Efficient communication with a ponderous brain

Performance errors in Optimality Theory

Tamás Biró

birot@nytud.hu

Radboud University Nijmegen

January 22, 2008

Overview

- Competence and performance: errors, irregularities, learning and communication
- Optimality Theory (OT) as an optimization problem
- The Simulated Annealing for OT Algorithm (SA-OT)
- Example: string grammar
- Learning SA-OT
- Concluding remarks

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)

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

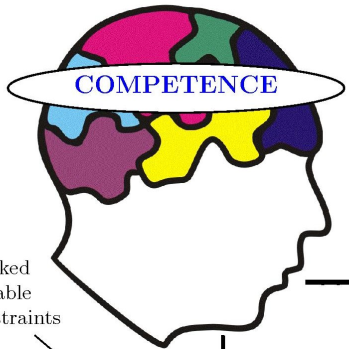
Given the grammar of a language, one can study the use of the language statistically in various ways; and the development of probabilistic models for the $use\ of\ language$ [emphasis added – T. B.] (as distinct from the syntactic structure of language) can be quite rewarding ...

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)

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)	??

Competence vs. *performance*



A grammar is a Harmony function on the candidate set, defined by the ranked constraints.
Global optimum: more harmonic than all other candidates.
Local optimum: more harmonic than its neighbours.

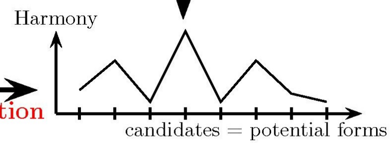
PERFORMANCE

Ranked violable constraints

Candidates (potential forms)

	/aat/	NoCoda	Parse	Onset
☞	[a.a<t>]		*	**
	[a.at]	*!		**
	[<a>at]	*!	*	*

implementation



Optimality Theory
grammar
competence model
grammatical form = ☞ (globally) optimal candidate

SA-OT
implementation
performance model
produced forms = globally or locally optimal candidates

Competence vs. *performance*

Paul Smolensky:

"... *competence* can be understood as an idealization of actual behavior—*performance*—in which we have removed the effects of limitations on computational resources: generally speaking, space, time, and precision." (Smolensky et al.: *The Harmonic Mind*, 2006, vol. 1, p. 228.)

Competence = grammar: is a *function*

$$input \mapsto correct\ output/parse/struct.\ description$$

Performance: *algorithm* that finds it. Or doesn't.

Competence: performance run infinitely slowly.

Errors and irregularities

Divergence between competence and performance:

- ☞ *Grammatical forms* = globally optimal
- ! *Performance errors*: frequency diminishes at slow (careful) production (as in traditional simulated annealing).
- ∼ *Irregularities*: frequency does not diminish at slow (careful) production (due to *strict domination*).

Not all forms in a language need be analysed as grammatical!

Adequacy of a performance model

Performance model: an algorithm that realizes (implements) the grammar (*i.e.*, the model of competence), which

- usually finds the form grammatical w.r.t. grammar (☞),
- but also makes the same errors as humans do,
- with a similar frequency
- under various conditions (speech rate, style, etc.).

Moreover, *runtime* and *complexity* of algorithm is plausible.

Consequences for language acquisition

Child is exposed to teacher's *performance distribution* (derived from teacher's competence + production mechanism):

- Grammatical forms, performance errors and irregular forms
- produced with different frequencies
- under various circumstances (time pressure, stylistic and sociolinguistic variations, etc.—parameters of SA-OT)

Can she reproduce the teacher's underlying *competence*?

Consequences for communication

- Mind is willing to make errors in order to produce outputs faster.
- Do these errors lead to misunderstandings?
- Efficiency of communication: max. speed, min. misinterpretation.

