Evolutionary Optimality Theory

Stanford University
December 6, 2002

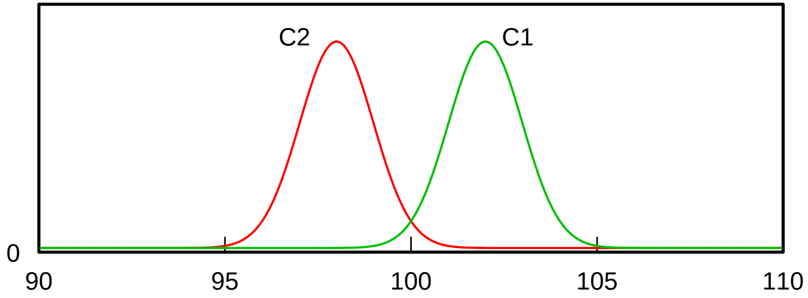
Gerhard Jäger
Zentrum für Allgemeine Sprachwissenschaft Berlin
University of Potsdam
`jaeger@zas.gwz-berlin.de`
www.ling.uni-potsdam.de/~jaeger/

1. Overview

- Stochastic Optimality Theory
- unidirectional learning
- bidirectional learning and iconicity
- Differential Case Marking

2. Stochastic Optimality Theory (StOT)

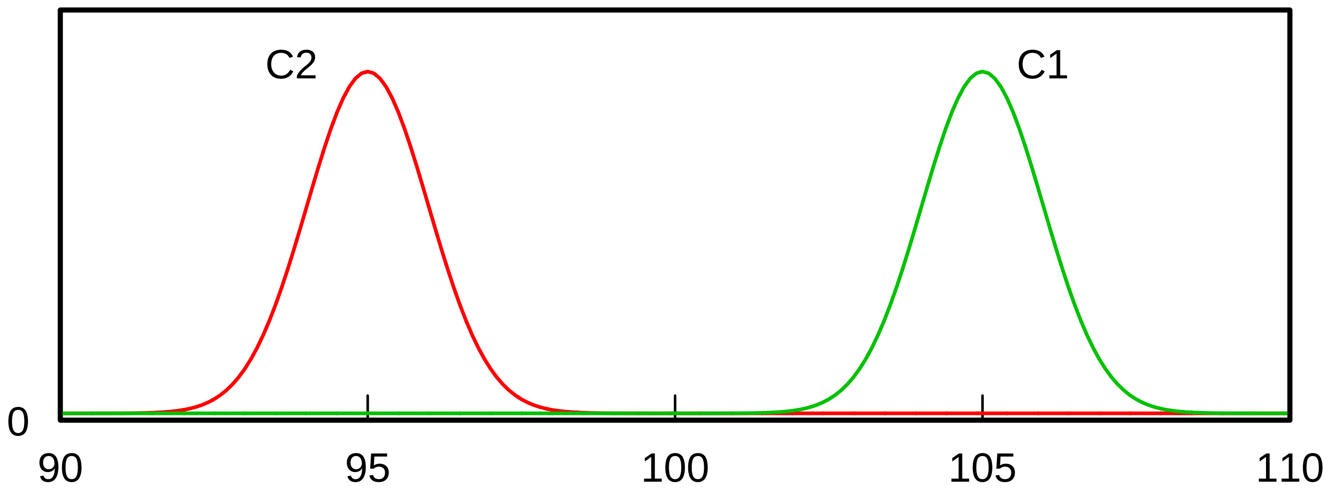
- probabilistic grammar
- assigns probability distribution over possible meanings for a given form (and vice versa)
- Two modifications of standard OT (cf. Boersma 1998)
  1. **constraint ranking on a continuous scale** distance between constraints matters
  2. **stochastic evaluation** actual ordering of constraints varies, with probabilities depending on continuous ranking



- Absolute size of the distance between conflicting constraints determines their interaction:
  - difference between mean values > 10 units:

C$_1$ dominates C$_2$ categorically

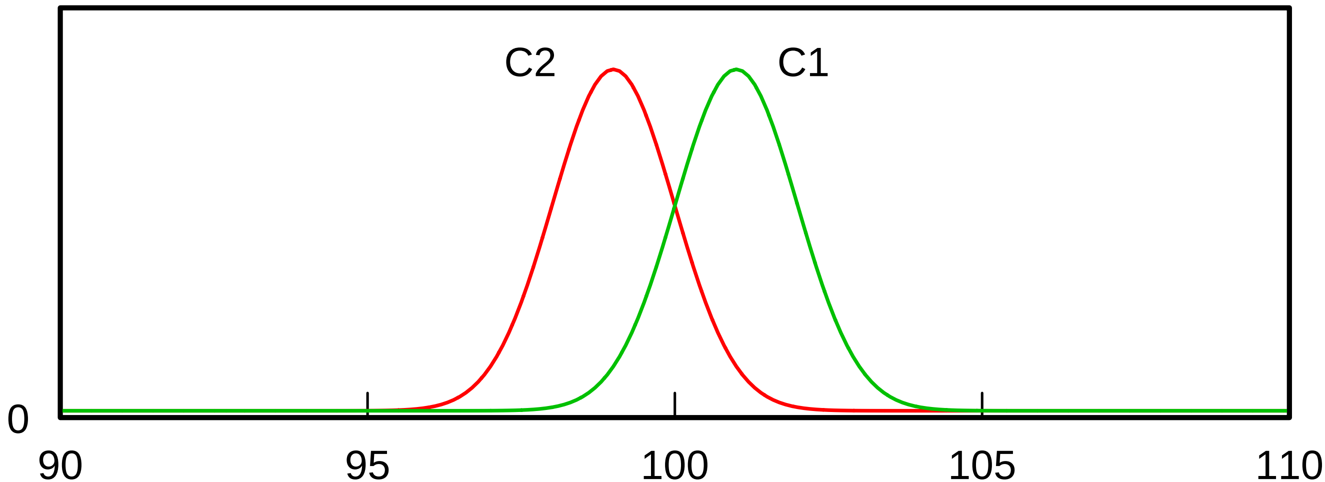
$$p(C_2 > C_1) < 10^{-10}$$



- difference ≈ 2:

  preference for obeying C_1, but obeying C_2 is still grammatical

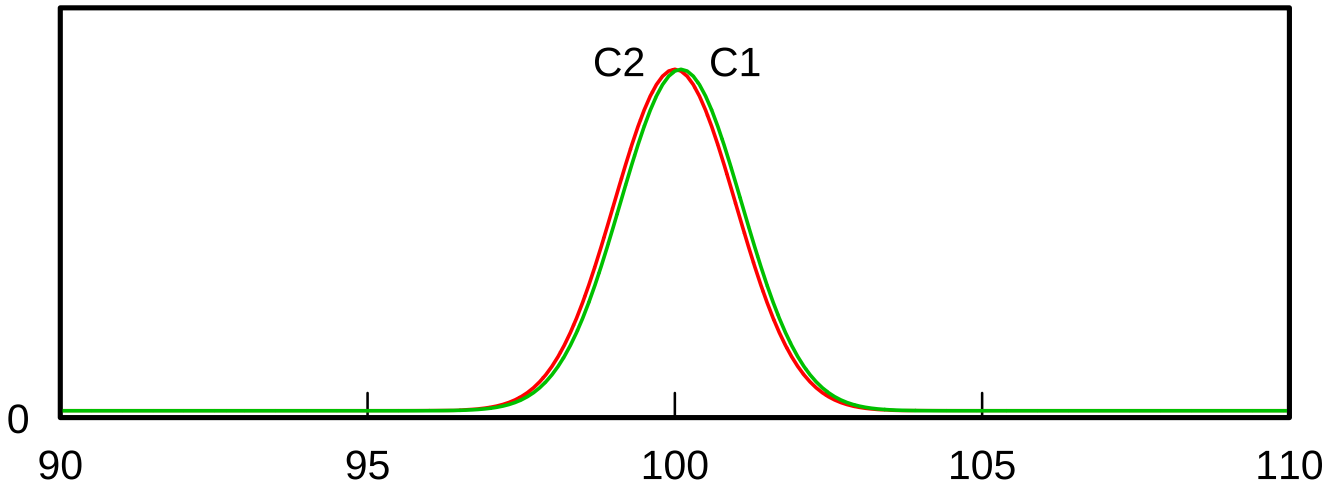
$$p(C_2 > C_1) \approx 30\%$$



- Both constraints are roughly equally ranked:

