Errors in language production, language learning and language change

Some computational experiments with Optimality Theory

Tamás Biró

ACLC, University of Amsterdam (UvA)

OAP dag, December 16, 2011

Example: sentential negation (Jespersen's cycle)

	pre-verbal	discontinuous	post-verbal
French	Jeo ne dis	Je ne dis pas	Je dis pas
English	Ic ne secge	Ic ne seye not	I say not
	1. *SN V*	2. *SN V SN*	3. *V SN*

To explain:

- Typology: pre-verbal, discontinuous, post-verbal,
- ... as well as mixed types.
- Diachronic change (a.k.a. language evolution).

Example: sentential negation (Jespersen's cycle)

	pre-verbal	discontinuous	post-verbal
French	Jeo ne dis	Je ne dis pas	Je dis pas
English	Ic ne secge	Ic ne seye not	I say not
	1. *SN V*	2. *SN V SN*	3. *V SN*

To explain:

- Typology: pre-verbal, discontinuous, post-verbal,
- ... as well as mixed types.
- Diachronic change (a.k.a. language evolution).

Example: sentential negation (Jespersen's cycle)

	pre-verbal	discontinuous	post-verbal
French	Jeo ne dis	Je ne dis pas	Je dis pas
English	Ic ne secge	Ic ne seye not	I say not
	1. *SN V*	2. *SN V SN*	3. *V SN*

To explain:

- Typology: pre-verbal, discontinuous, post-verbal,
- ... as well as mixed types.
- Diachronic change (a.k.a. language evolution).

Example: sentential negation (Jespersen's cycle)

	pre-verbal	discontinuous	post-verbal
French	Jeo ne dis	Je ne dis pas	Je dis pas
English	Ic ne secge	Ic ne seye not	I say not
	1. *SN V*	2. *SN V SN*	3. *V SN*

To explain:

- Typology: pre-verbal, discontinuous, post-verbal,
- ... as well as mixed types.
- Diachronic change (a.k.a. language evolution).

Research questions:

The role of *errors* = the results of imperfect mental computation.

- "Performance errors": ungrammatical but produced.
- Learning in the presence of "performance errors".
- "Performance errors" as a driving force behind language change.
- Another reason for making errors during learning.

References

- A. Lopopolo and T. Biró. 'Language Evolution and SA-OT: The case of sentential negation'. *Computational Linguistics in the Netherlands Journal* 1(2011):21–40.
- T. Biró. 'Towards a Robuster Interpretive Parsing: Learning from overt forms in Optimality Theory'. Submitted to *Journal of Logic, Language and Information*.

Research questions:

The role of *errors* = the results of imperfect mental computation.

- "Performance errors": ungrammatical but produced.
- Learning in the presence of "performance errors".
- "Performance errors" as a driving force behind language change.
- Another reason for making errors during learning.

References

- A. Lopopolo and T. Biró. 'Language Evolution and SA-OT: The case of sentential negation'. *Computational Linguistics in the Netherlands Journal* 1(2011):21–40.
- T. Biró. 'Towards a Robuster Interpretive Parsing: Learning from overt forms in Optimality Theory'. Submitted to *Journal of Logic, Language and Information*.

Research questions:

The role of *errors* $=$ the results of imperfect mental computation.

- “Performance errors”: ungrammatical but produced.
- Learning in the presence of “performance errors”.
- “Performance errors” as a driving force behind language change.
- Another reason for making errors during learning.

References

- A. Lopopolo and T. Biró. ‘Language Evolution and SA-OT: The case of sentential negation’. *Computational Linguistics in the Netherlands Journal* 1(2011):21–40.
- T. Biró. ‘Towards a Robuster Interpretive Parsing: Learning from overt forms in Optimality Theory’. Submitted to *Journal of Logic, Language and Information*.

