Learning an Optimality Theoretical grammar with a structured candidate set

Tamás Bíró

Work presented developed at:
Humanities Computing
CLCG
University of Groningen

Present affiliation:
Theoretical Linguistics
ACLC
University of Amsterdam

birot@nytud.hu

Bielefeld, January 10, 2007

Learning (an Optimality Theoretical grammar with a structured candidate set)

Tamás Bíró

Work presented developed at:
Humanities Computing
CLCG
University of Groningen

Present affiliation:
Theoretical Linguistics
ACLC
University of Amsterdam

`birot@nytud.hu`

Bielefeld, January 10, 2007

*Learning (an Optimality Theoretical grammar) (with a structured candidate set)

Tamás Bíró

Work presented developed at:
Humanities Computing
CLCG
University of Groningen

Present affiliation:
Theoretical Linguistics
ACLC
University of Amsterdam

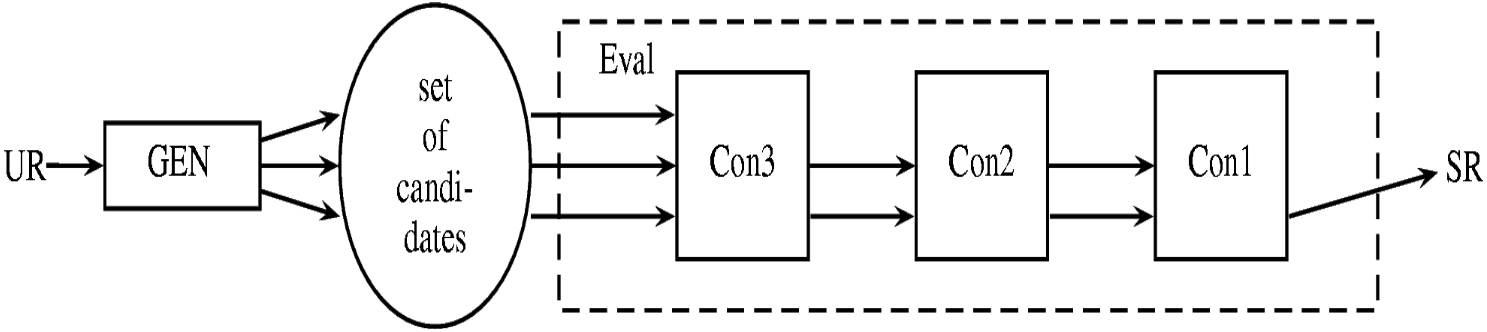
birot@nytud.hu

Bielefeld, January 10, 2007

Overview

- Optimality Theory (OT) in a nutshell
- Simulated Annealing for Optimality Theory (SA-OT)
- Examples
- Learnability?
- Conclusion

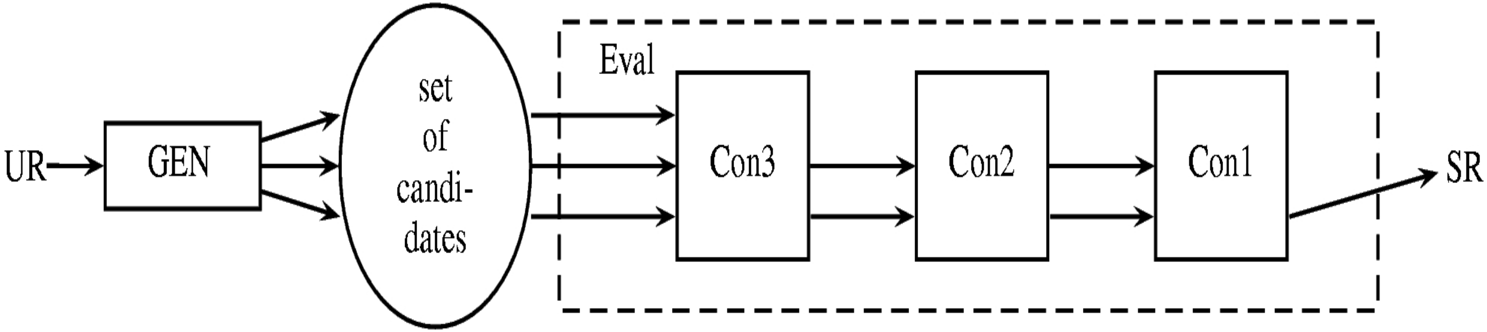
Optimality Theory in a nutshell



OT tableau: search the best candidate w.r.t lexicographic ordering
(cf. $abacus$, $abolish$,..., $apple$,..., $zebra$)

