Michael Becker, UMass Amherst
michael@linguist.umass.edu

NECPhon 1 @ UMass
November 10, 2007

Prudent error-driven learning with OT-CC[*]

Highlights:

- Theoretical problem: OT-CC offers a theory of GEN that allows us to derive outputs, but does so at the cost of a candidate set that grows factorially with the number of required repairs.
- Under-reported empirical observation: Children prudently *select* the words that they attempt to produce, avoiding words that would surface too unfaithfully. The kinds of words that children attempt develop gradually, just like the kinds of words that children produce, only a little earlier.
- I propose that the factorial explosion is the cause for children's selection of targets. Choosing targets that involve less repairs is a way to limit the explosion of candidate chains. I offer a version of error-driven learning that incorporates OT-CC's GEN and the need to mitigate the factorial explosion.

1 Error-driven learning with RCD

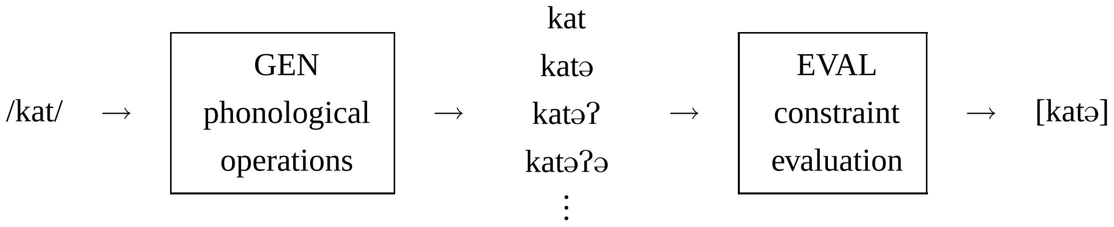
In error-driven learning (Tesar & Smolensky 1998), the learner runs a form through their current grammar, and compares it with the surface form. Example:

(1) $\mathcal{H}_0$: *CODA ≫ DEP
Adult form: [kat]
Current grammar produces: [katə]
[kat] ≠ [katə] → make winner-loser pair
Demote *CODA

[*] Thanks to Outi Bat-El for data and insightful discussion. Thanks to Joe Pater and John McCarthy for feedback and valuable comments. Remaining errors are not worth losing sleep over.

In classic OT, GEN is perfectly well-defined as a function, but it was never fully described as an algorithm (although some success was met in Tesar 1995; Eisner 2003; Riggle 2004). Therefore, we don't know how [katə] is generated as the output of /kat/. All we know is that once [katə] is generated, EVAL will determine that it's the winner.

(2)



2 Producing candidate sets using OT-CC

OT-CC (McCarthy 2007a) is based on Harmonic Serialism, a derivational version of OT (Prince & Smolensky 1993/2004:94-95). In OT-CC, the candidate set is finite and can be generated algorithmically.

2.1 Harmonic improvement

Moreton (2004): In OT, the winner is either completely faithful to the input, or less marked than the input.

(3) Given an input /A/ and an OT grammar, the output is either [A] or some [B] that is less marked than [A].

(4) [A] is the **the fully faithful candidate**, the most harmonic candidate that incurs no faithfulness violations.

(5) The output is the most harmonic candidate. If the output is different from the fully faithful candidate → the output is less faithful and less marked than the fully faithful candidate.

(6)

		/kat/	*CODA	DEP
a.		kat	*!	
b.	☞	katə		*

2.2 OT-CC, Optimality Theory with Candidate Chains

OT-CC (McCarthy 2007a) is a theory of phonology that builds on Moreton's "harmonic improvement", and adds the idea that improving the input is done one step at a time.

In this theory, a candidate is not just a surface form, it is a **chain** of forms that starts with the input and derives the output one step at a time.