Research questions:

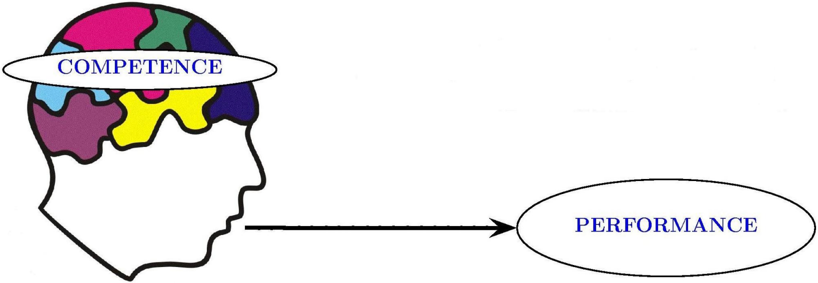
The role of *errors* = the results of imperfect mental computation.

- "Performance errors": ungrammatical but produced.
- Learning in the presence of "performance errors".
- "Performance errors" as a driving force behind language change.
- Another reason for making errors during learning.

References

- A. Lopopolo and T. Biró. 'Language Evolution and SA-OT: The case of sentential negation'. *Computational Linguistics in the Netherlands Journal* 1(2011):21–40.
- T. Biró. 'Towards a Robuster Interpretive Parsing: Learning from overt forms in Optimality Theory'. Submitted to *Journal of Logic, Language and Information*.