	c_n	c_{n-1}	...	c_{k+1}	c_k	c_{k-1}	c_{k-2}	...
w	2	0		1	2	3	0	
w'	2	0		1	3 !	1	2	
w"	3 !	0		1	3	1	2	

Optimality Theory in a nutshell

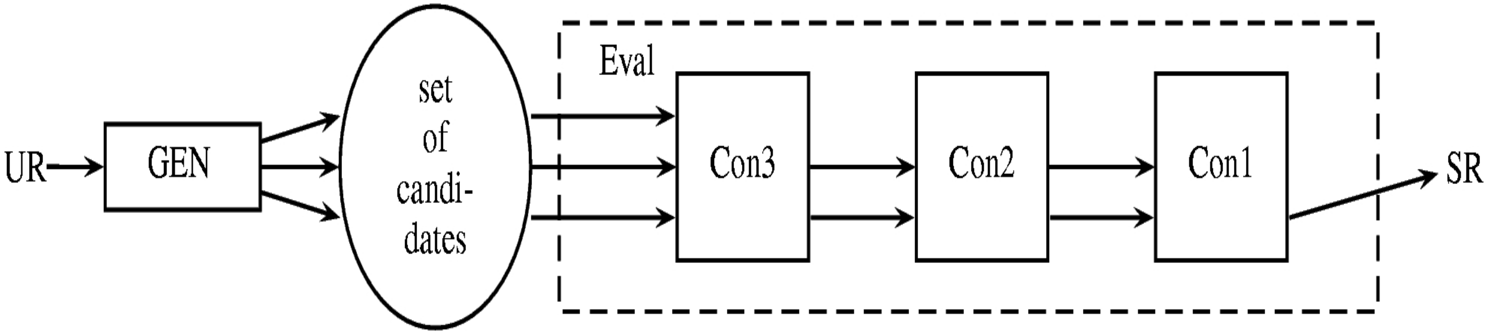


- Pipe-line *vs.* optimize the Eval-function

- Gen: $UR \mapsto \{w | w \text{ } is \text{ } a \text{ } candidate \text{ } corresponding \text{ } to \text{ } UR\}$

  E.g. assigning Dutch metrical foot structure & stress:
  fototoestel $\mapsto$ {fototoe(stél), (fotó)(tòestel), (fó)to(toestèl),...}

Optimality Theory: an optimisation problem



$$UR \mapsto \{w | w \textit{ is a candidate corresponding to } UR\}$$
$$E(w) = \Big(C_N(w), C_{N-1}(w), ..., C_0(w)\Big) \in \mathbb{N}_0^{N+1}$$
$$SR(UR) = \text{argopt}_{w \in Gen(UR)} E(w)$$

Optimisation: with respect to lexicographic ordering

OT is an optimization problem

The question is:
How can the optimal candidate be found?

- Finite-State OT (Ellison, Eisner, Karttunen, Frank & Satta, Gerdemann & van Noord, Jäger...)
- chart parsing (dynamic programing) (Tesar & Smolensky; Kuhn)

These are perfect for language technology. But we would like a psychologically adequate model of linguistic performance (e.g. errors): Simulated Annealing.

How to find optimum: gradient descent

```
w := w_init ;
Repeat
        Randomly select w' from the set Neighbours(w);
        Delta :=  E(w') - E(w) ;
        if   Delta < 0   then   w := w' ;
        else
               do nothing

        end-if


Until stopping condition = true

Return w          # w is an approximation to the optimal solution
```

The Simulated Annealing Algorithm

```
w := w_init ;       t := t_max  ;
Repeat
        Randomly select w' from the set Neighbours(w);
        Delta :=  E(w') - E(w) ;
        if   Delta < 0   then   w := w' ;
        else
               generate random r uniformly in range (0,1) ;
               if   r < exp(-Delta / t)   then   w := w' ;
        end-if

   t := alpha(t)                           # decrease t
Until stopping condition = true

Return w          # w is an approximation to the optimal solution
```

Gradient descent for OT?

