Acquiring Competence from Performance Data

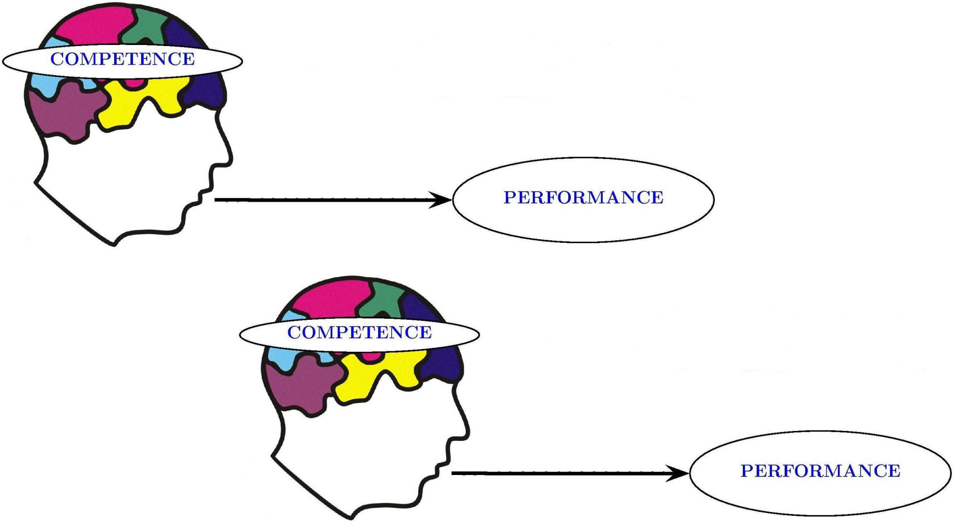
Online learnability of OT and HG with simulated annealing

Tamás Biró

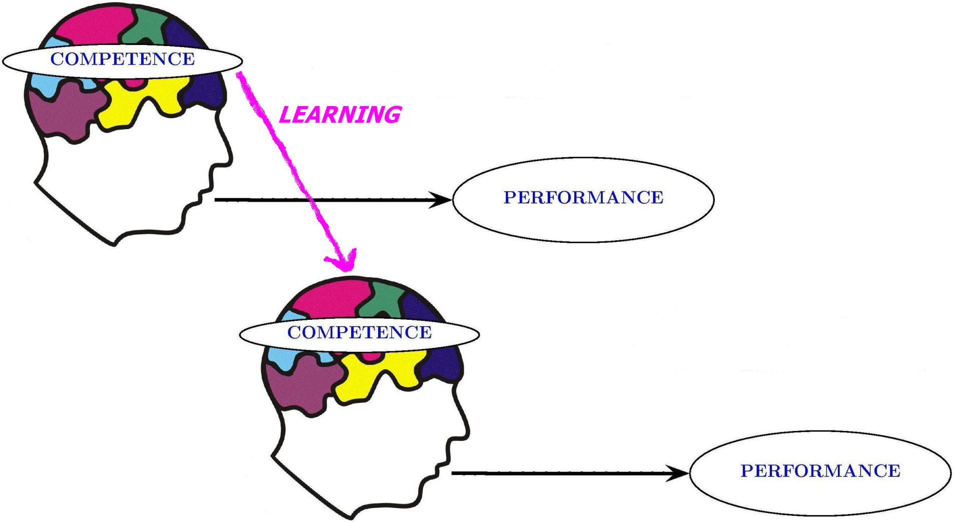
ACLC, University of Amsterdam (UvA)