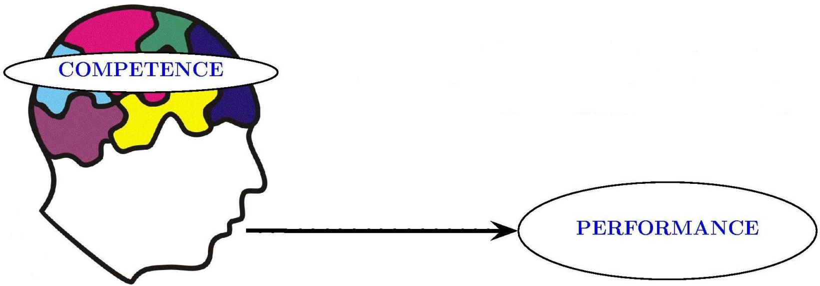
Errors of the mental computation



static knowledge

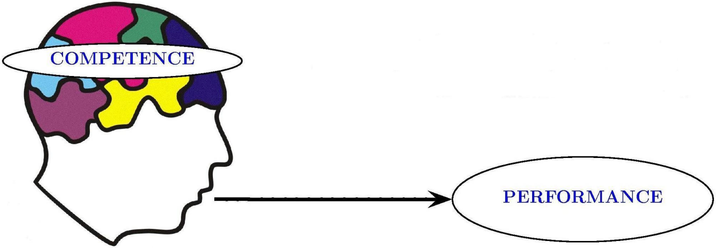
processes in the brain

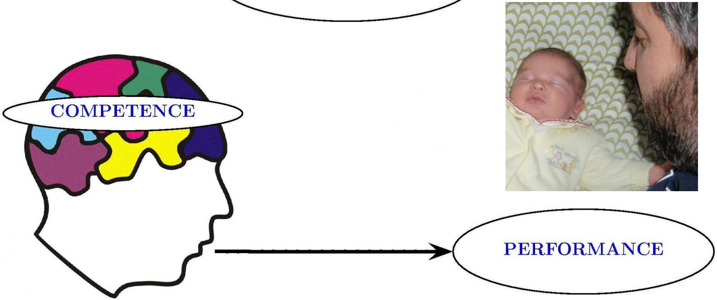
Errors of the mental computation



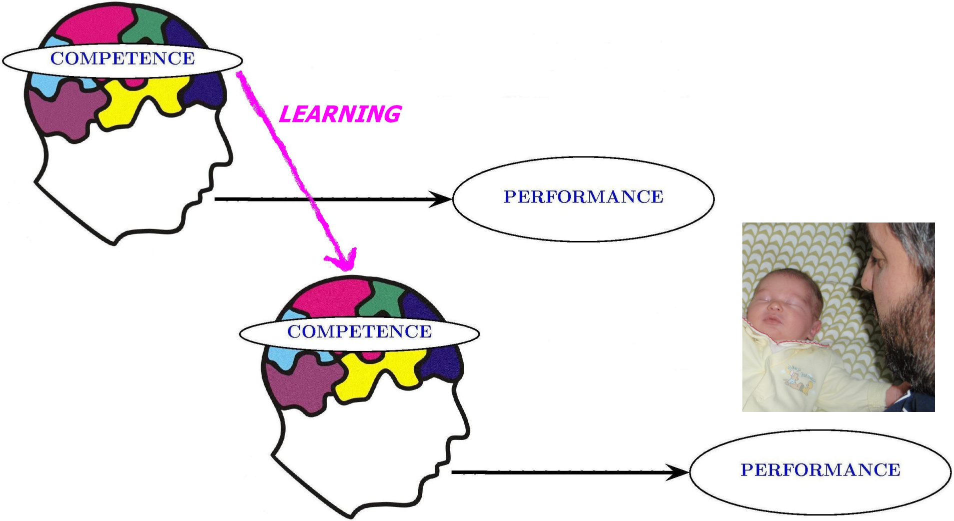
static knowledge processes in the brain

The language acquisition problem

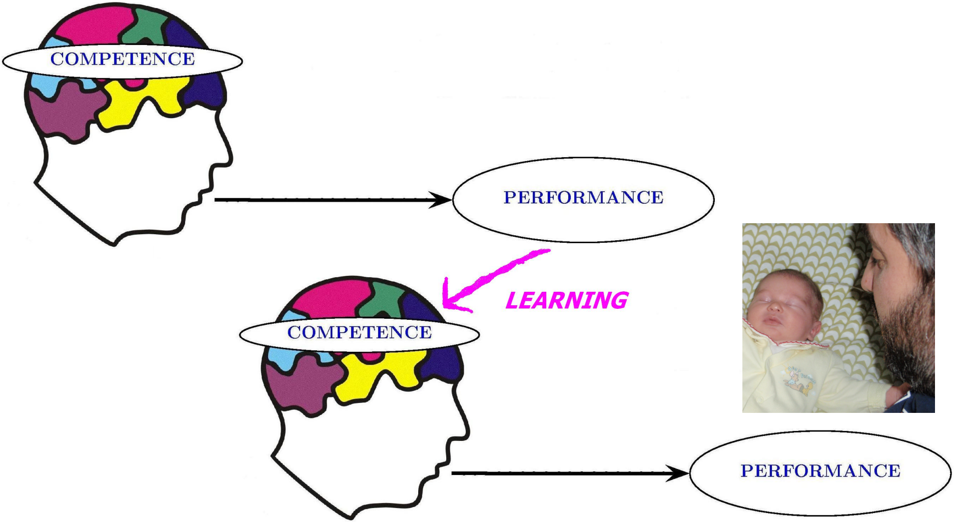




Learning from competence?



Learning from performance!



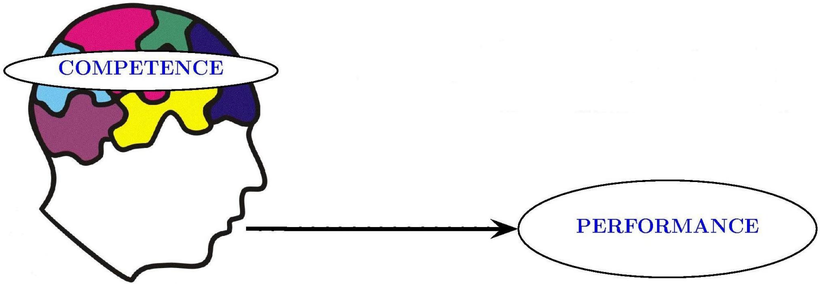
Overview

1. Modelling linguistic performance
2. Issues in learning and iterated learning
3. The problem of the overt forms
4. Conclusions

Overview

1. Modelling linguistic performance
2. Issues in learning and iterated learning
3. The problem of the overt forms
4. Conclusions