- McCarthy (2006): *persistent OT* (*harmonic serialism*, cf. Black 1993, McCarthy 2000, Norton 2003).

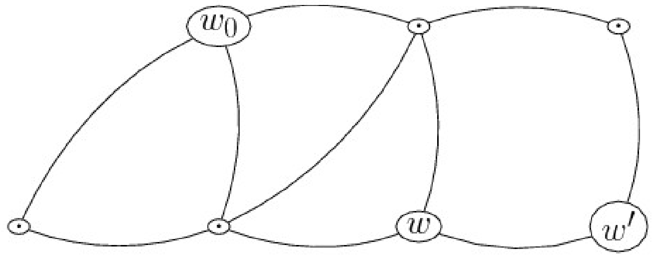
- Based on a remark by Prince and Smolensky (1993/2004) on a "restraint of analysis" as opposed to "freedom of analysis".

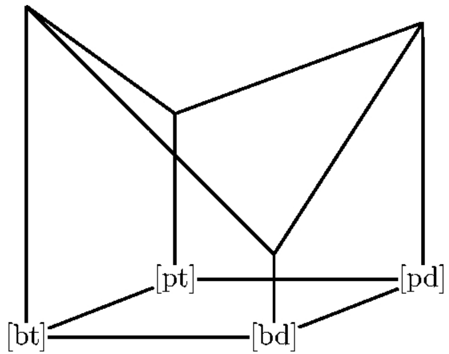
- Restricted Gen $\rightarrow$ Eval $\rightarrow$ Gen $\rightarrow$ Eval $\rightarrow$... (n times).

- Gradual progress toward (locally) max. harmony.

- Employed to simulate traditional derivations, opacity.

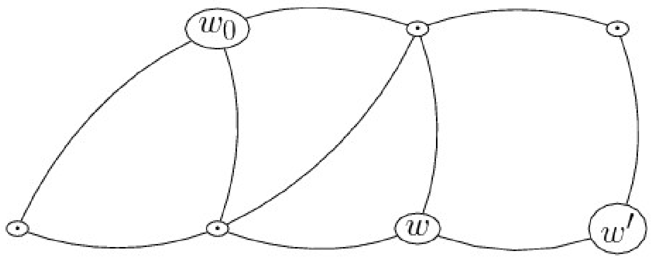
Simulated Annealing for OT

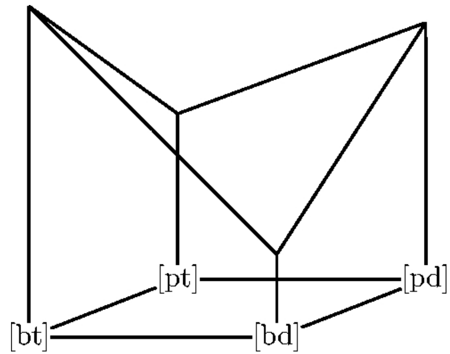




- Neighbourhood structure on the candidate set.
- Landscape's vertical dimension = harmony; random walk.
- If neighbour more optimal: move.
- If less optimal: move in the beginning, don't move later.

Simulated Annealing for OT





- Neighbourhood structure → local optima.
- System can get stuck in local optima: alternation forms.
- Precision of the algorithm depends on its speed (!!).
- Many different scenarios.

