Simulated Annealing for OT:
A performance model for phonology

Tamás Bíró

Work presented developed at:
Humanities Computing
CLCG
University of Groningen

Present affiliation:
Theoretical Linguistics
ACLC
University of Amsterdam

birot@nytud.hu

Workshop on Computing and Phonology – Groningen, Dec. 8, 2006

Overview

- *Competence* vs. *performance* in traditional generative linguistics
- The SA-OT Algorithm
- Evaluating the SA-OT Algorithm

Competence vs. *performance*

The Chomskyan dichotomy:

"Linguistic theory is concerned primarily with an ideal speaker-listener, in a completely homogeneous speech-community, who knows its language perfectly and is unaffected by such grammatically irrelevant conditions as memory limitations, distractions, shifts of attention and interest, and errors (random or characteristic) in applying his knowledge of the language in actual performance. ... We thus make a fundamental distinction between *competence* (the speaker-hearer's knowledge of his language) and *performance* (the actual use of language in concrete situations)."
(Chomsky: *Aspects*, 1965, pp. 3-4; cited in FRW, p. 27)

Competence vs. *performance*

The Chomskyan dichotomy:

- memory limitations,
- distractions,
- shifts of attention and interest,
- and errors (random or characteristic)

in applying his knowledge of the language in actual performance. ...

Competence vs. *performance*

Or what about:

- when losing one's teeth
- fatigue, alcohol
- speech rate, memory limitation
- conditional corpus frequency of forms
- gradient grammaticality of forms

Possible arguments

- Presence of factors that are seen as linguistic in other context.
- Hard constraints in one language may appear as statistical preference in another language (in general / under certain circumstances).
- Functional motivations for features of grammars.

The meta-scientific side

Whenever you do not know how to approach the phenomenon, you want to argue for it to be irrelevant.

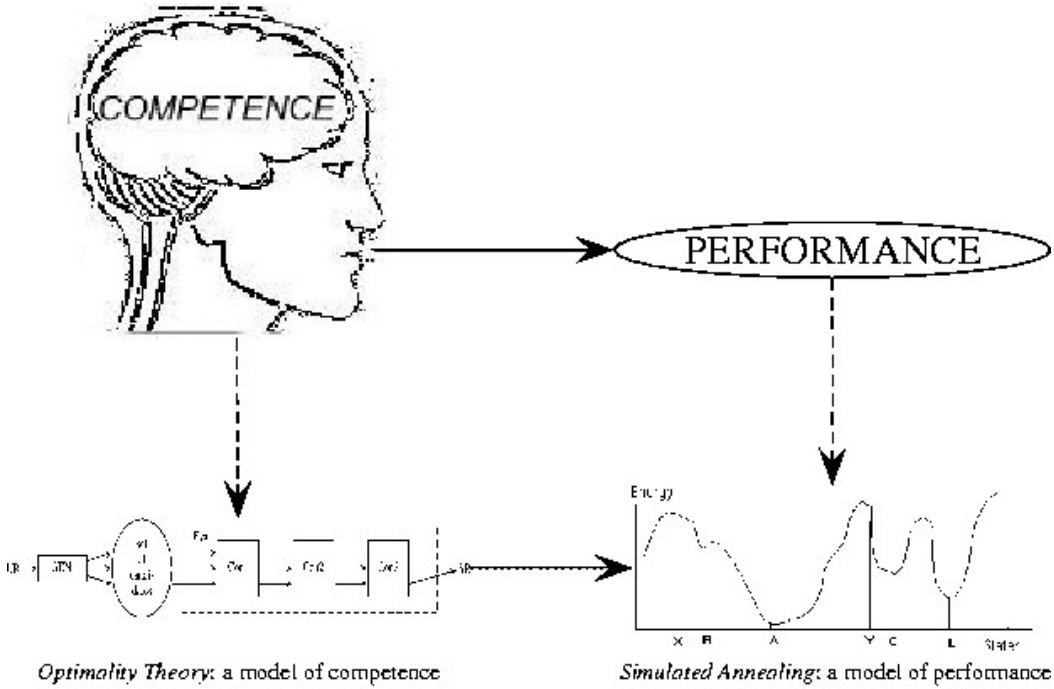
But if you have a neat model for a phenomenon, you want to deal with that phenomenon.