free variation

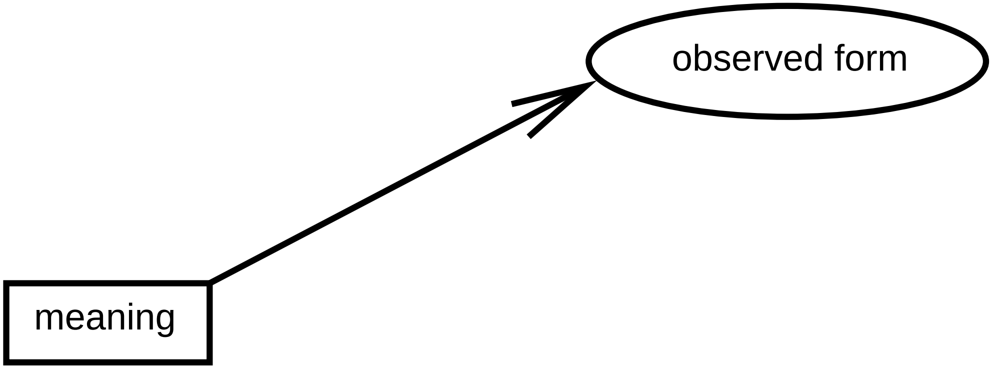
$$p(C_2 > C_1) = 50\%$$

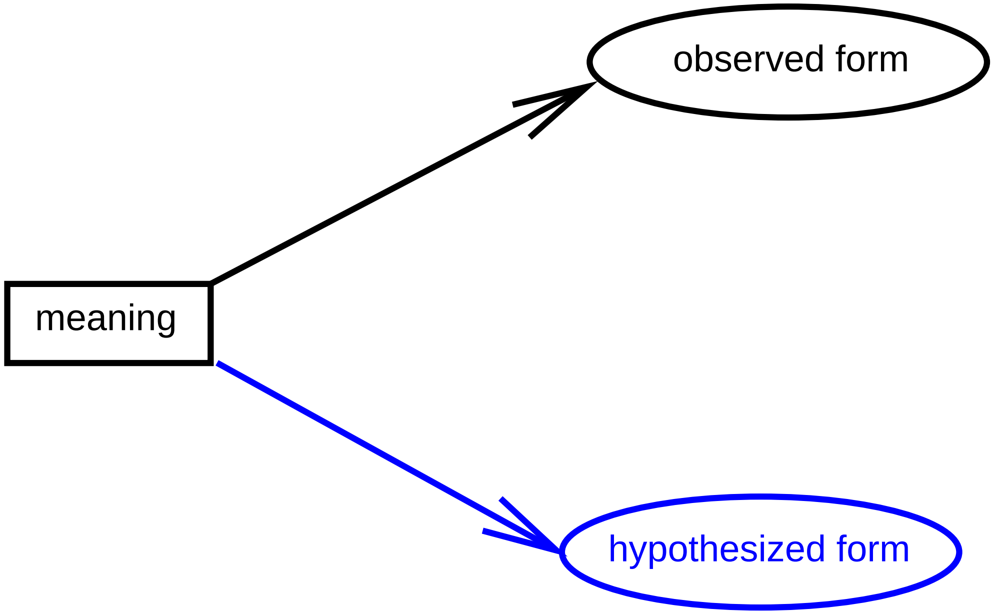


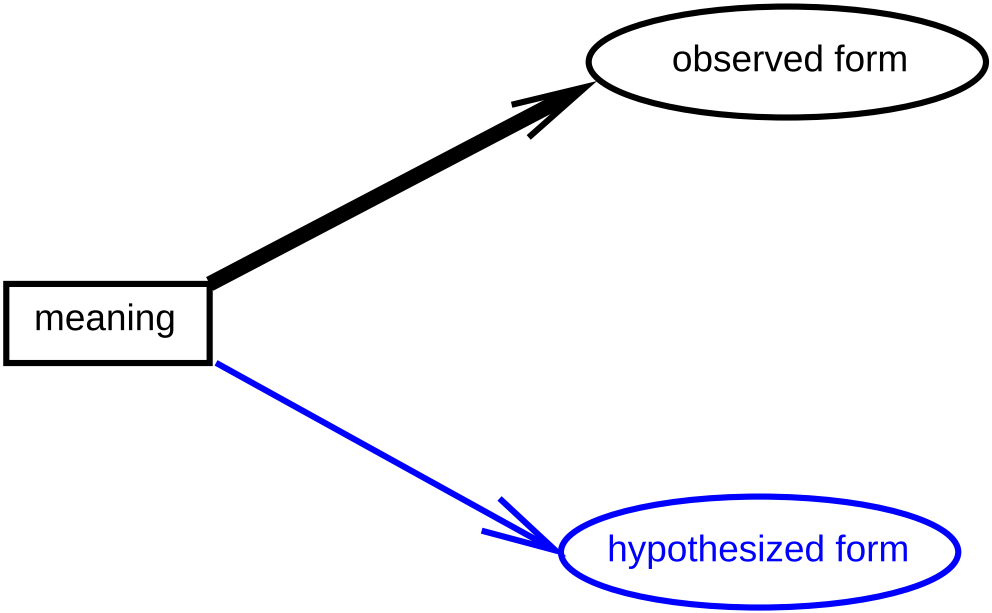
3. The Gradual Learning Algorithm (GLA)

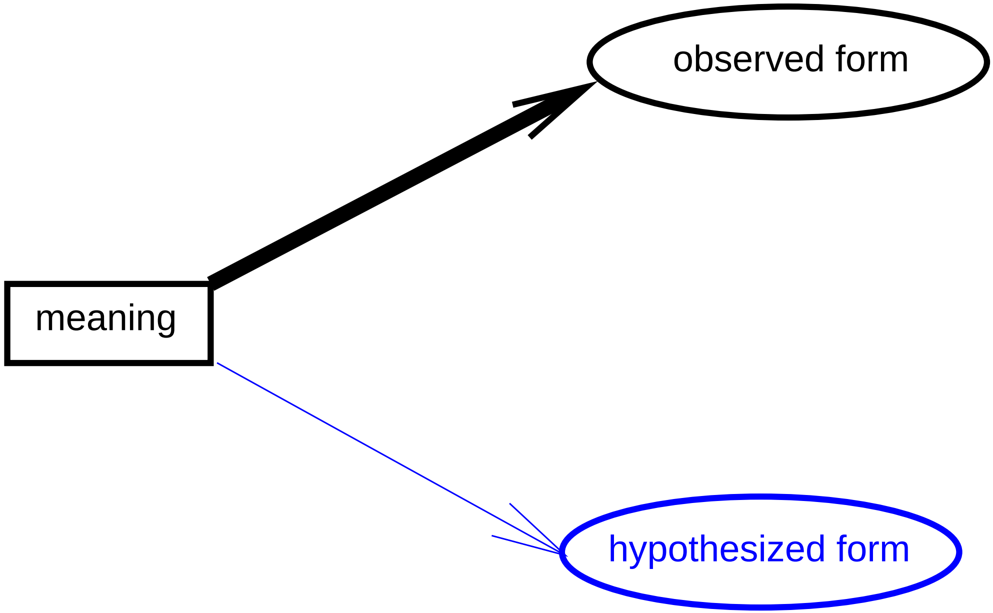
- Function from (analyzed) corpus to StOT-Grammar
- error-driven
- outputs grammar that reproduces statistical patterns in the training corpus

$\boxed{\text{meaning}}$


observed form
meaning


observed form
meaning
hypothesized form


observed form
meaning
hypothesized form


observed form
meaning
hypothesized form

Six stages:

- **Initial state** Constraints begin with a ranking that is hypothesized by the linguist (and plays no significant role for learning result)
- **Step 1: A datum** Algorithm is presented with a learning datum—a fully specified input-output pair $\langle i, o \rangle$
- **Step 2: Generation**
  - For each constraint, a noise value is drawn from the normal distribution and added to its current ranking. This yields the *selection point*.
  - Constraints are ranked by descending order of the selection points. This yields a linear order of the constraints.
  - Based on this constraint ranking, the grammar generates an output o' for the input i.