Domains for temperature and constraints

- Temperature: $T = \langle K_T, t \rangle \in \mathbb{Z} \times \mathbb{R}^+$ (or "$\mathbb{Z}$" $\times \mathbb{R}^+$).
- Constraints associated with domains of K_T:

	$--$	C_0	C_1	C_2
. . .	$K = -1$	$K = 0$	$K = 1$	$K = 2$
. . .	... 0.5 1.0 1.5 2.0 2.5	... 0.5 1.0 1.5 2.0 2.5	... 0.5 1.0 1.5 2.0 2.5	... 0.5 1.0 1.5 2.0 2.5

Rules of moving

RULES OF MOVING from w to w'
at temperature $T = \langle K_T, t \rangle$:

If w' is better than w: move! $P(w \to w'|T) = 1$

If w' loses due to fatal constraint C_k:

If $k > K_T$: don't move! $P(w \to w'|T) = 0$

If $k < K_T$: move! $P(w \to w'|T) = 1$

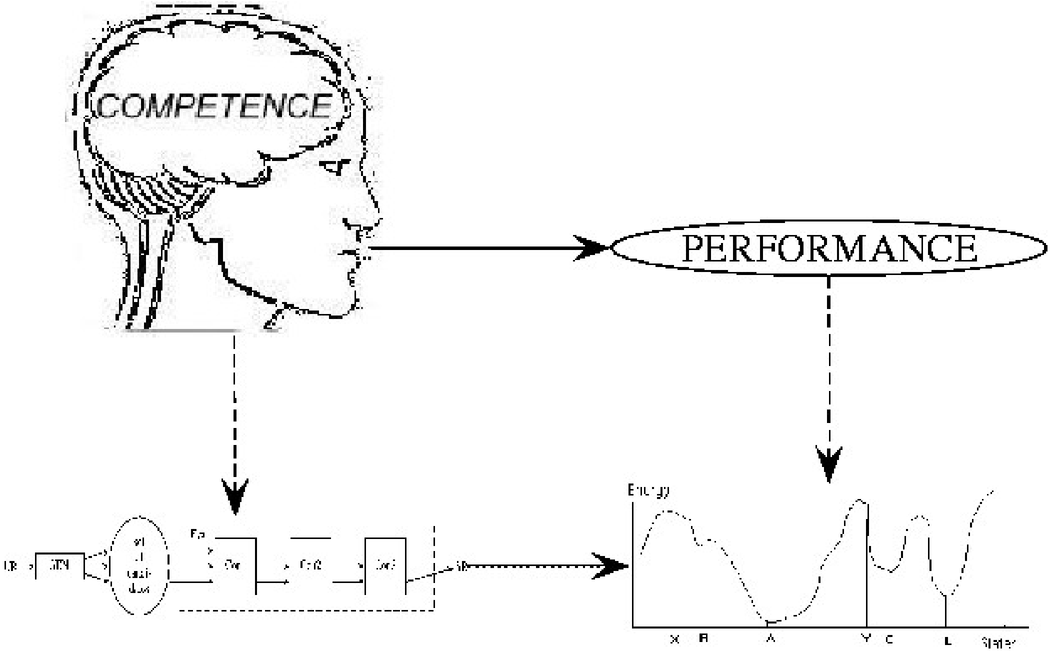
If $k = K_T$: move with probability

$$P = e^{-(C_k(w') - C_k(w))/t}$$

The SA-OT algorithm

```
w := w_init ;
for K = K_max to K_min step K_step
      for t = t_max to t_min step t_step
             CHOOSE   random w' in neighbourhood(w) ;
             COMPARE  w' to w:  C := fatal constraint
                                d := C(w') - C(w);
             if d <= 0 then w := w';
             else               w := w' with probability
                          P(C,d;K,t) = 1          , if C < K
                                     = exp(-d/t) , if C = K
                                     = 0          , if C > K
      end-for
end-for
return w
```

SA-OT as a model of linguistic performance



Optimality Theory: a model of competence

Simulated Annealing: a model of performance

Proposal: three levels

Level	its product	its model	the product in the model
Competence in narrow sense: static knowledge of the language	grammatical form	standard OT grammar	globally optimal candidate
Dynamic language production process	acceptable or attested forms	SA-OT algorithm	local optima
Performance in its outmost sense	acoustic signal, etc.	(phonetics, pragmatics)	??

The Art of Using Simulated Annealing Optimality Theory

- Take a traditional OT model
- Add *convincing* neighbourhood structure to candidate set
- Local (non-global) optima = alternation forms
- Run simulation (e.g., `http://www.let.rug.nl/~birot/sa-ot`):
  - Slowly: likely to return only the grammatical form
  - Quickly: likely to return local (non-global) optima

Parameters of the algorithm

- t_{step} (and t_{max}, t_{min})
- K_{max} (and K_{min})
- K_{step}
- w_0 (inital candidate)
- Topology (neighbourhood structure)
- Constraint hierarchy

How to make the topology convincing?