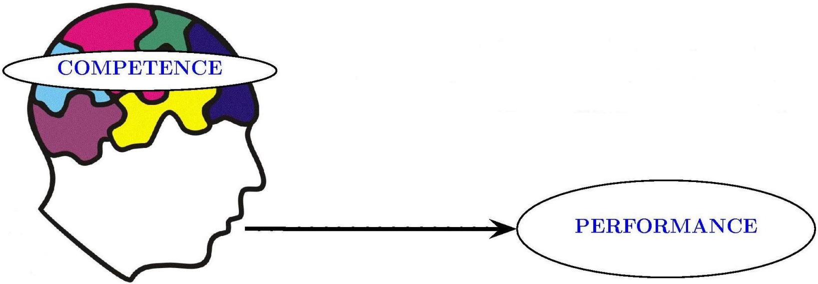
Errors of the mental computation



static knowledge processes in the brain

Optimality Theory Simulated Annealing for OT

Errors of the mental computation

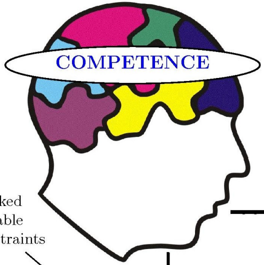


static knowledge Optimality Theory

processes in the brain Simulated Annealing for OT

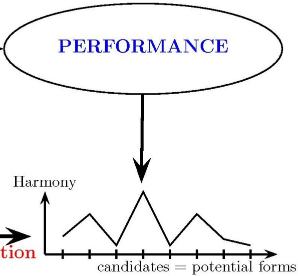
Errors of the mental computation

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.



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

Modelling linguistic competence

FAITH[NEG] ≫ *NEGATION ≫ NEGATIONFIRST ≫ NEGATIONLAST

	/pol = neg/	Faith[Neg]	*Neg	NegFirst	NegLast
	[V]	*		*	*
☞	[SN V]		*		*
	[V SN]		*	*	
	[SN V SN]		**		
	[V SN SN]		**	*	
	[SN SN V]		**		*
	[SN V SN SN]		***		
	...				

Lopopolo and Biró (2011), based on Henriëtte de Swart (2010).

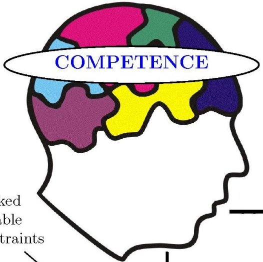
Modelling linguistic competence

FAITH[NEG] ≫ NEGATIONFIRST ≫ *NEGATION ≫ NEGATIONLAST

	/pol = neg/	Faith[Neg]	NegFirst	*Neg	NegLast
	[V]	*	*		*
☞	[SN V]			*	*
	[V SN]		*	*	
	[SN V SN]			**	
	[V SN SN]		*	**	
	[SN SN V]			**	*
	[SN V SN SN]			***	
	...				

Lopopolo and Biró (2011), based on Henriëtte de Swart (2010).

Errors of the mental computation



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.

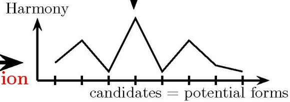
PERFORMANCE

Ranked violable constraints

Candidates (potential forms)

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

implementation

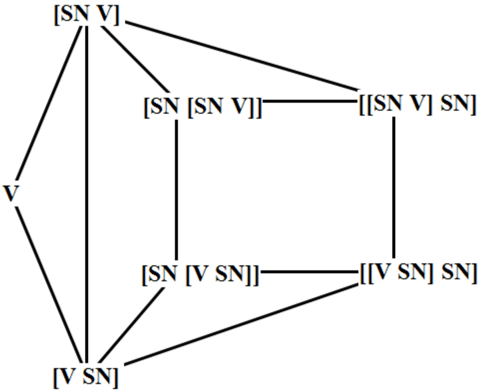


Optimality Theory
grammar
competence model
grammatical form = ☞ (globally) optimal candidate

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

Modelling linguistic performance

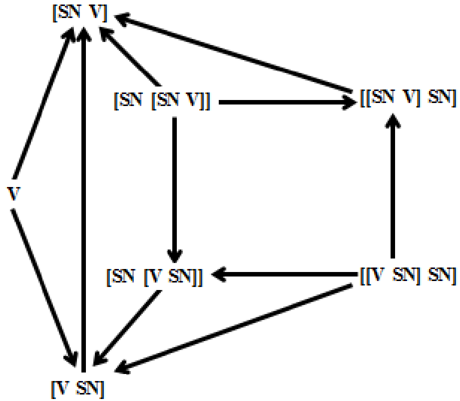
A topology (neighborhood structure) on the candidate set:



Locally optimal forms: are predicted to be the produced forms.

