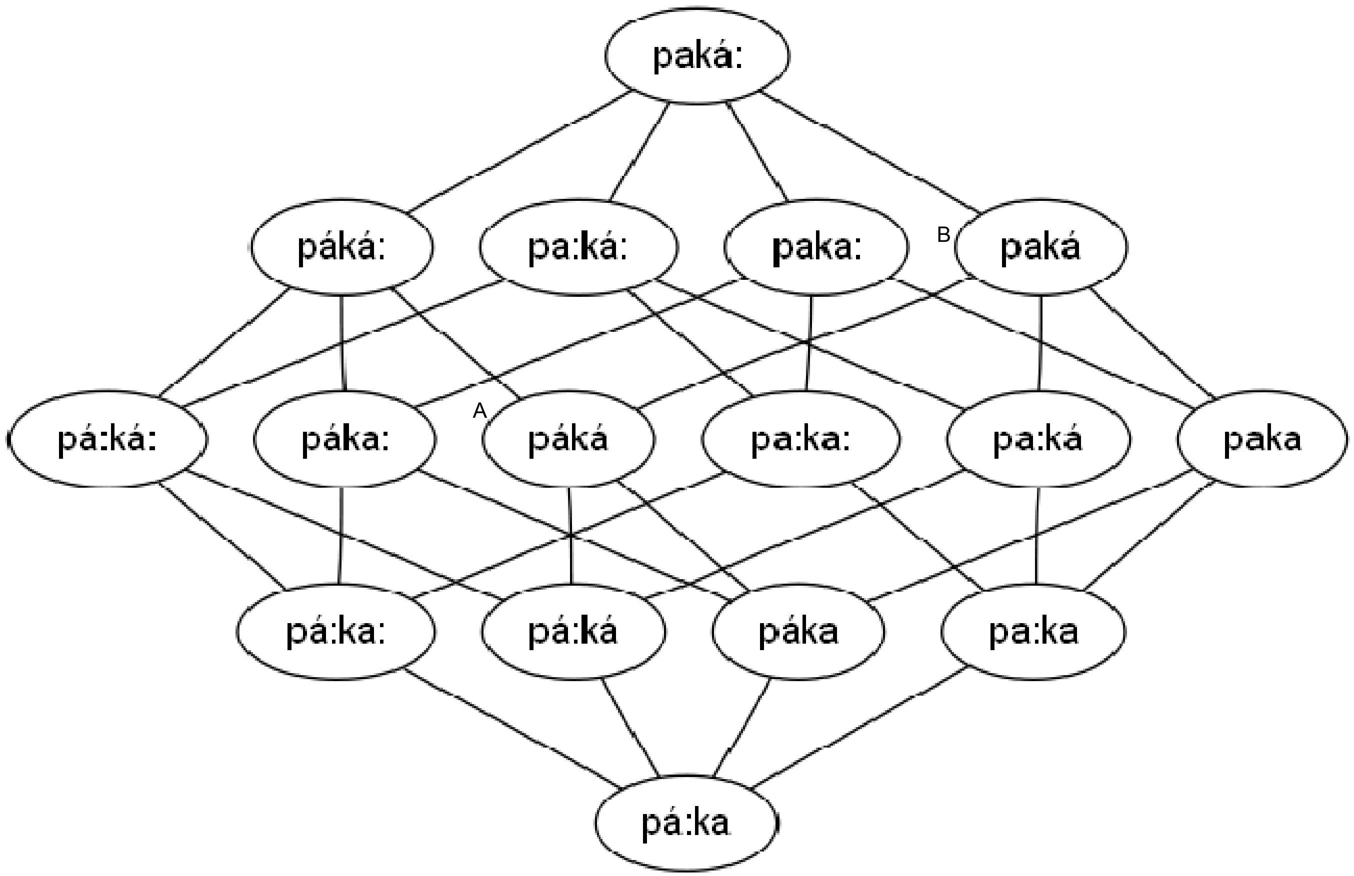
- Simplified:
  - for every grammatical candidate A→X of the map:
  - if input B is more similar to X than A is,
  - then B also maps to X.