A connected (weighted) "graph"; universal;...

- Observation-driven strategies:
  - Many phenomena in many languages
    or even better: cross-linguistic typologies
  - Based on existing theories based on cross-linguistic observations (cf. Hayes's *metrical stress theory*)

- Theory-driven strategies:
  - Principles (e.g. minimal set of basic transformations)
  - Psycholinguistically relevant notions of similarity, etc.

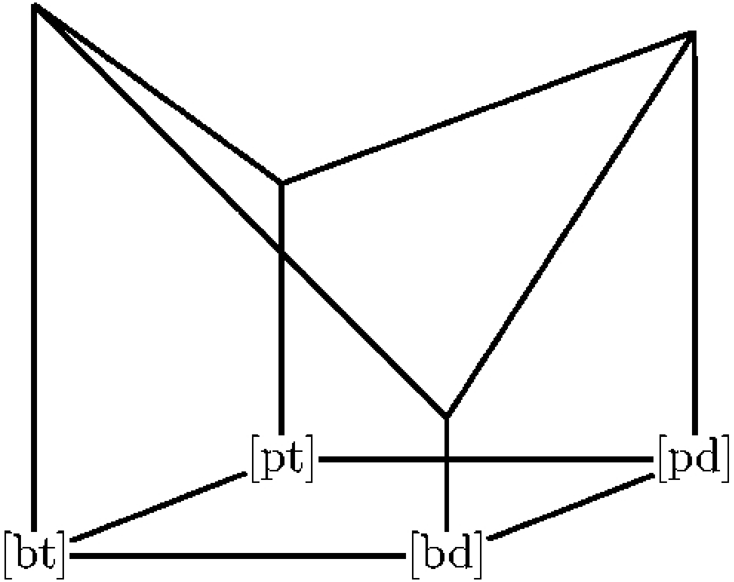
Example: Fast speech: Dutch metrical stress

fo.to.toe.stel 'camera'	*uit.ge.ve.rij* 'publisher'	*stu.die.toe.la.ge* 'study grant'	*per.fec.tio.nist* 'perfectionist'
susu	ssus	susuu	usus
fó.to.tòe.stel fast: *0.82* slow: *1.00*	*ùit.gè.ve.ríj* fast: *0.65* / **0.67** slow: *0.97* / **0.96**	*stú.die.tòe.la.ge* fast: *0.55* / **0.38** slow: *0.96* / **0.81**	*per.fèc.tio.níst* fast: *0.49* / **0.13** slow: *0.91* / **0.20**
fó.to.toe.stèl fast: *0.18* slow: *0.00*	*ùit.ge.ve.ríj* fast: *0.35* / **0.33** slow: *0.03* / **0.04**	*stú.die.toe.là.ge* fast: *0.45* / **0.62** slow: *0.04* / **0.19**	*pèr.fec.tio.níst* fast: *0.39* / **0.87** slow: *0.07* / **0.80**

Simulated / **observed** (Schreuder) frequencies.

In the simulations, $T_{step} = 3$ used for fast speech and $T_{step} = 0.1$ for slow speech.

Example: Irregularities



- Local optimum that is not avoidable.

Example: string-grammar 1

- *Candidates:* $\{0, 1, ..., P-1\}^L$
  E.g. $(L = P = 4)$: 0000, 0001, 0120, 0123,... 3333.

- *Neighbourhood structure:* w and w' neighbours iff one basic step transforms w to w'.

- Basic step: change exactly one character $\pm 1, \mod P$ (cyclicity).

- Each neighbour with equal probability.