Consequences for language change

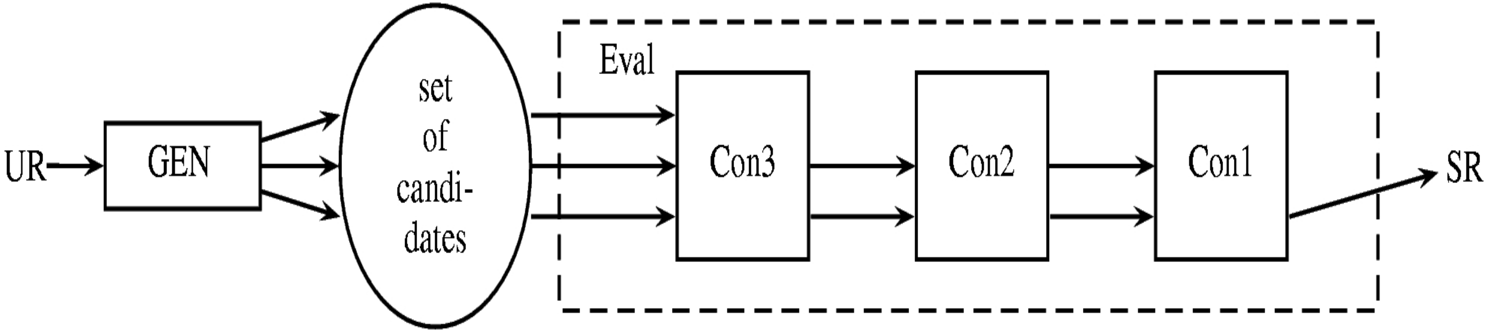
- Imperfect language aquisition (iterative learning)

- Maximizing efficiency of communication (evolutionary approach: speed vs. misinterpretations)

Overview

- Competence and performance: errors, irregularities, learning and communication
- Optimality Theory (OT) as an optimization problem
- The Simulated Annealing for OT Algorithm (SA-OT)
- Example: string grammar
- Learning SA-OT
- Concluding remarks

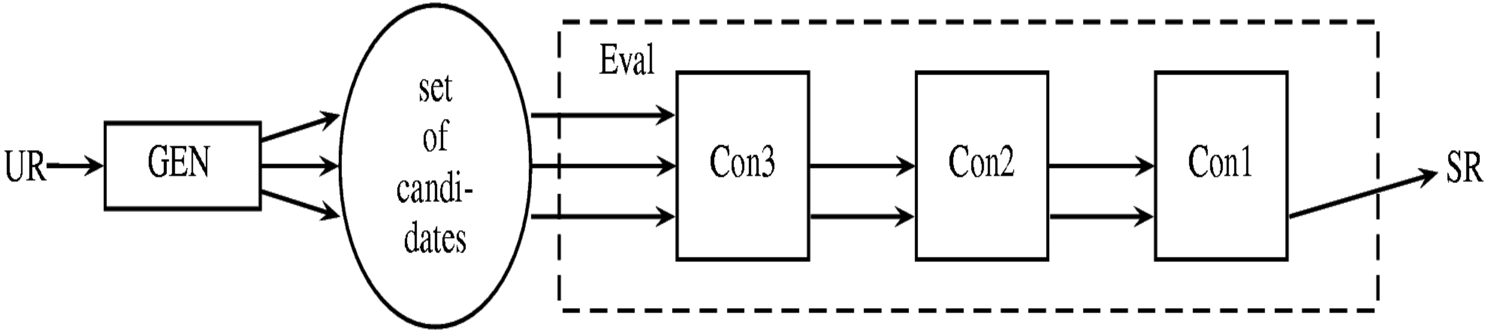
Optimality Theory in a nutshell



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	

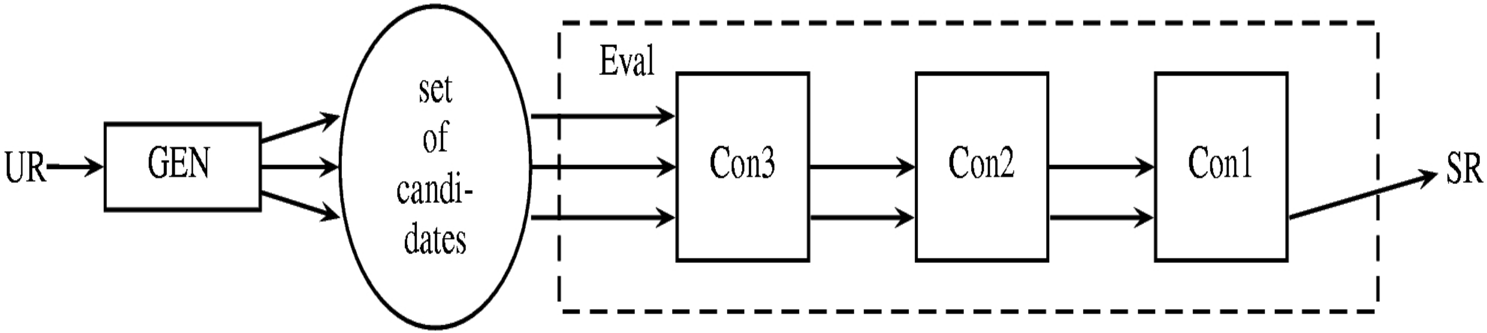
Optimality Theory in a nutshell



- 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 optimization problem



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

Optimization: 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...)