- **Step 3: Comparison** If $o = o'$, nothing happens. Otherwise, the algorithm compares the constraint violations of the learning datum $\langle i, o \rangle$ with the self-generated pair $\langle i, o' \rangle$.
- **Step 5: Adjustment**
  - All constraints that favor $\langle i, o \rangle$ over $\langle i, o' \rangle$ are *increased* by some small predefined numerical amount ("plasticity").
  - All constraints that favor $\langle i, o' \rangle$ over $\langle i, o \rangle$ are *decreased* by the plasticity value.
- **Final state** Steps 1 – 4 are repeated until the constraint values stabilize.

4. Bidirectionality

4.1. Bidirectional evaluation

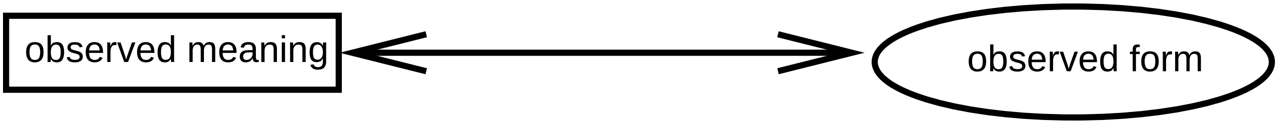
- OT-grammar defines ranking of possible forms for a given meaning and vice versa
- StOT-grammar defines probability distribution over OT-grammars
- licit meaning-form association for a given grammar must be optimal for both speaker and hearer (cf. Blutner 2000, Zeevat 2000, Beaver 2000)