Greater Similarity

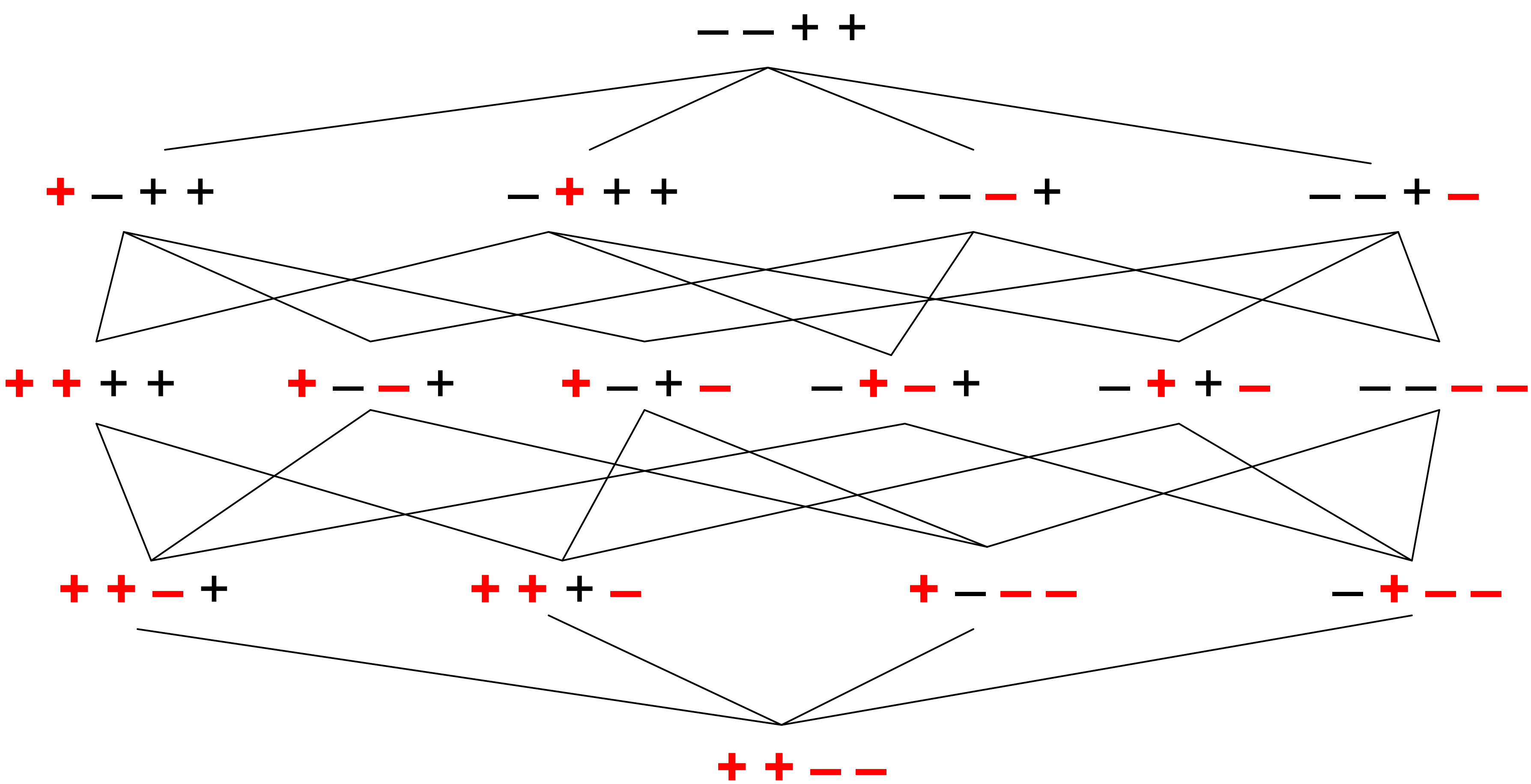
- Candidate B→X has **greater similarity** than candidate A→X if every disparity in B→X has an identical corresponding disparity in A→X.
  - The relation is only defined for pairs of candidates sharing the same output.