(7) Given an input /A/ and a surface form [B], the winner is a
candidate chain such that:

- The first link in the chain is [A]
- The last link in the chain in [B]
- Every link in the chain is more harmonic than the preceding link
- Every link in the chain adds exactly one basic unfaithful phonological operation = one Localized Unfaithful Mapping (LUM)

(8) Example: given the input /kat/ and the grammar *CODA ≫ DEP,
the chain <kat, katə> is the winner, since

- [kat] is the fully faithful candidate
- [katə] is more harmonic than [kat] given the grammar
- [kat] → [katə] adds exactly one LUM: epenthesis of a schwa

(9) Given the input /kat/ and the grammar *CODA ≫ DEP, *VTV

- <kat, katə, kadə> is the winner
- *<kat, katə> is a possible chain (but not the winner)
- **<kat, kadə> is not (epenthesis and voicing done at once)[1]
- **<kat, kad, kadə> is not (not harmonically improving)

The basic phonological operations include epenthesis of one segment, deletion of one segment, and change of one feature. The operations derive the input from the output one step at time.

[1] One star marks a losing chain, two stars mark an ill-formed chain

2.3 Finite candidate sets

OT-CC candidate sets are finite if we are know that:

- Each chain is finitely long

- The number of chains is finite

(10) What are possible ways to make a chain infinitely long?

- Unbounded epenthesis
- Repeating forms in the chain

If these things don't happen, all chains are finitely long.

(11) Chains can't have unbounded epenthesis if we only allow faithfulness and markedness constraints: **<A, AA, AAA, AAAA, ...> (in terms of Moreton (2004), the grammar is "eventually idempotent".

- Markedness constraints can't cause unbounded epenthesis, because they only look at the output. They can only demand epenthesis up to a certain size (e.g. minimal word).

- Faithfulness constraints demand input-output *identity*, so they can't cause epenthesis.

(12) Forms can't repeat in a chain: **<..., A, ..., B, ..., A, ...>

If A follows B in a chain, then A is more harmonic than B
If B follows A in a chain, then B is more harmonic than A
Both statements can't be true.

(13) The number of chains is finite because the number of operations is finite.

Starting with the trivial one-link chain (the faithful candidate), the finite set of operations apply to it, forming a finite number of two-link chains. From those, a finite number of three-link chains will be created, etc., until chains can't get any longer.

3 Learning with a theory of GEN

Now that we have a theory of GEN, we can use it to run forms through the grammar:

(14) $\mathcal{H}_0$: *CODA $\gg$ DEP
Adult form: [kat]
Current grammar produces the winning chain <kat, katə>
[kat] $\neq$ [katə] → make winner-loser pair
Demote *CODA

The problem: When a derivation from the input to the winner involves repairs that don't interact between them, i.e. repairs that can apply in any order, the number of chains grows factorially with the number of repairs.
Example: Children acquiring Hebrew go through a stage where adult *a.vo.(ká.do)* 'avocado' is produced as *(ká.do)*. Deleting three segments can give rise to up to $3! = 6$ possible winning chains:[2]

(15) <a.vo.(ká.do), vo.(ká.do), o.(ká.do), (ká.do)>
<a.vo.(ká.do), vo.(ká.do), v.(ká.do), (ká.do)>
<a.vo.(ká.do), a.o.(ká.do), o.(ká.do), (ká.do)>
<a.vo.(ká.do), a.o.(ká.do), a.(ká.do), (ká.do)>
<a.vo.(ká.do), av.(ká.do), a.(ká.do), (ká.do)>
<a.vo.(ká.do), av.(ká.do), v.(ká.do), (ká.do)>

Generally, n unordered repairs give rise to $n!$ winning chains[3].
My proposal: Children accept the harsh reality of factorial explosion, and mitigate the problem by avoiding derivations that will require too many chains.

[2] The number of steps involved in a given derivation is a theoretical matter. Specifically for deletions that aren't crucially ordered, I assume that deletion happens one segment at a time. See however, McCarthy (2007c), for a proposal that segments get deleted one feature at a time, and McCarthy (2007b), who proposes that any amount of deletion can happen in one step of the derivation.

[3] There can also be up to $(n-1)! + (n-2)! + ...$ losing chains, but those add up to less than $n!$. If repairs are ordered, the number of chains grows linearly (n repairs give rise to n chains). This means that the number of chains doesn't necessarily correlate with the depth of the derivation, since processes that are crucially ordered do not increase the number of chains.