The meta-scientific side

Given the grammar of a language, one can study the use of the language statistically in various ways; and the development of probabilistic models for the $use\ of\ language$ *[emphasis added – T. B.] (as distinct from the syntactic structure of language) can be quite rewarding ...*

One might seek to develop a more elaborate relation between statistical and syntactic structure than the simple order of approximation model we have rejected. I would certainly not care to argue that any such relation is unthinkable, but I know of no suggestion to this effect that does not have some obvious flaws. (Chomsky: Syntactic structures, 1957, p. 17, n. 4; cited in FRW, p. 28.)

Compromise?



Optimality Theory: a model of competence

Simulated Annealing: a model of performance

FRW p. 44.

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)	??

OT is an optimization problem

- Let's have OT as a competence-model.
- Then, what would model the dynamic language production process?
- An algorithm that finds the candidate that is predicted to be the grammatical form = the optimal candidate of the candidate set

Hence, the task is find an optimization algorithm.

- Finite-State OT, chart parsing (dynamic programing)?
- We need and adequate model of performance: e.g. errors

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
```

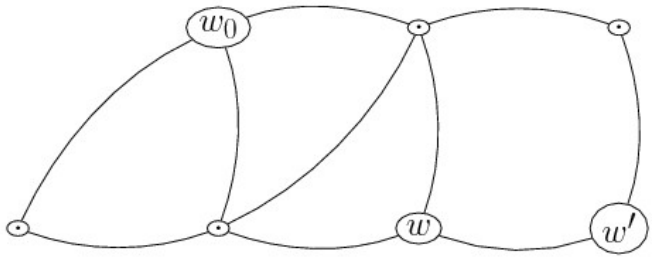
The Simulated Annealing Algorithm

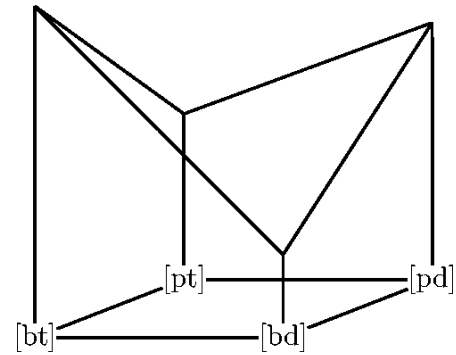
```
w := w_init ;       t := t_max  ;
Repeat
        Randomly select w' from the set Neighbours(w);
        Delta :=  E(w') - E(w) ;
        if   Delta < 0   then   w := w' ;
        else
               generate random r uniformly in range (0,1) ;
               if   r < exp(-Delta / t)   then   w := 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 max. harmony candidate.
- Employed to simulate traditional derivations, opacity.

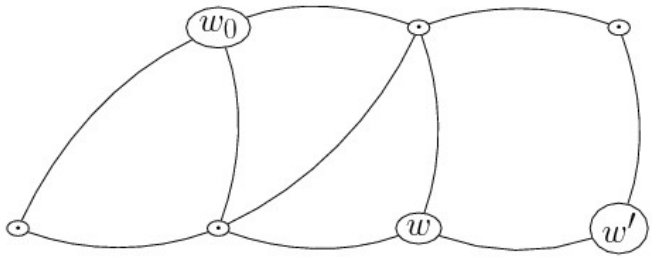
Simulated Annealing for OT

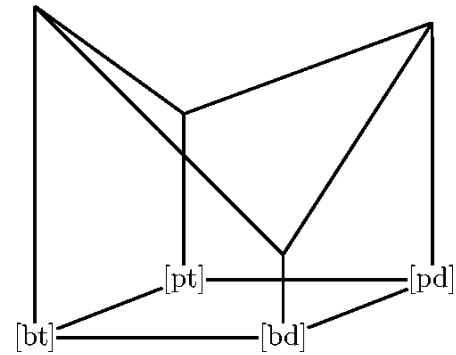




- Neighbourhood structure on the candidate set
- Landscape's vertical dimension = harmony; random walk
- If neighbour more optimal: move.
- If less optimal: move in the beginning, don't move later