		(+/–stress +/–length)
A→X	*páká → paká:*	[+ – + –] → [– – + +]
B→X	*paká → paká:*	[– – + –] → [– – + +]

Relative Similarity (up = greater similarity)



Relative Similarity (+/–stress +/–length)



(A Piece of) An Output-Driven Map

- *páká:* → *páka* 2 disparities
- *páká* → *páka* 1 disparity
- *páka:* → *páka* 1 disparity
- *páka* → *páka* 0 disparities (Identity Mapping)

- Output conditions force /*páká:*/ to accept 2 disparities to reach the "nearest" phonotactically valid output.
- Inputs with greater similarity to [*páka*] require only a strict subset of those disparities to reach [*páka*].
- Output-driven: simply removing some obstacles to an output ensures reaching that same output.

A Non-Output-Driven Map

• *páká:* → *páka*	2 disparities
• *páká* → *paká*	Different Output!
• *páka:* → *páka*	1 disparity
• *páka* → *páka*	0 disparities (Identity Mapping)

The Identity Map Property

- The Identity Map Property
  - All grammatical outputs "map to themselves".
  - Common assumption, especially with respect to phonotactic learning.
- All output-driven maps have the Identity Map Property.
  - No input is more similar to an output X than X itself.
  - If any input maps to X, then X maps to X.
- Consequence: maps with chain shifts are not output-driven.

Output-Driven Maps in Optimality Theory

- An OT system is guaranteed to define only output-driven maps if two main conditions are met:
  - *Gen* must be **correspondence uniform**.
  - All constraints of Con must be **output-driven-preserving**.
  - These are sufficient conditions.
- Correspondence uniformity is fully consistent with a standard “freedom of analysis” view of *Gen*.

Output-Driven-Preserving Constraints

- The details of output-driven-preserving (ODP) are technical, but require of a constraint C:
  - if B→X has greater similarity than A→X,
  - and B→Y has fewer violations of C than B→X,
  - then A→Y must have fewer violations of C than A→X.
- Consequence: all markedness constraints are ODP.
- “Basic” IO faithfulness constraints are ODP.
  - Max, Dep, Ident
  - See Tesar (2008) for proofs of the (non)ODP status of a variety of constraints.

Exploiting ODM Structure in Learning

- ODM structure can be exploited in the learning of both:
  - underlying feature values
  - ranking information

- The primary benefit: computational efficiency
  - Converts exponential search to linear search

Phonotactic Learning

- Phonotactic learning commonly uses underlying forms that are (effectively) identical to the observed output.
  - Identity mappings for observed words
  - Prince & Tesar (2004), Hayes (2004)

- The Identity Map property follows from ODM structure.
  - Phonotactic learning can be done as before.

Phonotactic Ranking Information (Lang. A)

		WSP	ML	MR	*V:	FS	FL
r1s1	*páka ~ paká*		W	L		W	
r1s3	*paká ~ páka*		L	W		W	
r1s4	*paká: ~ paká*				L		W

Applying Biased Constraint Demotion:

WSP $\gg$ FS $\gg$ {ML,MR} $\gg$ FL $\gg$ *V:

Learning Underlying Feature Values

- ODM: A→X entails B→X
- Contrapositive: NOT (B→X) entails NOT (A→X)
  - If a given input cannot map to the output, then all inputs with lesser similarity (additional disparities) cannot map to that output.