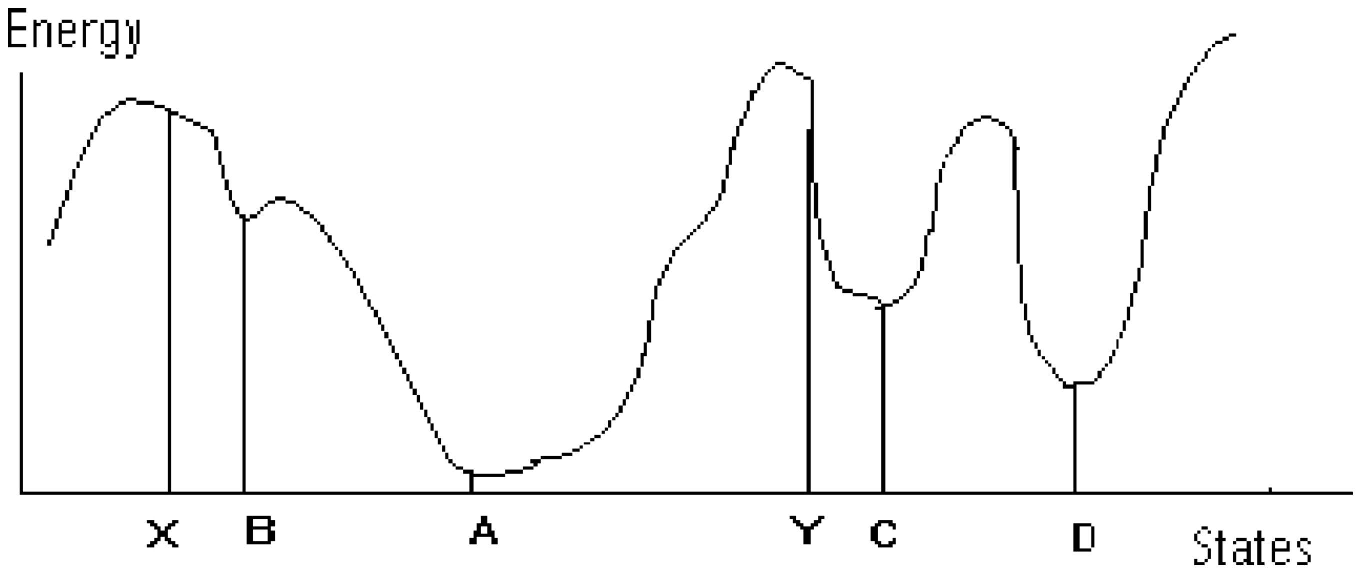
- chart parsing (dynamic programming) (Tesar & Smolensky; Kuhn)

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

Overview

- Competence and performance: errors, irregularities, learning and communication
- Optimality Theory (OT) as an optimization problem
- The Simulated Annealing for OT Algorithm (SA-OT)
- Example: string grammar
- Learning SA-OT
- Concluding remarks

Hill climbing: finding the global optimum



Global minimum: A
Local minima: B, C, D

Gradient descent from X brings to B, from Y brings to C.