Computational Linguistics in the Netherlands, February 5, 2010

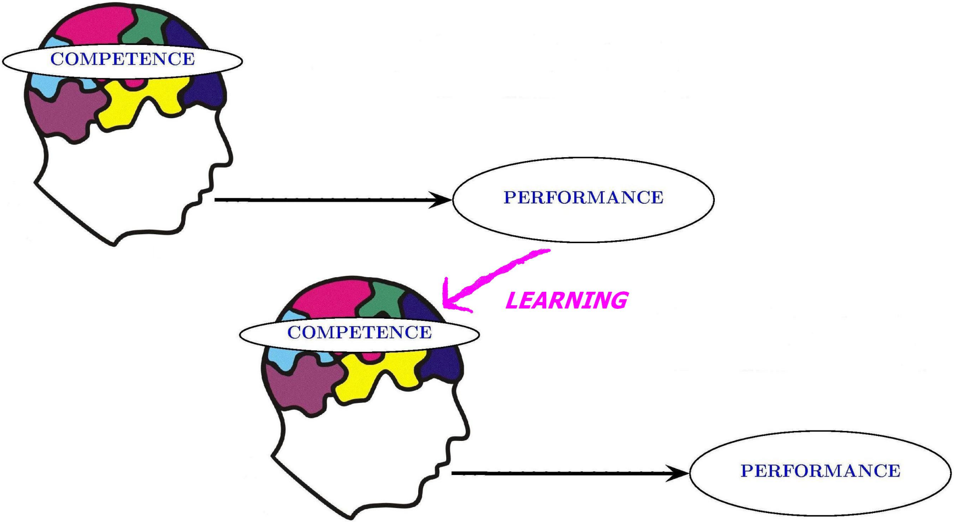
The language acquisition problem



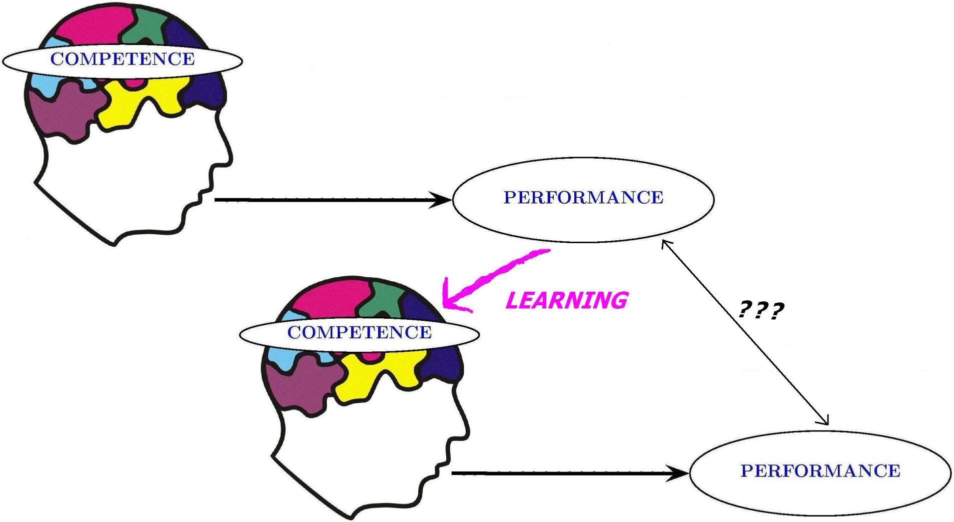
Learning from competence?



Learning from performance!



Distance of teacher's and learner's performance



Overview

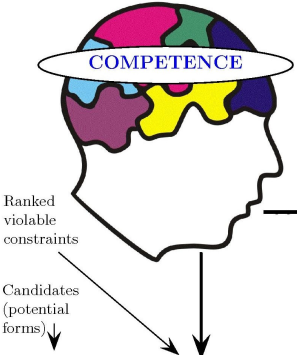
1. Modelling linguistic performance
2. Learning
3. Results
4. Conclusions

Overview

1. Modelling linguistic performance
2. Learning
3. Results
4. Conclusions

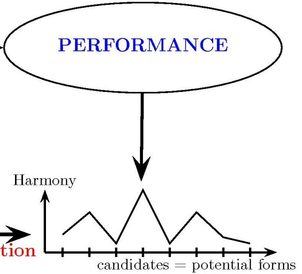
Errors and mental computations

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.



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

implementation



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

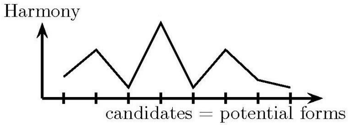
SA-OT
implementation
performance model
produced forms = globally or locally optimal candidates

Competence and performance models

$$\mathrm{SF}(U) = \underset{w \in \mathrm{Gen}(U)}{\arg\operatorname{opt}} \; H(w)$$

Competence models:

- $C_i(w)$ elementary functions on the candidates ("constraints" – a misnomer).
- Optimality Theory: $H(w) = (C_n(w), ..., C_1(w))$
  $\arg\operatorname{opt}$: lexicographic order.
- q-Harmony Grammar: $H(w) = C_n(w) \cdot q^n + ... + C_i(w) \cdot q$.
  Large q: OT-like strict domination.
  Small q: ganging-up effects.



Performance models:

- Exhaustive search: returns global optimum.
- Simulated annealing: returns some local optimum.
  - Run slowly: frequently the globally optimal one.
  - Run quickly: global opt. less frequent, more often performance errors.