Example: string-grammar 2

Markedness Constraints ($w = w_0 w_1 ... w_{L-1}$, $0 \leq n < P$):

- No-n: $*n(w) := \sum_{i=0}^{L-1} (w_i = n)$
- No-initial-n: $*\textsc{Initial}n(w) := (w_0 = n)$
- No-final-n: $*\textsc{Final}n(w) := (w_{L-1} = n)$
- Assimilation $\textsc{Assim}(w) := \sum_{i=0}^{L-2} (w_i \neq w_{i+1})$
- Dissimilation $\textsc{Dissim}(w) := \sum_{i=0}^{L-2} (w_i = w_{i+1})$

Example: string-grammar 3

- Faithfulness to UR σ:

$$\textsc{Faith}_{\sigma}(w) = \sum_{i=0}^{L-1} d(\sigma_i, w_i)$$

where $d(a,b) = \min(|a-b|, |b-a|)$

(binary square, feature-combination?)

Example: string-grammar 4

$L = P = 4$, $T_{max} = 3$, $T_{min} = 0$, $K_{step} = 1$.

Each of the 256 candidates used 4 times as w_0.

Grammar:

*0 $\gg$ Assim $\gg$ Faithf$_{\sigma=0000}$ $\gg$ *Init1 $\gg$*Init0 $\gg$ *Init2 $\gg$ *Init3 $\gg$ *Fin0 $\gg$ *Fin1 $\gg$ *Fin2 $\gg$ *Fin3 $\gg$ *3 $\gg$ *2 $\gg$ *1 $\gg$ Dissim

Globally optimal form: 3333

Many other local optima, e.g.: 1111, 2222, 3311, 1333, etc.

Example: string-grammar 5

Output frequencies for different T_{step} values:

output	0.0003	0.001	0.003	0.01	0.03	0.1
1111	0.40	0.40	0.36	0.35	0.32	0.24
3333	0.39	0.39	0.41	0.36	0.34	0.21
2222	0.14	0.14	0.15	0.18	0.19	0.17
3311	0.04	0.04	0.04	0.05	0.06	0.05
1133	0.03	0.04	0.04	0.04	0.05	0.04
others	–	–	–	–	0.04	0.29

Learnability?

Why? What to learn?? Learning: for whom?

- For a linguist: find parameters exactly matching the observations to make the publication nice.
- For language technology (create a complex, but nicely working system; but OT not very used in NLP).
- For a cognitive scientist: a language acquisition model.

Parameters of the algorithm (reminder)

- t_{step} (and t_{max}, t_{min})
- K_{max}
- K_{step}
- Topology (neighbourhood structure)
- w_0 (initial candidate)
- Constraint hierarchy

Learnability?

What to learn??

- Find values for the parameters of the algorithm (e.g. T_{step}) that return the same frequencies:
  - Nice for the linguist;
  - But this is learning performance.

- Learn the underlying OT grammar, "despite" perf. errors:
  - Exact quantitative match is not the goal;
  - This is learning competence, also a relevant issue.

A first trial for learning

- Observed forms must be local optima.
- Guess: more harmonic form has higher frequency.
  - This is what SA-OT would like to achieve but it doesn't. What are the consequences?
  - Cf. Coetzee's OT proposal.

I employed *Recursive Constraint Demotion* (RCD), an off-line standard learning algorithm.

Result: a grammar that produces a superset of forms.

Preliminary remarks on learning / acquisition

- No negative evidence = hard to make something not a local optimum.
- Already infants "measure" frequencies (e.g., Gervain). Even at pre-production age: off-line algorithm makes sense.
- Children's forms: superset of adults' forms.

Future work: gradually refine the grammar learned by RCD to reach adult grammar (existing on-line algorithms: EDCD, GLA).

What does SA-OT offers to standard OT?

- A new approach to account for variation:
  - Non-optimal candidates also produced (cf. Coetzee);
  - As opposed to: more candidates with same violation profile; more hiararchies in a grammar.

- A topology (neighbourhood structure) on the candidate set.

- Additional ranking arguments (cf. McCarthy 2006) $\longrightarrow$ learning algorithms (in progress).

- Arguments for including losers (never winning candidates).

Summary of SA-OT

- Implementation: can OT be useful to language technology? is OT cognitively plausible?
- A model of variation / performance phenomena.
- *Errare humanum est* – a general cognitive principle: the role of heuristics.
- Learning is being worked on.
- Demo at `http://www.let.rug.nl/~birot/sa-ot`.

Thank you for your attention!

Tamás Bíró

birot@nytud.hu