Testing Individual Disparities

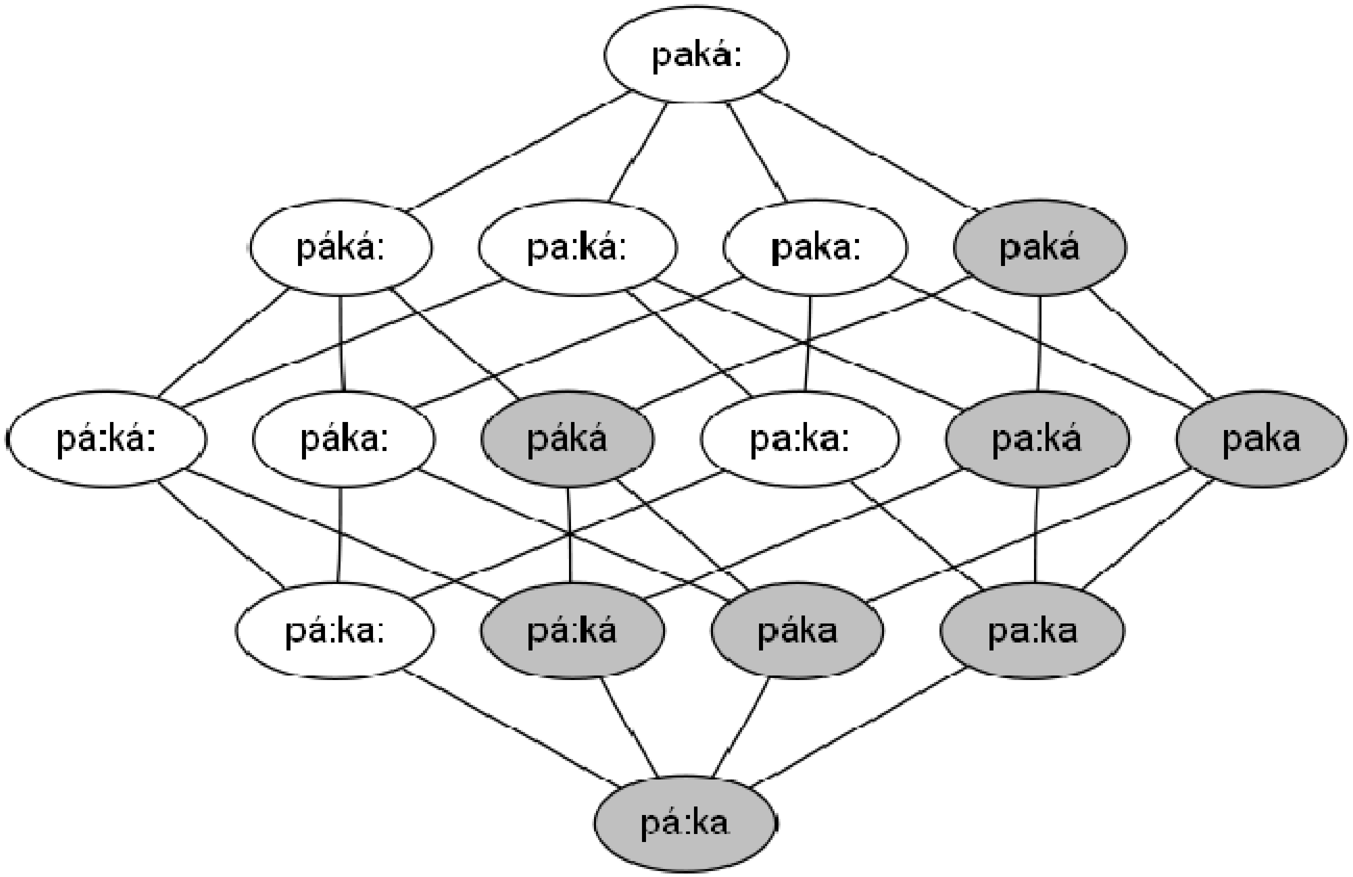
- Observed output (r1s4): *paká:*
- What is the underlying length of suffix s4?
- *paká→paká:* disparity for s4 length only.
- If *paká→paká:* is inconsistent
  - no other input with s4 set to short maps to *paká:*
  - s4 can be set to long (+).

paká→paká: is Inconsistent

/paká/	WSP	ML	MR	*V:	FS	FL
paká:		*		*		*
paká		*				
ERC *paká: ~ paká*				L		L

paká→paká: is harmonically bound.

Setting s4 to +long



Exponential to Linear

- The learner only needs to test one input for each unset underlying feature.
  - Set one underlying feature to **mismatch** the output, set the others to **match** the output.
- The number of inputs to be evaluated is linear in the number of unset features, rather than exponential (even at the outset of learning).
- Complication: the number of inputs to be evaluated increases for some multi-word sets (more on this later).