How to find optimum: Gradient Descent 1

```
w := w_init ;
Repeat
        w':= best element of 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
```

How to find optimum: Gradient Descent 2

```
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
```

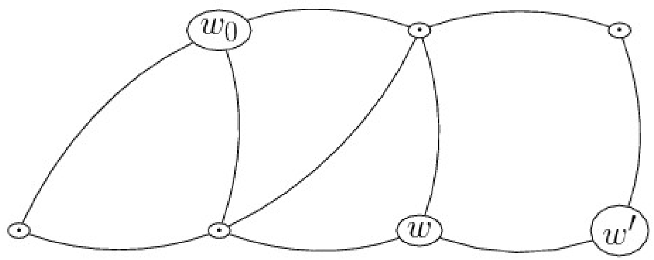
The Simulated Annealing Algorithm

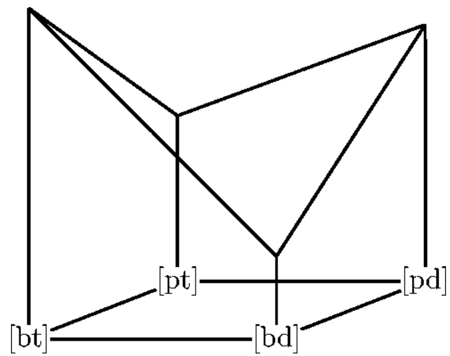
```
w := w_init ;       t := t_max   ;
Repeat
        Randomly select w' from the set Neighbours(w);
        Delta :=  E(w') - E(w) ;
        if   Delta < 0   then   w := w' ;
        else
               generate random r uniformly in range (0,1) ;
               if   r < exp(-Delta / t)   then   w :=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

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

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

Deterministic 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.

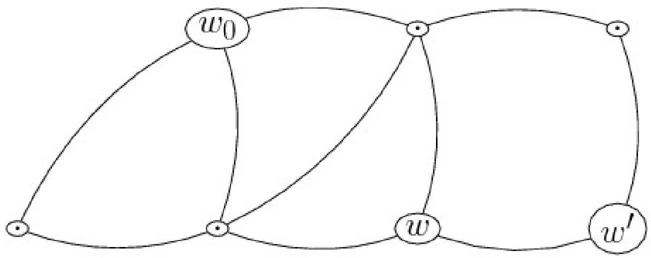
Simulated Annealing for OT – general idea

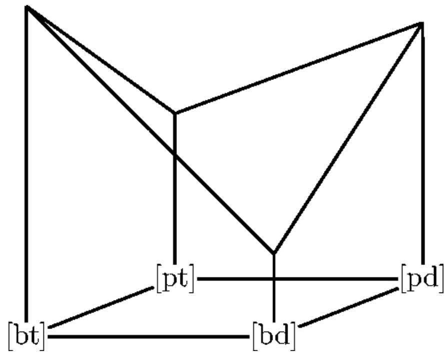




- Neighbourhood structure on the candidate set.
- Landscape's vertical dimension $=$ harmony; random walk.
- If neighbour more optimal: move.
- If less optimal: move in the beginning, don't move later.

Simulated Annealing for OT – general idea





- Neighbourhood structure $\rightarrow$ local optima.
- System stuck in local optima: alternation forms, errors
- Precision of the algorithm depends on its speed (!!).
- Many different scenarios.

Sim. annealing with non-real valued target function

- Exponential weights if upper bound on $C_i(w)$ violation levels:

$$E(w) = C_N(w) \cdot q^N + C_{N-1}(w) \cdot q^{N-1} + ... + C_1(w) \cdot q + C_0(w)$$

- Polynomials:

$$E(w)[q] = C_N(w) \cdot q^N + C_{N-1}(w) \cdot q^{N-1} + ... + C_1(w) \cdot q + C_0(w)$$

- Ordinal weights:

$$E(w) = \omega^N C_N(w) + ... + \omega C_1(w) + C_0(w)$$

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

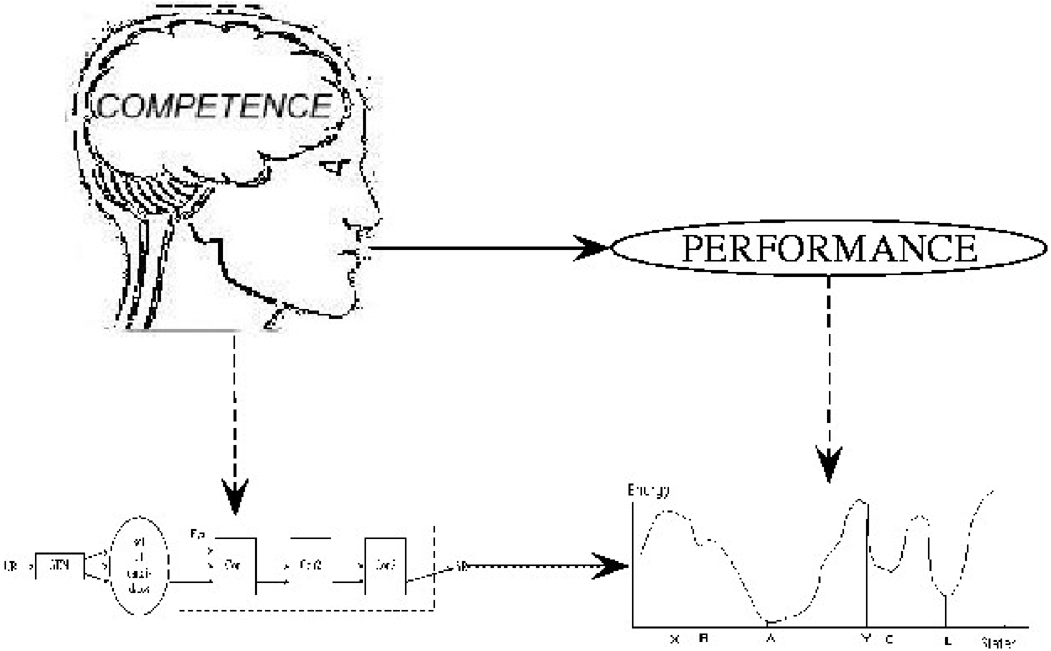
If $k = K_T$: move with probability

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

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
```