Overview

1. Modelling linguistic performance
2. Learning
3. Results
4. Conclusions

Online learning algorithms

Constraint C_i has rank r_i.

In each learning cycle: learning data (*winner*) produced by teacher compared to form produced by learner (*loser*).

Update rule: update the rank r_i of every constraint C_i, depending on whether C_i prefers the winner or the loser.

- Boersma (1997): increase rank by ϵ if winner-preferring; decrease rank by ϵ if loser-preferring constraint.
- Magri (2009): increase rank of all winner-preferring constraints by ϵ; decrease rank of highest ranked loser-preferring constraint by $W \cdot \epsilon$, where W is the number of winner-preferring constraints.

Learn until performance converges

- Convergence of performance, and not of competence. Child may acquire different grammar.
- Sample of teacher vs. sample of learner (sample size $= 100$).
- **Convergence criterion:** *JSD* between sample produced by target grammar and sample produced by learner's current grammar $\leq$ average *JSD* of two samples produced by target grammar.

Jensen-Shannon divergence: measures the "distance" of two distributions

$$JSD(P||Q) = \frac{D(P||M) + D(Q||M)}{2}$$

where $D(P||Q) = \sum_x P(x) \log \frac{P(x)}{Q(x)}$ (relative entropy, Kullback-Leibler divergence), $M(x) = \frac{P(x)+Q(x)}{2}$.

- Symmetric: $JSD(P||Q) = JSD(Q||P)$. Non-negative: $JSD(P||Q) \geq 0$. $JSD(P||Q) \leq 1$.
- $JSD(P||Q) = 0$ if and only if $P(x) = Q(x),\ \forall x$. $JSD(P||Q) = 1$ if and only if $P(x) \cdot Q(x) = 0,\ \forall x$.
- Same language: $JSD(L_t||L_l) = 0$. Not a single overlap: $JSD(L_t||L_l) = 1$.

Overview

1. Modelling linguistic performance
2. Learning
3. Results
4. Conclusions

Results: number of learning steps until convergence

- 2000 times learning (rnd target, rnd underlying form) per grammar type × production method × learning method.
- Measure the number of learning steps until convergence.
- Distribution of the number of required learning steps:

		OT	10-HG	4-HG	1.5-HG
gramm.	M	*13* ; 27 ; *45*	*13* ; 28 ; *46*	*12* ; 27 ; *48*	*15* ; 30 ; *47*
	B	*23* ; 43 ; *65*	*22* ; 41 ; *64*	*22* ; 42 ; *64*	*23* ; 40 ; *60*
sa,	M	*53* ; 109 ; *233*	*63* ; 140 ; *328*	*60* ; 148 ; *366*	*83* ; 199 ; *508*
$t_{\text{step}} = 0.1$	B	*80* ; 171 ; *462*	*92* ; 240 ; *772*	*92* ; 239 ; *785*	*117* ; 290 ; *694*
sa,	M	*64* ; 131 ; *305*	*62* ; 134 ; *304*	*63* ; 137 ; *329*	*72* ; 163 ; *437*
$t_{\text{step}} = 1$	B	*90* ; 212 ; *560*	*92* ; 233 ; *572*	*84* ; 212 ; *646*	*101* ; 242 ; *616*

(*1st quartile* ; median ; *3rd quartile*)

Methodological notes

Paradigm:

- Measure number of learning steps until converging performance.
- Statistics on the distribution of the required learning step number.
- Under different learning conditions.
- Distributions have extremely long tails.
  Significance of differences: using non-parametric tests.

Does learning speed depend on initial grammar? On learning data?
Run two learners learning the same target grammar:

- with same initial grammar: strong correlation in nr. of learning steps.
  Learning data not the same: slightly decreased correlation.
- with different initial grammars: correlation (almost) lost.

Long tail: children must start with same initial grammar, but need not receive same (correct or erroneous) data (if learning algorithm is correct).

Conclusions

Proposed paradigm for the learnability of a grammar framework:

- Competence = grammar framework (e.g., OT or HG).
- Performance = imperfect implementation of competence model.
- Learning from performance data, only partially reflecting competence.
- Learner does not have access to teacher's competence directly: converge on performance.
- Convergence measure using *Jensen-Shannon divergence*.
- Argument for same initial grammar in children?

Implemented on OTKit.

Thank you for your attention!

Tamás Biró:
t.s.biro@uva.nl

Tools for Optimality Theory
http://www.birot.hu/OTKit/

Work supported by:

Amsterdam Center for Language and Communication ACLC

NWO Netherlands Organisation for Scientific Research