Modelling linguistic performance

FAITH[NEG] ≫ *NEGATION ≫ NEGATIONFIRST ≫ NEGATIONLAST

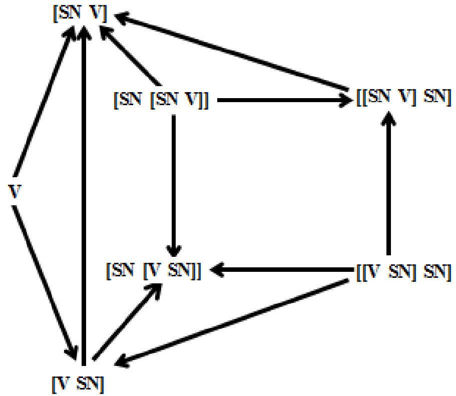


Hierarchy 1: *Neg >> NegFirst >> NegLast

Locally optimal forms: ☞ [SN V].

Modelling linguistic performance

FAITH[NEG] $\gg$ NEGATIONFIRST $\gg$ *NEGATION $\gg$ NEGATIONLAST



Hierarchy 2: NegFirst >> *Neg >> NegLast

Locally optimal forms: ☞ [SN V] and ∼ [SN [V SN]].

	Hierarchy	competence	performance
1.	*Neg $\gg$ NegFirst $\gg$ NegLast	pre-verbal	pre-verbal
2.	NegFirst $\gg$ *Neg $\gg$ NegLast	pre-verbal	pre-V and discont.
3.	NegFirst $\gg$ NegLast $\gg$ *Neg	discontinuous	discontinuous
4.	NegLast $\gg$ NegFirst $\gg$ *Neg	discontinuous	discontinuous
5.	NegLast $\gg$ *Neg $\gg$ NegFirst	post-verbal	discont. and post-V
6.	*Neg $\gg$ NegLast $\gg$ NegFirst	post-verbal	post-verbal

Observerd typology: 3 pure types and 2 mixed types.
Predicted typology:

- Traditional OT (H. de Swart): 3 pure types.
- Stochastic OT (H. de Swart): 3 pure types and 3 mixed types.
- SA-OT (Lopopolo and Biró): 3 pure types and 2 mixed types.

	Hierarchy	competence	performance
1.	*Neg $\gg$ NegFirst $\gg$ NegLast	pre-verbal	pre-verbal
2.	NegFirst $\gg$ *Neg $\gg$ NegLast	pre-verbal	pre-V and discont.
3.	NegFirst $\gg$ NegLast $\gg$ *Neg	discontinuous	discontinuous
4.	NegLast $\gg$ NegFirst $\gg$ *Neg	discontinuous	discontinuous
5.	NegLast $\gg$ *Neg $\gg$ NegFirst	post-verbal	discont. and post-V
6.	*Neg $\gg$ NegLast $\gg$ NegFirst	post-verbal	post-verbal

Observerd typology: 3 pure types and 2 mixed types.
Predicted typology:

- Traditional OT (H. de Swart): 3 pure types.
- Stochastic OT (H. de Swart): 3 pure types and 3 mixed types.
- SA-OT (Lopopolo and Biró): 3 pure types and 2 mixed types.

	Hierarchy	competence	performance
1.	*Neg $\gg$ NegFirst $\gg$ NegLast	pre-verbal	pre-verbal
2.	NegFirst $\gg$ *Neg $\gg$ NegLast	pre-verbal	pre-V and discont.
3.	NegFirst $\gg$ NegLast $\gg$ *Neg	discontinuous	discontinuous
4.	NegLast $\gg$ NegFirst $\gg$ *Neg	discontinuous	discontinuous
5.	NegLast $\gg$ *Neg $\gg$ NegFirst	post-verbal	discont. and post-V
6.	*Neg $\gg$ NegLast $\gg$ NegFirst	post-verbal	post-verbal

Observerd typology: 3 pure types and 2 mixed types.
Predicted typology:

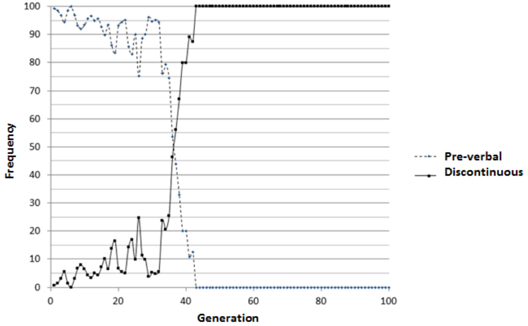
- Traditional OT (H. de Swart): 3 pure types.
- Stochastic OT (H. de Swart): 3 pure types and 3 mixed types.
- SA-OT (Lopopolo and Biró): 3 pure types and 2 mixed types.

Overview

1. Modelling linguistic performance
2. Issues in learning and iterated learning
3. The problem of the overt forms
4. Conclusions

Iterated learning: reproducing language change (?)

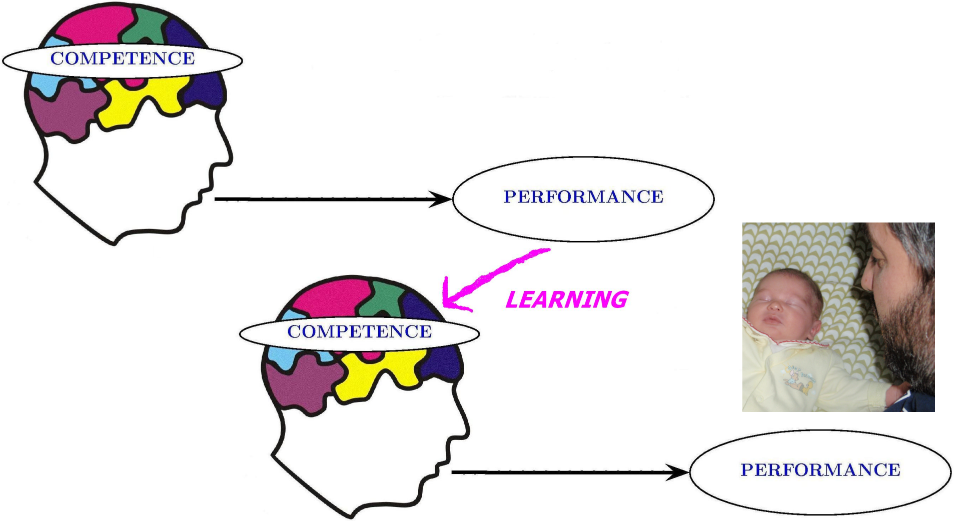
Five agents in each generation. Generations 0 to 100.
Each agent learns from every agent in the previous generation.
Negation types in the “simulated historical corpus”:



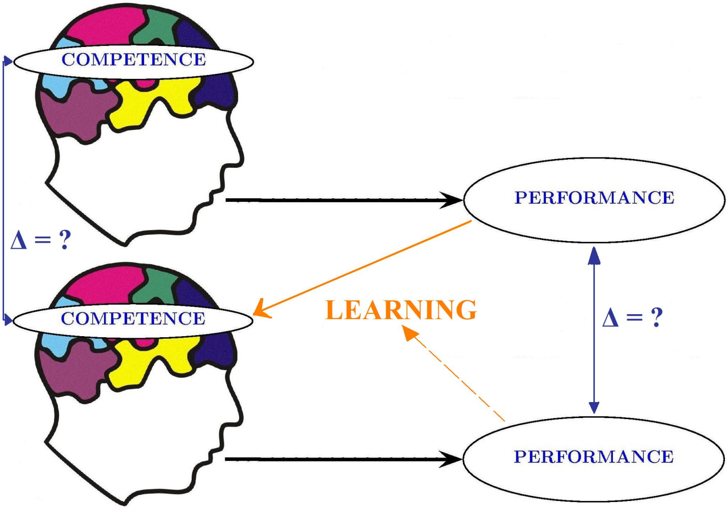
A. Lopopolo and T. Biró. ‘Language Evolution and SA-OT: The case of sentential negation’.

Computational Linguistics in the Netherlands Journal 1(2011):21–40.

Learning from performance!