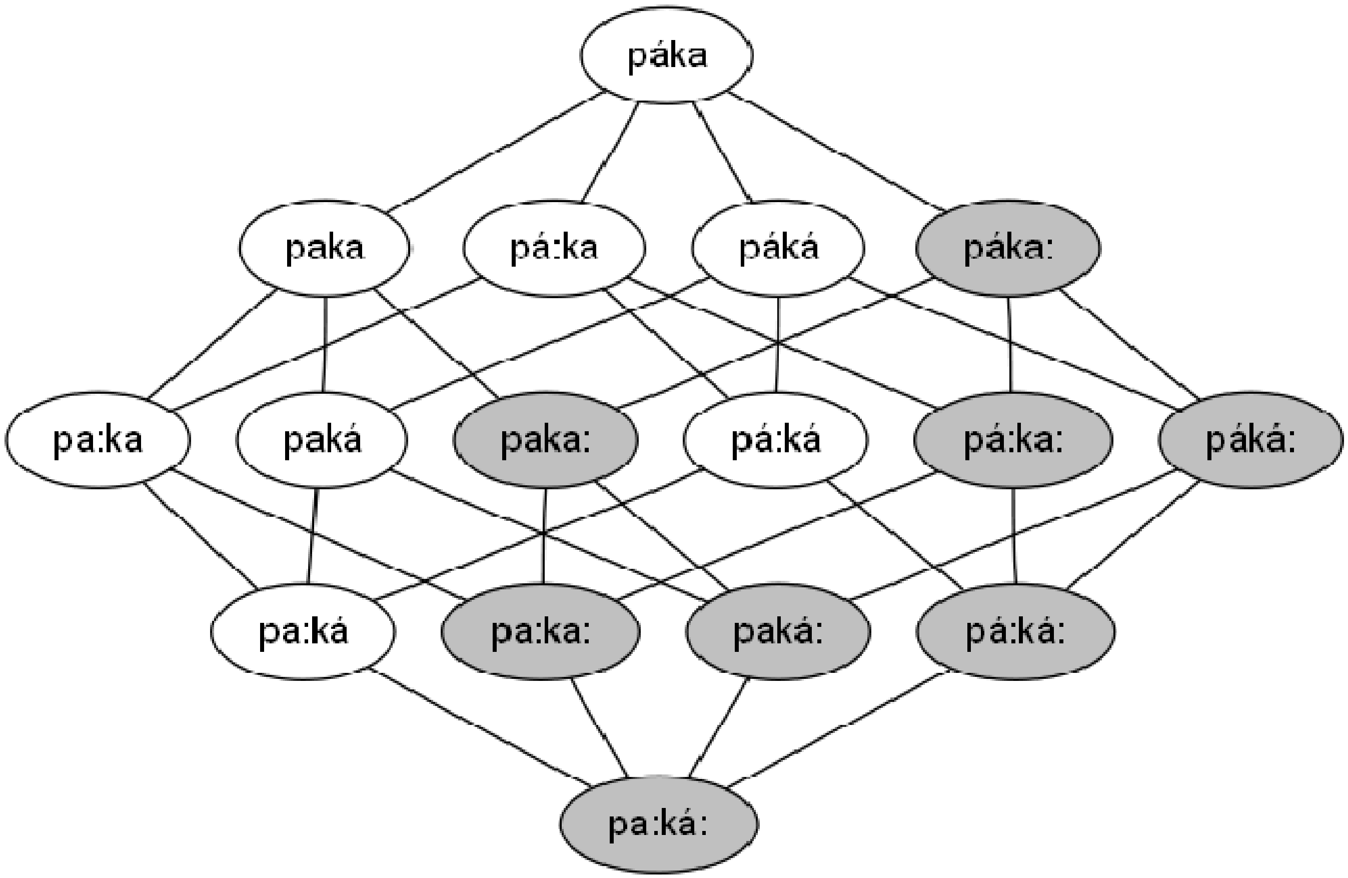
Ranking Information with ODM

- Once a feature has been set, the value is fixed for any word containing that morpheme.
- Further ranking information can be obtained from forms in which a set feature is not faithfully preserved (Tesar 2006b).