SA-OT as a model of linguistic performance



Optimality Theory: a model of competence

Simulated Annealing: a model of performance

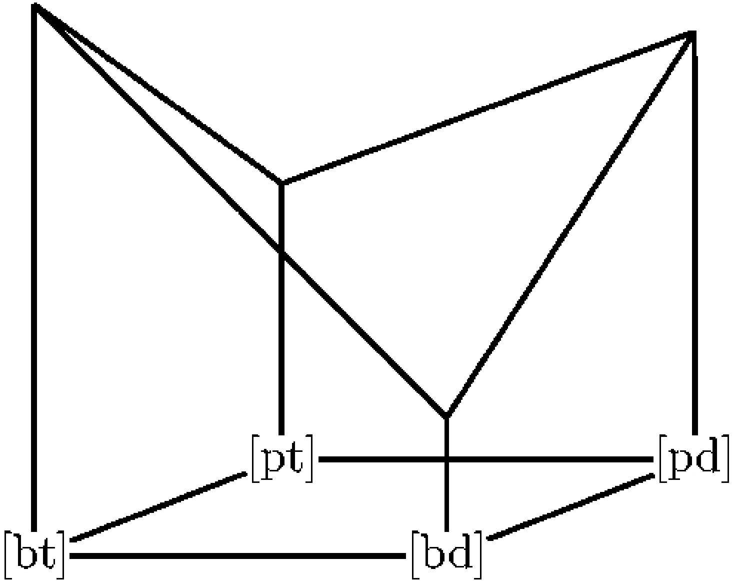
Proposal: three levels

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

Irregularities



- Local optimum that is not avoidable ($op\ die$).

Parameters of the algorithm

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

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 relying 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.

Overview

- Competence and performance: errors, irregularities, learning and communication
- Optimality Theory (OT) as an optimization problem
- The Simulated Annealing for OT Algorithm (SA-OT)
- Example: string grammar
- Learning SA-OT
- Concluding remarks

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 ± 1, $\mod P$ (cyclicity). Neighbours of 0022: 1022, 0012, 0322,...

- 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:

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

Example: string-grammar 4

$\mathcal{H}$: no0 $\gg$ ass $\gg$ Faith$_{\sigma=0000}$ $\gg$ ni1 $\gg$ ni0 $\gg$ ni2 $\gg$ ni3 $\gg$ nf0 $\gg$ nf1 $\gg$ nf2 $\gg$ nf3 $\gg$ no3 $\gg$ no2 $\gg$ no1 $\gg$ dis

Globally optimal form: ☞ 3333. But 13 local optima: 2222, $\{1,3\}^4$.