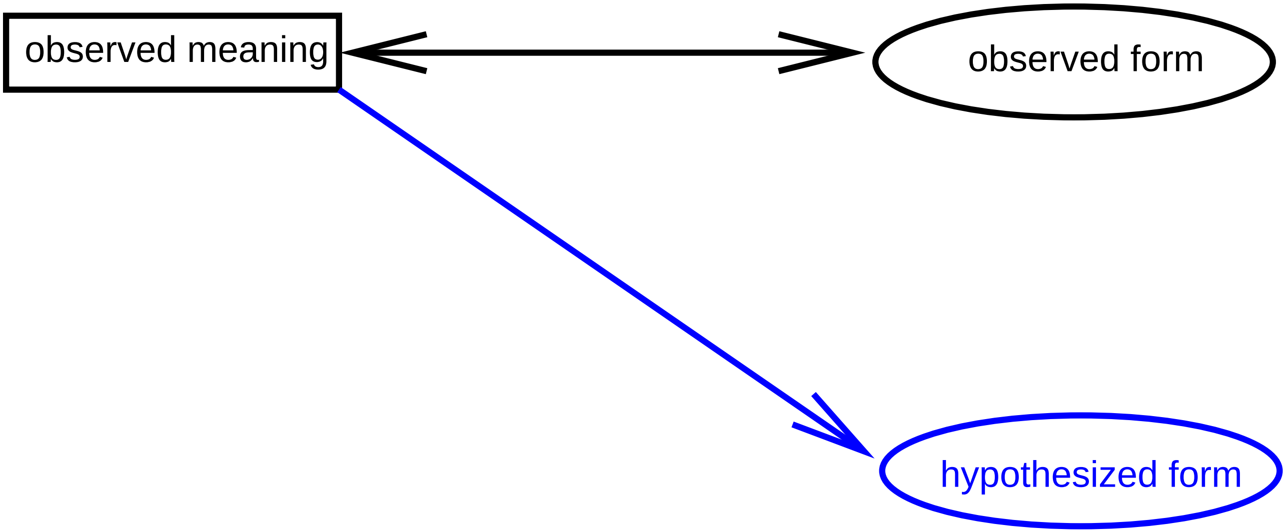
Definition 1

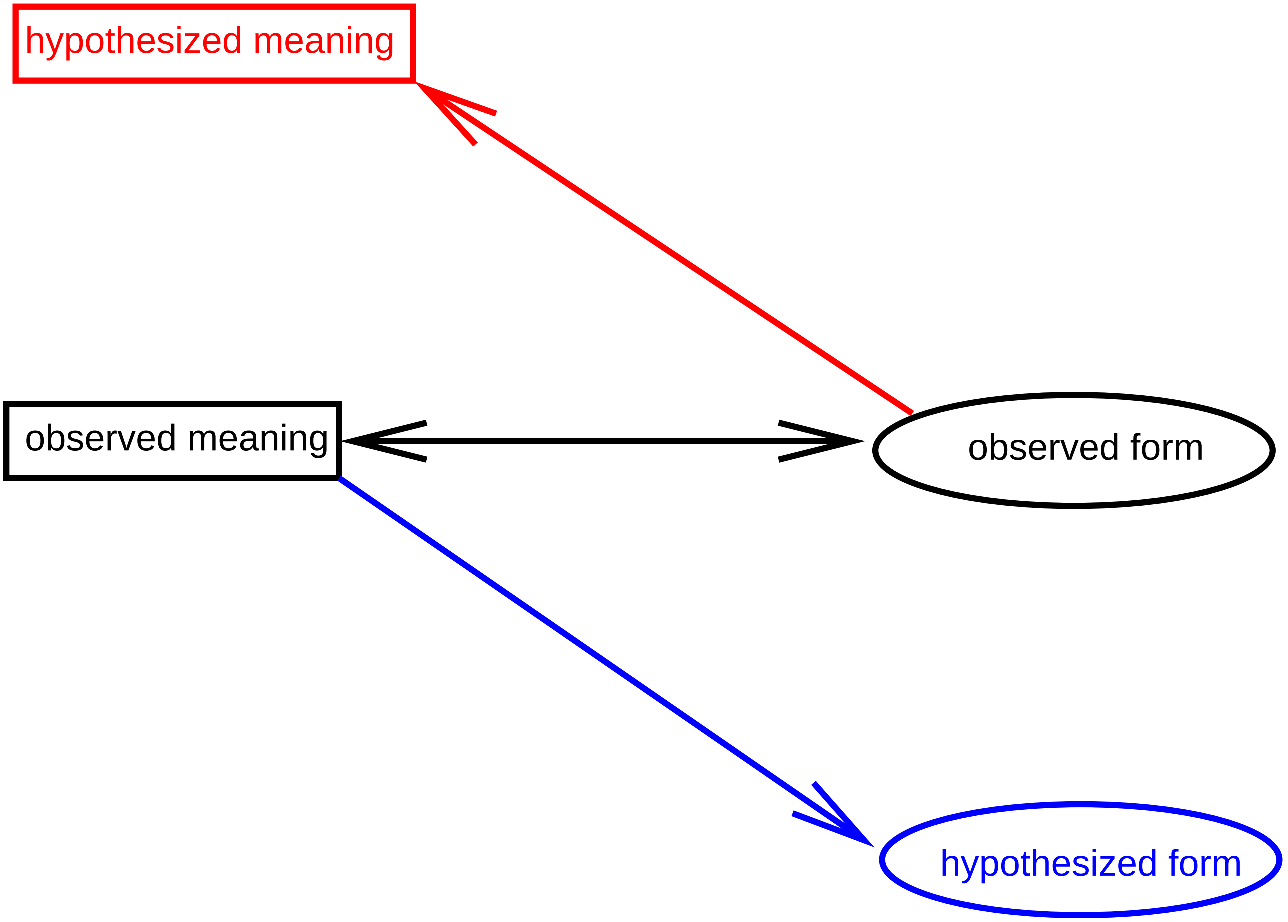
- *A form-meaning pair* $\langle f, m \rangle$ *is* hearer-optimal *iff* $\langle f, m \rangle \in \mathbf{GEN}$ *and there is no alternative meaning* m' *such that* $\langle f, m' \rangle \in \mathbf{GEN}$ *and* $\langle f, m' \rangle < \langle f, m \rangle$*.*
- *A form-meaning pair* $\langle f, m \rangle$ *is* optimal *iff either it is hearer-optimal and there is no alternative form* f' *such that* $\langle f', m \rangle$ *is hearer-optimal and* $\langle f', m \rangle < \langle f, m \rangle$*, or there is no hearer-optimal* $\langle f', m \rangle$*, and there is no* $\langle f', m \rangle \in \mathbf{GEN}$ *such that* $\langle f', m \rangle < \langle f, m \rangle$*.*