Nonfaithful Features

- Observed output (r3s4): [*páka*]
- s4 has already been set to +long.
  - Lexicon: r3 /?,?/ s4 /?,+/
- Minimal disparity mapping: *páka:* → *páka*
  - NOT an identity mapping.

Available Inputs for r3s4



Ranking Info from r3s4

/*páka:*/	WSP	ML	MR	*V:	FS	FL
páka						*
páka:	*			*		
ERC	W			W		L
paká: ~ paká				L		W
Fusion	W			L		L

WSP ≫ FL ≫ *V:

Obtained despite incomplete input knowledge.

Single Word Learning

		WSP	ML	MR	*V:	FS	FL
r1s1	*páka ~ paká*		W	L		W	
r1s3	*paká ~ páka*		L	W		W	
r1s4	*paká: ~ paká*				L		W
r3s4	*/páka:/ páka ~ páka:*	W			W		L

WSP ≫ FS ≫ {ML,MR} ≫ FL ≫ *V:

r1 /?,–/ r2 /?,+/ r3 /?,–/ r4 /?,+/

s1 /?,?/ s2 /?,?/ s3 /?,–/ s4 /?,+/

Multi-word Sets

- Processing multiple words simultaneously has value when a morpheme is shared between the words.
  - Needed to obtain key lexical information (Tesar 2006b).
  - The shared morpheme must have the same UF for all words in the set.
- Alternating feature within the set: no single underlying value will match the surface everywhere.
- All values of the alternating feature must be tried in combination with each other tested feature value.
  - Exponential growth only with respect to unset features alternating within the words being processed.

Contrast Pair (Tesar 2006a)

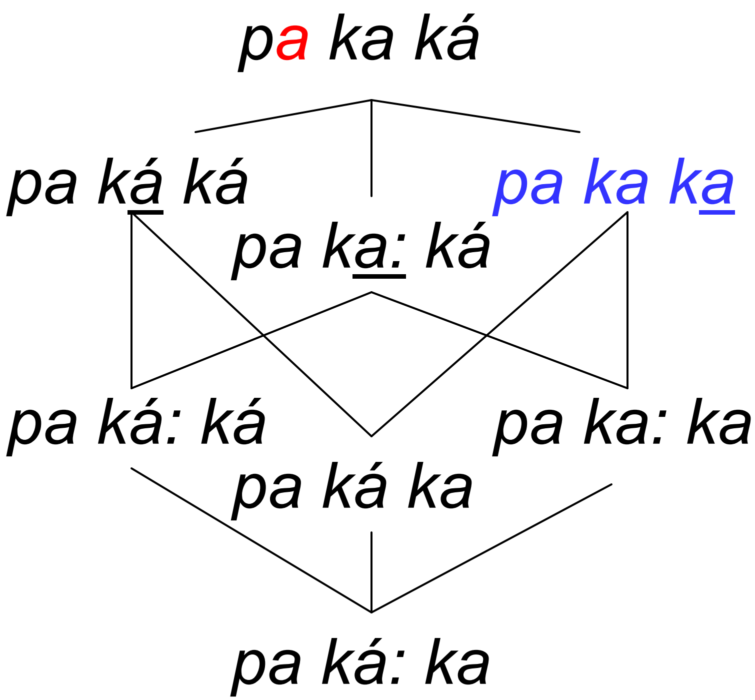
- r1s1 [*páka*] r1s3 [*paká*]
  - r1 alternates in stress
- Lexicon: r1 /?,–/ s1 /?,?/ s3 /?,–/
- Testing the stress feature for s3 involves both:
  - r1 /**–**,–/ s1 /–,–/ s3 /–,–/
  - r1 /**+**,–/ s1 /–,–/ s3 /–,–/
- If both lexical hypotheses are inconsistent, we can set s3 to be +stress.