Output frequencies for different t_{step} (=inverse speed) values:

output	t_step = 1	t_step = 0.1	t_step = 0.01	t_step = 0.001
☞ 3333	0.1174 ± 0.0016	0.2074 ± 0.0108	0.2715 ± 0.0077	0.3107 ± 0.0032
∼ 1111	0.1163 ± 0.0021	0.2184 ± 0.0067	0.2821 ± 0.0058	0.3068 ± 0.0058
∼ 2222	0.1153 ± 0.0024	0.2993 ± 0.0092	0.3787 ± 0.0045	0.3602 ± 0.0091
! 1133	0.0453 ± 0.0018	0.0485 ± 0.0038	0.0328 ± 0.0006	0.0105 ± 0.0014
! 3311	0.0436 ± 0.0035	0.0474 ± 0.0054	0.0344 ± 0.0021	0.0114 ± 0.0016
! others	0.5608	0.1776	< 0.0002	–

$L = P = 4$, $T_{max} = 3$, $T_{min} = 0$, $K_{step} = 1$. Each candidate 4 times as w_0.

Overview

- Competence and performance: errors, irregularities, learning and communication
- Optimality Theory (OT) as an optimization problem
- The Simulated Annealing for OT Algorithm (SA-OT)
- Example: string grammar
- Learning SA-OT
- Concluding remarks

Learning

Learning algorithms in Optimality Theory:

- Off-line learning algorithms: Recursive Constraint Demotion
  - Initial grammar from observations in pre-linguistic infants?
  - Produces typical "children errors": extra local optima
- On-line learning algorithms: Error Driven Constraint Demotion, Gradual Learning Algorithm.
  - Grammar improving gradually in childhood?

Learning

Assumptions and heuristics behind learning algorithms:

- Traditional OT: observed form is optimal
- SA-OT: observed form is locally optimal
- Moreover: more frequent form is more harmonic
  Not always true in trad. OT; even less true in SA-OT.

Nevertheless, some success!

Same performance, different competence after RCD

target	after RCD	after GLA
No0 15	No0 15	No0 15.000000
Ass 14	Ass 12	Ass 14.000004
Fai 13	Fai 4	Fai 6.100000
Ni1 12	Ni1 8	Ni1 10.400004
Ni0 11	Ni0 13	Ni0 13.000000
Ni2 10	Ni2 5	Ni2 7.100000
Ni3 9	Ni3 3	Ni3 -1.500000
Nf0 8	Nf0 14	Nf0 14.000000
Nf1 7	Nf1 10	Nf1 6.300000
Nf2 6	Nf2 6	Nf2 8.100000
Nf3 5	Nf3 2	Nf3 3.600000
No3 4	No3 7	No3 3.000000
No2 3	No2 11	No2 13.100004
No1 2	No1 9	No1 10.900006
Dis 1	Dis 1	Dis -1.000000

Constraint name + its rank.

	target	after RCD	after GLA
	315 2222	322 2222	298 2222
	210 1111	221 1111	238 1111
☞	196 3333	200 3333	225 3333
	55 3111	53 1133	54 3311
	49 1133	50 3111	45 1133
	48 1333	45 1113	41 1333
	48 3331	45 3311	40 1113
	46 1113	42 3331	37 3331
	42 3311	32 1333	35 3111
	4 1331	4 1131	3 1331
	4 3133	4 1331	3 3113
	3 3113	2 1311	2 3133
	2 1311	2 3113	2 3313
	2 3313	2 3133	1 1131

Absolute frequency + output form.

GLA corrects children speech errors

target	after RCD	after GLA
No0 15	No0 2	No0 12.500014
Ass 14	Ass 14	Ass 12.299995
Fai 13	Fai 7	Fai -0.500000
Ni1 12	Ni1 9	Ni1 10.800001
Ni0 11	Ni0 15	Ni0 15.000000
Ni2 10	Ni2 13	Ni2 11.499998
Ni3 9	Ni3 11	Ni3 10.699999
Nf0 8	Nf0 8	Nf0 13.300017
Nf1 7	Nf1 6	Nf1 7.900000
Nf2 6	Nf2 1	Nf2 -12.200008
Nf3 5	Nf3 3	Nf3 9.000000
No3 4	No3 4	No3 3.700000
No2 3	No2 12	No2 11.400000
No1 2	No1 5	No1 3.300000
Dis 1	Dis 10	Dis 11.700005

target	after RCD	after GLA
302 2222	279 1111	292 2222
235 1111	277 2222	230 1111
☞ 208 3333	277 3333	194 3333
49 1113	39 1133	73 3311
49 3311	39 3311	52 1133
45 1333	**38 2200**	47 1113
44 1133	37 3111	45 3111
40 3111	35 1333	38 1333
37 3331	2 1113	37 3331
5 1331	1 1000	5 1331
4 3313		4 3133
2 3113		3 3313
2 3133		2 1311
1 1131		2 3113
1 1311		

GLA decreases freq. of 1111 and 3333. “Child form” 2200: extra local optimum.