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

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! $P(w \rightarrow w'|T) = 1$

If w' loses due to fatal constraint C_k:

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

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

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
```

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 (some more technical details needed...):
  - Slowly: likely to return only the grammatical form
  - Quickly: likely to return local (non-global) optima

A note on topologies

Three possible future strategies:

- Argue that the topology follows from theory
- Topology required cross-linguistically
- Topology follows from general principles, e.g. psychologically motivated similarity measures, or minimal set of basic operations.

- Similarity on surface or similarity in structure?

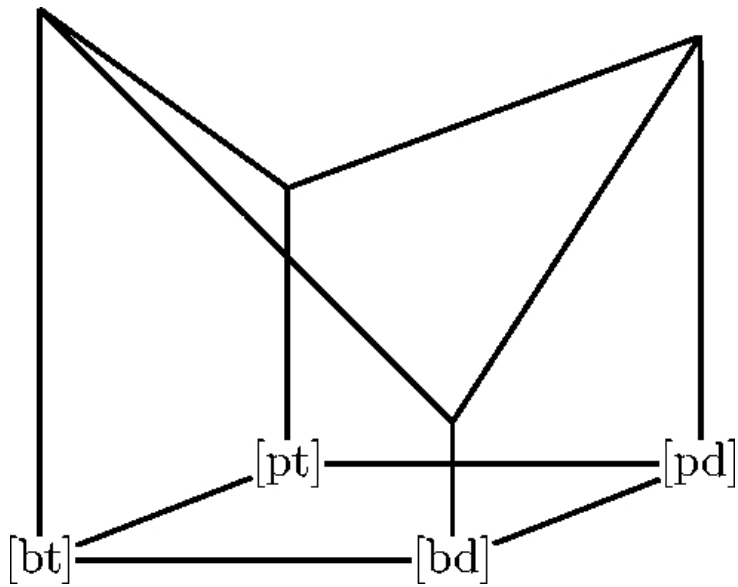
- Fine tuning quantitative side?

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	*úit.gè.ve.rìj* fast: 0.65 slow: 0.97	*stú.die.tòe.la.ge* fast: 0.55 slow: 0.96	*per.féc.tio.nìst* fast: 0.49 slow: 0.91
fó.to.toe.stèl fast: 0.18 slow: 0.00	*úit.ge.ve.rìj* fast: 0.35 slow: 0.03	*stú.die.toe.là.ge* fast: 0.45 slow: 0.04	*pér.fec.tio.nìst* fast: 0.39 slow: 0.07

Data from *M. Schreuder* & *D. Gilbers* (2004): The Influence of Speech Rate on Rhythm Patterns; *M. Schreuder*: Prosodic Processes in Language and Music, 2005; FRW p. 155.

Irregularities



Cf. FRW Chapter 6

- Local optimum that is not avoidable.

More parameters

/pd/	Dep	RegrAss	ProgrAss	*Voice
[pd]		*	*	*
[pt]		*		
[bt]		*	*	*
[bd]			*	**
[p$^n d$]	n			*
[p$^n d$t]	n			
[b$^n d$t]	n			*
[b$^n d$d]	n			**

(FRW 6.5 improved)

Parameters of the algorithm

- t_{step} (and t_{max}, t_{min})
- K_{max}
- K_{step}
- Topology (neighbourhood structure)

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 (FRW Chapter 7, cf. McCarthy 2006)
- 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.
- Demo at http://www.let.rug.nl/ birot/sa-6t

Thank you for your attention!