4 Children's target selection

In early stages of acquisition, children pare down all their words to a single strong syllable (S), then allow words with a following weak syllable (SW, trochee). This is a commonly reported pattern cross-linguistically (e.g. Pater 1997; Vihman et al. 1998).

(16) Shaxar's monosyllabic **productions** from polysyllabic targets (Adam & Bat-El 2007)

Period	Age	SW		WS	
		target	%mono-σ	target	%mono-σ
I	1;02.00-1;03.05	9	56%	7	86%
II	1;03.14-1;04.24	43	14%	29	48%
III	1;05.04-1;05.08	39	10%	29	28%
IV	1;05.15-1;05.29	35	11%	57	28%
V	1;06.02-1;06.20	49	2%	55	29%
VI	1;06.26	26	0%	53	11%
VII	1;07.02-1;07.09	51	4%	99	4%

For Shaxar, producing SW targets rarely involves deletion after period I, but producing WS targets commonly involves deletion well into period V. Adam & Bat-El's (2007) observation: Shaxar shows the gradual acceptance of WS words not only in his productions, but also in his attempts (see also Schwartz 1988).

(17) Shaxar's **attempts** at major class words (Adam & Bat-El 2007)

Period	Age	targets	SW	WS	% WS
I	1;02.00-1;03.05	16	9	7	44%
II	1;03.14-1;04.24	72	43	29	40%
III	1;05.04-1;05.08	68	39	29	43%
IV	1;05.15-1;05.29	92	35	57	62%
V	1;06.02-1;06.20	104	49	55	53%
VI	1;06.26	79	26	53	67%
VII	1;07.02-1;07.09	150	51	99	66%

This child gradually attempts subsets of Hebrew that more closely resemble the adult language, which has WS (final stress) in ∼75% of its major class words (Bolozky & Becker 2006).

Classical OT naturally captures the constraints on the child **productions**: MARKEDNESS $\gg$ MAX causes as much deletion as needed to satisfy markedness. But this does not capture the constraints on **attempts**, as there is no way to express the cost of deletion.
In OT-CC, more deletion causes more chains, so there is a cost to massive deletion. The explosion of chains can represent a measure of the processing load of a given input-output mapping.

5 Prudent learning

The learning algorithm:

(18)

1. Prepare the data for processing

   (a) Accept a batch of input forms.

   (b) Run each form through the current grammar, with GEN turned off, so only trivial single-link chains are created. Syllaby and run through EVAL.

   (c) Order the forms in the current batch such that the least marked forms are processed first, i.e. in decreasing harmony.

2. Apply prudent error-driven learning:

   (a) Select the first form from the batch, run through GEN and EVAL.

   (b) Compare the winner with the adult form. If not identical, add winner-loser pair to the Support (or Cache, see Tessier 2007), and run BCD.

   (c) If GEN created more than x chains, go back to step 1. Otherwise, keep going until all the batch is processed.

Example:
Suppose a child picks out the following nouns from a stream of Hebrew they're exposed to: *ba.ná.na* 'banana', *mit.ri.yá* 'umbrella', *gé.ʃem* 'rain'. The child constructs a faithful candidate from each form in the batch:

(19)

	*LAPSE	INITIAL-σ́	MAX
a. ba.ná.na		*	
b. mit.ri.yá	*	*	
c. gé.ʃem			

With the current grammar, the batch is reordered on a scale of harmony: *géʃem*, *banána*, *mitriyá*. These are run through GEN and EVAL:

(20) 1. *géʃem* is processed, one chain created: <*géʃem*>.

2. *banána* is processed, several chains created, e.g. <*ba.ná.na*, *bná.na*, *bá.na*>, <*ba.ná.na*, *a.ná.na*, *ná.na*>, etc. W-L pair created:

	INITIAL-σ́	MAX
ba.ná.na ≻ **bá.na*, **ná.na*	L	W

INITIAL-σ́ demoted below MAX.

3. *mit.ri.yá* is known to be worse than *ba.ná.na*, which involved several chains, so it may not be attempted at all, and it's back to step 1, to collect more data from the environment. If it is attempted, it will likely involve more chains than *ba.ná.na* did.