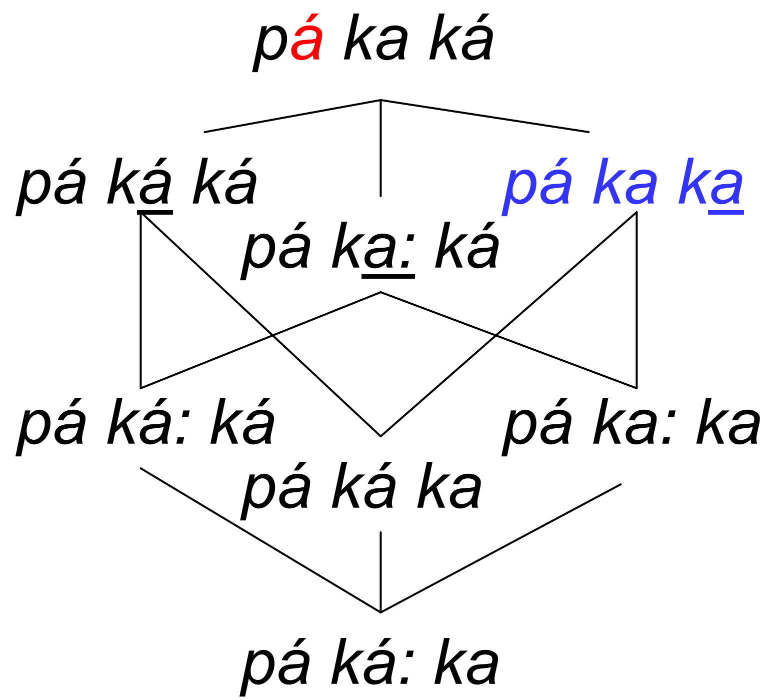
Contrast Pair r1s1 / r1s3

r1s1: /?,– ?,?/ → [+,– –,–] → *páka*

r1s3: /?,– ?,–/ → [–,– +,–] → *paká*

(r1 s1 s3)





Morphemic Contrast at Work

- Both s3 –stress hypotheses guaranteed to fail:
  - *pa ka k<u>a</u>* *pá ka k<u>a</u>* (r1 s1 s3)
  - s1 and s3 must differ underlyingly.

- Set s3 to be +stress.

More Ranking Information: r3s3

- r3s3: *páka*
- Chosen because s3's stress is unfaithfully mapped.

/*pákál*/		WSP	ML	MR	*V:	FS	FL
páka	*winner*			*			
paká	*loser*		*				
ERC			W	L			

- Learned: ML $\gg$ MR
- Now the other underlying stress features can be set.

Learned Grammar for Lang. A

Ranking: WSP $\gg$ FS $\gg$ ML $\gg$ MR $\gg$ FL $\gg$ *V:

Lexicon:

r1 /–,–/	r2 /–,+/	r3 /+,–/	r4 /+,+/
s1 /–,?/	s2 /–,?/	s3 /+,–/	s4 /+,+/

- s1 and s2 are homophonous.
- An unaccented suffix is never stressed, and its length is always neutralized.

Unset, Not Underspecified

- Unset features: can be set to any value without affecting the morpheme's behavior.
  - NOT "essential" underspecification (Archangeli 1988; Kiparsky 1982).
- "(non)contrastive" is a property of individual features in phonological context.
  - Length is noncontrastive in s1 and s3, but contrastive for the other morphemes.

On-line Learning, With Memory

- At a given time, the learner can process whatever forms they have reliably observed.
  - Need to store some output forms in memory, in order to analyze them into morphemes.
- No need to wait until all words of a paradigm have been stored.
  - No need for Initial Lexicon Construction (Tesar et al 2003).
  - Contrast pairs can be formed if/when the relevant contrasting words have been stored.

Contact with Acquisition

- This learning approach predicts dependencies.
  - Learn length feature values for a few key morphemes.
  - Then learn WSP $\gg$ FL $\gg$ *V:
  - Then learn stress feature values for a few key morphemes.
  - Then learn ML $\gg$ MR
- These dependencies should constrain order of acquisition.

Conclusions

- Output-Driven Maps are a good next approximation to the structure of basic phonology.
- ODM structure makes much more efficient learning possible.
  - Reduces from exponential to linear
  - Both underlying forms and ranking information.
- Future:
  - can ODM structure be expanded to include attested instances of non-OD phenomena, while remaining exploitable in learning?
  - translating UF-ranking dependencies into predictions about acquisition data.