Language acquisition with online learning algorithms



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.

- Run until convergence of performance, and not of competence.
- Distance of teacher sample vs. learner sample measured by JSD:

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), $M(x) = \frac{P(x)+Q(x)}{2}$.

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.

- Run until convergence of performance, and not of competence.
- Distance of teacher sample vs. learner sample measured by JSD:

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), $M(x) = \frac{P(x)+Q(x)}{2}$.

Results: number of learning steps until convergence

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

production	update rule	OT	10-HG	4-HG
grammatical	Magri	*13* ; 27 ; *45* ; 67	*13* ; 28 ; *46* ; 70	*12* ; 27 ; *48* ; 69
	GLA	*23* ; 43 ; *65* ; 102	*22* ; 41 ; *64* ; 107	*22* ; 42 ; *64* ; 107
SA-OT,	Magri	*53* ; 109 ; *233* ; 497	*63* ; 140 ; *328* ; 1681	*60* ; 148 ; *366* ; 1517
$t_{\text{step}} = 0.1$	GLA	*80* ; 171 ; *462* ; 1543	*92* ; 240 ; *772* ; 7512	*92* ; 239 ; *785* ; 8633
SA-OT,	Magri	*64* ; 131 ; *305* ; 1022	*62* ; 134 ; *304* ; 1127	*63* ; 137 ; *329* ; 1278
$t_{\text{step}} = 1$	GLA	*90* ; 212 ; *560* ; 1966	*92* ; 233 ; *572* ; 3116	*84* ; 212 ; *646* ; 3005

(*1st quartile* ; median ; *3rd quartile* ; 90th percentile)

Overview

The problem of the overt forms

- Generation n produces [SN [V SN]] and utters “SN V SN”.
- Generation $n+1$ hears “SN V SN”.
  Is it [SN [V SN]] or [[SN V] SN]?
- In general, huge amount of crucial information for the reconstruction of a grammar is covert.
  - Co-indexation: *He$_i$ looks like him$_{i/j}$*.
  - Foot structure: *banána* proof for *ba[nána]* or *[baná]na*?
  - Basic word order: *John loves Mary* proof for SVO or OVS?
- Does it mislead learning?

The problem of the overt forms

- Generation n produces [SN [V SN]] and utters "SN V SN".
- Generation $n+1$ hears "SN V SN".
  Is it [SN [V SN]] or [[SN V] SN]?
- In general, huge amount of crucial information for the reconstruction of a grammar is covert.
  - Co-indexation: *He*$_i$ *looks like him*$_{i/j}$.
  - Foot structure: *banána* proof for *ba[nána]* or *[baná]na*?
  - Basic word order: *John loves Mary* proof for SVO or OVS?
- Does it mislead learning?

The problem of the overt forms

	Learner → *Neg	V-right	← Teacher V-left
L ☞ [SN V]	*		*
L's target [[SN V] SN]	**		*
T ☞ [SN [V SN]]	**	*	

- Learner: *Neg ≫ V-right ≫ V-left. Produces [SN V].
- Teacher: V-left ≫ V-right ≫ *Neg. Produces [SN [V SN]].
- Learner hears "SN V SN". Would like to change her grammar to produce ... [[SN V] SN] or [SN [V SN]]?
- Form [[SN V] SN] is still better than [SN [V SN]] in her grammar, so she takes it as the target for learning,
- ... and fails to learn the target language.

The problem of the overt forms

Learner → ← Teacher

	*Neg	V-right	V-left
L ☞ [SN V]	*		*
L's target [[SN V] SN]	**		*
T ☞ [SN [V SN]]	**	*	

- Learner: *Neg $\gg$ V-right $\gg$ V-left. Produces [SN V].
- Teacher: V-left $\gg$ V-right $\gg$ *Neg. Produces [SN [V SN]].
- Learner hears "SN V SN". Would like to change her grammar to produce ... [[SN V] SN] or [SN [V SN]]?
- Form [[SN V] SN] is still better than [SN [V SN]] in her grammar, so she takes it as the target for learning,
- ... and fails to learn the target language.