Given the initial grammar of *LAPSE ≫ INITIAL-σ́ ≫ MAX, *mit.ri.yá* reduces to *yá*, i.e. deletion of 5 segments, for up to $5! = 120$ chains. There are two possible ways in which processing *ba.ná.na* before *mit.ri.yá* can be beneficial:

(21) The number of chains created when deriving *ba.ná.na* can warn about the cost of processing *mit.ri.yá*, causing avoidance.

(22) If the processing of *ba.ná.na* causes demotion of INITIAL-σ́ below MAX, then the input *mit.ri.yá* will only reduce to *ti.yá*, not *yá*. Deleting three segments rather than five means that the number of chains will be closer to $3! = 6$ rather than $5! = 120$.

This scenario is simplified, since real children attempt increasingly complex forms very gradually. Plain error-driven learning can learn too fast, skipping stages that children take a while to go through (Tessier 2007, and see also a solution in terms of Harmonic Grammar, Jesney & Tessier 2007, *later today*).

6 Conclusions

- I presented data about the under-reported phenomenon of children's target selection, pointing out that children avoid words whose phonological structure would require too much deviation from the adult form.

- I proposed *prudent learning*, an error-driven learning algorithm that derives target selection from the cost of chain explosion in OT-CC.

- Consequently, chain explosion in OT-CC is no longer viewed as a problem, but rather as a desired property that supplies a formal expression for the observed phenomenon.

References

Adam, Galit & Outi Bat-El (2007). The trochaic bias is universal: Evidence from Hebrew. Handout from Generative Approaches to Language Acquisition, Barcelona.

Bolozky, Shmuel & Michael Becker (2006). Living Lexicon of Hebrew Nouns. Ms. UMass Amherst.

Eisner, Jason (2003). Simpler and more general minimization for weighted finite-state automata. In *Proceedings of the Joint Meeting of the Human Language Technology Conference and the North American Chapter of the Association for Computational Linguistics (HLT-NAACL)*. Edmonton, 64–71.

Jesney, Karen & Anne-Michelle Tessier (2007). Re-evaluating learning biases in harmonic grammar. In Michael Becker (ed.) *papers in theoretical and computational phonology, UMOP 36*, GLSA.

McCarthy, John J. (2007a). *Hidden Generalizations: Phonological Opacity in Optimality Theory*. London: Equinox Publishing Company.

McCarthy, John J. (2007b). The serial interaction of stress and syncope. Ms. UMass Amherst.

McCarthy, John J. (2007c). Slouching toward optimality. *Phonological Studies (Journal of the Phonological Society of Japan)* **7**.

Moreton, Elliott (2004). Non-computable functions in optimality theory. In John J. McCarthy (ed.) *Optimality Theory in Phonology*, Blackwell. 141–163.

Pater, Joe (1997). Minimal violation and phonological developement. *language acquisition* **6**. 201–253.

Prince, Alan & Paul Smolensky (1993/2004). *Optimality Theory: Constraint Interaction in Generative Grammar*. Oxford: Blackwell. [ROA-537].

Riggle, Jason (2004). *Generation, Recognition, and Learning in Finite State Optimality Theory*. Ph.D. dissertation, UCLA.

Schwartz, Richard G. (1988). Phonological factors in early acquisition. In Michael D. Smith & John L. Locke (eds.) *The Emergent Lexicon: The Child's Development of a Linguistic Vocabulary*, Academic Press, New York, chap. 7. 185–222.

Tesar, Bruce (1995). *Computational Optimality Theory*. Ph.D. dissertation, University of Colorado at Boulder. ROA 90-0000.

Tesar, Bruce & Paul Smolensky (1998). Learnability in optimality theory. *Linguistic Inquiry* **29**. 229–268.

Tessier, Anne-Michelle (2007). *Biases and stages in phonological acquisition*. Ph.D. dissertation, University of Massachusetts, Amherst.

Vihman, Marilyn, Rory DePaolis & Barbara Davisin (1998). Is there a "trochaic bias" in early word learning? Evidence from infant production in english and french. *Child Development* **69**. 935–949.