GLA does not converge towards target

target	after RCD	after GLA
No0 15	No0 11	No0 20.200031
Ass 14	Ass 15	Ass 21.200026
Fai 13	Fai 8	Fai 7.600000
Ni1 12	Ni1 7	Ni1 5.299999
Ni0 11	Ni0 5	Ni0 13.300017
Ni2 10	Ni2 4	Ni2 7.500000
Ni3 9	Ni3 10	Ni3 -0.100000
Nf0 8	Nf0 14	Nf0 19.300018
Nf1 7	Nf1 12	Nf1 0.599994
Nf2 6	Nf2 3	Nf2 9.500005
Nf3 5	Nf3 2	Nf3 1.600000
No3 4	No3 1	No3 -14.000012
No2 3	No2 13	No2 18.900017
No1 2	No1 9	No1 7.400000
Dis 1	Dis 6	Dis -0.200000

	target	after RCD	after GLA
	300 2222	232 1111	250 1111
	231 1111	216 2222	239 2222
☞	214 3333	207 3333	226 3333
	53 1133	201 0000	170 0000
	51 3311	49 1133	36 2200
	50 3331	34 3311	33 1133
	46 1333	30 0022	31 0022
	38 3111	29 2200	30 3311
	33 1113	14 1113	3 1333
	6 1331	11 3331	2 1113
	1 1131	1 1333	2 3111
	1 3113		2 3331

Constraint ranks diverge. 0000 is locally optimal.

Overview

- Competence and performance: errors, irregularities, learning and communication
- Optimality Theory (OT) as an optimization problem
- The Simulated Annealing for OT Algorithm (SA-OT)
- Example: string grammar
- Learning SA-OT
- Concluding remarks

Conclusions: language modelling

- Performance as the implementation of the grammar.
  E.g.: competence = OT, performance = SA-OT or ICS.

- Errors and irregularities are good for
  - asses the descriptive adequacy of the combined competence + performance model
  - language learning/acquisition, and evolution
    E.g.: some children language forms as extra local optima.

- *Errare humanum est*: heuristics in cognitive science.
  Nonetheless: efficient communication

What does SA-OT offer to standard OT?

- A new approach to account for variation / performance:
  - Non-optimal candidates also produced (cf. Coetzee);
  - As opposed to: more candidates with same violation profile; more hierarchies in a grammar.
- A topology (neighbourhood structure) on the candidate set.
- Additional ranking arguments $\longrightarrow$ learning.
- Godot-effect: role played by loser candidates.
- New implementation: cognitively plausible, lang. technology?

- Demo at `http://www.let.rug.nl/~birot/sa-ot`.
- (Plans to develop it further)

Language vs. Culture?

The tacit knowledge of a participant in a symbolic-cultural system is neither taught nor learned by rote. Rather each new participant [...] reconstructs the rules which govern the symbolic-cultural system in question. These reconstructions may differ considerably, depending upon such factors as the personal history of the individual in question. Consequently, the products of each individual's symbolic mechanism are idiosyncratic to some extent. (Lawson-McCauley, 1990, p. 68., italics original)

Said about culture, as a difference from language.
I have now argued: it also holds for language!

The *dis*-harmonic mind?

ICS (Integrated Connectionist/Symbolic Cognitive Architecture):

"[T]here is no symbolic algorithm whose internal structure can predict the time and the accuracy of processing; this can only be done with connectionist algorithms" (Smolensky and Legendre (2006): *The Harmonic Mind*, vol. 1, p. 91).

SA-OT:

- symbolic computation only
- predicts tme and accuracy of processing

Thank you for your attention!

Tamás Bíró

birot@nytud.hu