The problem of the overt forms

A (partial) solution:

- Learner hears “SN V SN”. Is it [[SN V] SN] or [SN [V SN]]?
- Since the learner really cannot know,
  she takes the (weighted) average of the violations by these forms,
- Teacher produces [SN [V SN]]. Learner produces [SN V].
  and updates the grammar in order to approach this average.

Learner → ← Teacher

		*Neg	V-right	V-left
L ☞	[SN V]	*		*
	[[SN V] SN]	**		*
T ☞	[SN [V SN]]	**	*	
L's target	“average”	2	0.5	0.5

The improved learning algorithm performs significantly better:
Biró. ‘Towards a Robuster Interpretive Parsing: Learning from overt forms in Optimality Theory’. Submitted to *Journal of Logic, Language and Information*.

The problem of the overt forms

A (partial) solution:

- Learner hears “SN V SN”. Is it [[SN V] SN] or [SN [V SN]]?
- Since the learner really cannot know,
  she takes the (weighted) average of the violations by these forms,
- Teacher produces [SN [V SN]]. Learner produces [SN V].
  and updates the grammar in order to approach this average.

Learner → ← Teacher

	*Neg	V-right	V-left
L ☞ [SN V]	*		*
[[SN V] SN]	**		*
T ☞ [SN [V SN]]	**	*	
L’s target “average”	2	0.5	0.5

The improved learning algorithm performs significantly better:
Biró. ‘Towards a Robuster Interpretive Parsing: Learning from overt forms in Optimality Theory’. Submitted to *Journal of Logic, Language and Information*.

Conclusions

The role of *errors* = the results of imperfect mental computation.

- OT as a model of competence (static knowledge), Simulated Annealing for OT as a model of the (eventually erroneous) computation in the brain (performance).
- “Performance errors” as driving force behind language change.
- Language learning until convergence on performance patterns (measured using Jensen-Shannon Divergence).
- Different learning methods need different numbers of learning step until convergence.
- Learning despite hidden (covert) information.

Conclusions

The role of *errors* = the results of imperfect mental computation.

- OT as a model of competence (static knowledge), Simulated Annealing for OT as a model of the (eventually erroneous) computation in the brain (performance).
- "Performance errors" as driving force behind language change.
- Language learning until convergence on performance patterns (measured using Jensen-Shannon Divergence).
- Different learning methods need different numbers of learning step until convergence.
- Learning despite hidden (covert) information.

Conclusions

The role of *errors* $=$ the results of imperfect mental computation.

- OT as a model of competence (static knowledge), Simulated Annealing for OT as a model of the (eventually erroneous) computation in the brain (performance).
- “Performance errors” as driving force behind language change.
- Language learning until convergence on performance patterns (measured using Jensen-Shannon Divergence).
- Different learning methods need different numbers of learning step until convergence.
- Learning despite hidden (covert) information.

Conclusions

The role of *errors* $=$ the results of imperfect mental computation.

- OT as a model of competence (static knowledge), Simulated Annealing for OT as a model of the (eventually erroneous) computation in the brain (performance).
- "Performance errors" as driving force behind language change.
- Language learning until convergence on performance patterns (measured using Jensen-Shannon Divergence).
- Different learning methods need different numbers of learning step until convergence.
- Learning despite hidden (covert) information.

Conclusions

The role of *errors* $=$ the results of imperfect mental computation.

- OT as a model of competence (static knowledge), Simulated Annealing for OT as a model of the (eventually erroneous) computation in the brain (performance).
- “Performance errors” as driving force behind language change.
- Language learning until convergence on performance patterns (measured using Jensen-Shannon Divergence).
- Different learning methods need different numbers of learning step until convergence.
- Learning despite hidden (covert) information.

Conclusions

The role of *errors* = the results of imperfect mental computation.

- OT as a model of competence (static knowledge), Simulated Annealing for OT as a model of the (eventually erroneous) computation in the brain (performance).
- “Performance errors” as driving force behind language change.
- Language learning until convergence on performance patterns (measured using Jensen-Shannon Divergence).
- Different learning methods need different numbers of learning step until convergence.
- Learning despite hidden (covert) information.

Thank you for your attention!

Tamás Biró:
`t.s.biro@uva.nl`

ot kit 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