References

Apoussidou, Diana. 2007. The Learnability of Metrical Phonology. PhD. dissertation, University of Amsterdam, Amsterdam.

Archangeli, Diana. 1988. Aspects of underspecification theory. *Phonology* 5:183-207.

Boersma, Paul. 2001. Phonology-semantics interaction in OT, and its acquisition. In *Papers in Experimental and Theoretical Linguistics*, ed. by Robert Kirchner, Wolf Wikeley, and Joe Pater, 24-35. Edmonton: University of Alberta. ROA-369.

Hale, Mark, and Reiss, Charles. 1997. Grammar Optimization: The simultaneous acquisition of constraint ranking and a lexicon. Ms., Concordia University, Montreal. ROA-231.

Hayes, Bruce. 2004. Phonological acquisition in Optimality Theory: The early stages. In *Constraints in Phonological Acquisition*, eds. René Kager, Joe Pater and Wim Zonneveld, 158-203. Cambridge: Cambridge University Press.

Jarosz, Gaja. 2006. Rich Lexicons and Restrictive Grammars - Maximum Likelihood Learning in Optimality Theory. PhD. dissertation, The Johns Hopkins University, Baltimore, MD. ROA-884.

Kiparsky, Paul. 1982. Lexical phonology and morphology. In *Linguistics in the Morning Calm*, ed. I. S. Yang, 3-91. Seoul: Hanshin.

McCarthy, John J., and Prince, Alan. 1993. Generalized alignment. In *Yearbook of Morphology*, ed. by Geert Booij and Jaap Van Marle, 79-154. Dordrecht: Kluwer.

McCarthy, John J., and Prince, Alan. 1995. Faithfulness and Reduplicative Identity. In *University of Massachusetts Occasional Papers 18: Papers in Optimality Theory*, ed. by Jill Beckman, Laura Walsh Dickey, and Suzanne Urbancyzk, 249-384. Amherst, MA: GLSA, University of Massachusetts.

Merchant, Nazarré. 2008. Discovering underlying forms: Contrast pairs and ranking. PhD. dissertation, Rutgers University, New Brunswick. ROA-964.

Merchant, Nazarré, and Tesar, Bruce. 2008. Learning underlying forms by searching restricted lexical subspaces. In *Proceedings of the Forty-First Conference of the Chicago Linguistics Society (2005), vol. II: The Panels*, 33-47. ROA-811.

Prince, Alan. 1990. Quantitative consequences of rhythmic organization. In *CLS26-II: Papers from the Parasession on the Syllable in Phonetics and Phonology*, ed. by Karen Deaton, Manuela Noske, and Michael Ziolkowski, 355-398. Chicago, IL: Chicago Linguistics Society.

Prince, Alan, and Tesar, Bruce. 2004. Learning phonotactic distributions. In *Constraints in Phonological Acquisition*, eds. René Kager, Joe Pater and Wim Zonneveld, 245-291. Cambridge: Cambridge University Press.

Rosenthall, Sam. 1994. Vowel/glide alternation in a theory of constraint interaction. PhD. dissertation, University of Massachusetts, Amherst.

Tesar, Bruce. 2006a. Faithful contrastive features in learning. *Cognitive Science* 30:863-903.

Tesar, Bruce. 2006b. Learning from paradigmatic information. In *Proceedings of the 36th Meeting of the North East Linguistics Society*, ed. by Christopher Davis, Amy Rose Deal, and Youri Zabbal, 619-638. GLSA. ROA-795.

Tesar, Bruce. 2008. Output-driven maps. Ms., Linguistics Dept., Rutgers University. ROA-956.

Tesar, Bruce, Alderete, John, Horwood, Graham, Merchant, Nazarré, Nishitani, Koichi, and Prince, Alan. 2003. Surgery in language learning. In *Proceedings of the Twenty-Second West Coast Conference on Formal Linguistics*, ed. by G. Garding and M. Tsujimura, 477-490. Somerville, MA: Cascadilla Press. ROA-619.

Tesar, Bruce, and Smolensky, Paul. 1996. Learnability in Optimality Theory (long version). Technical Report JHU-CogSci-96-3, The Johns Hopkins University, Baltimore, MD. ROA-156.