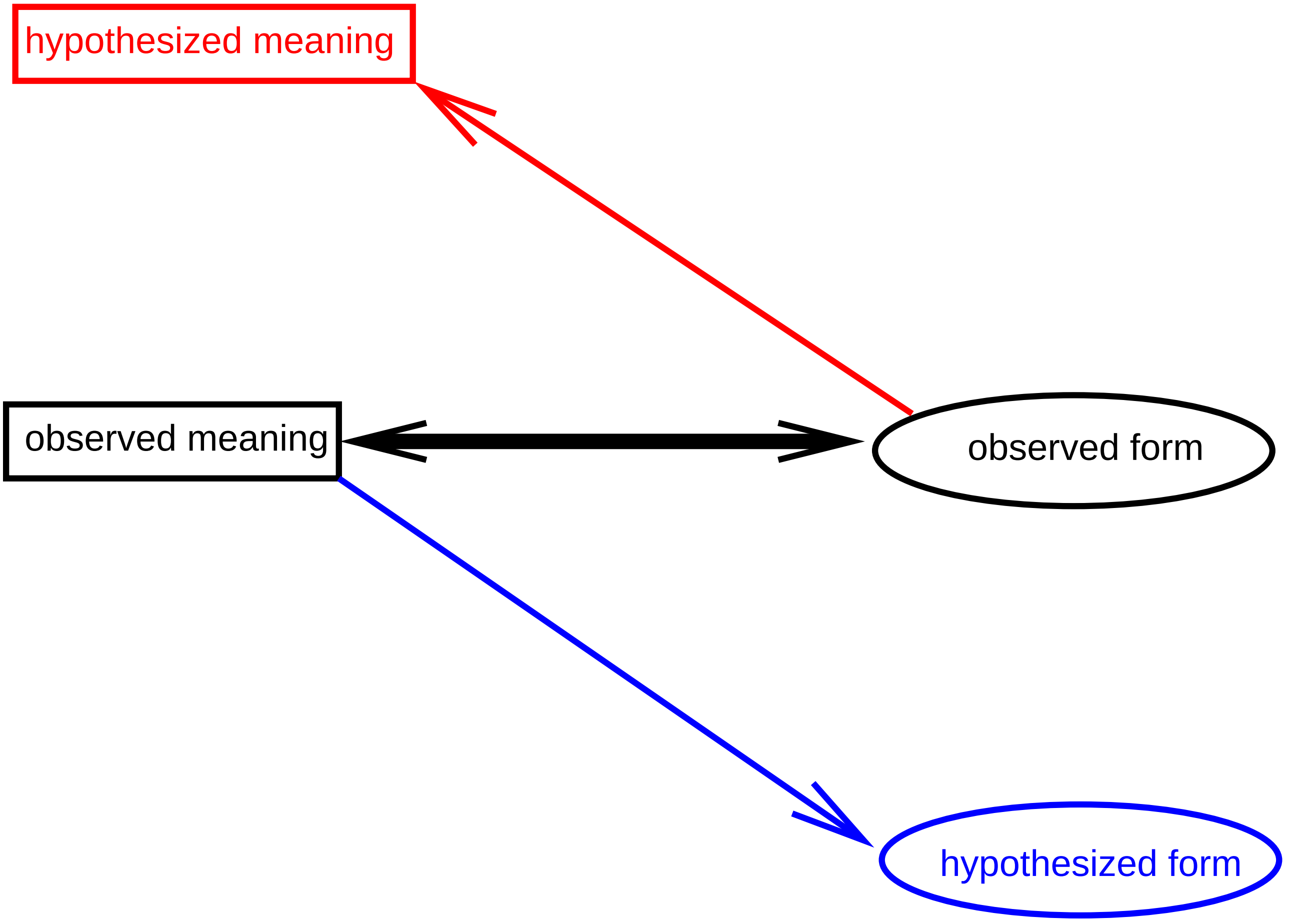
4.2. Bidirectional learning

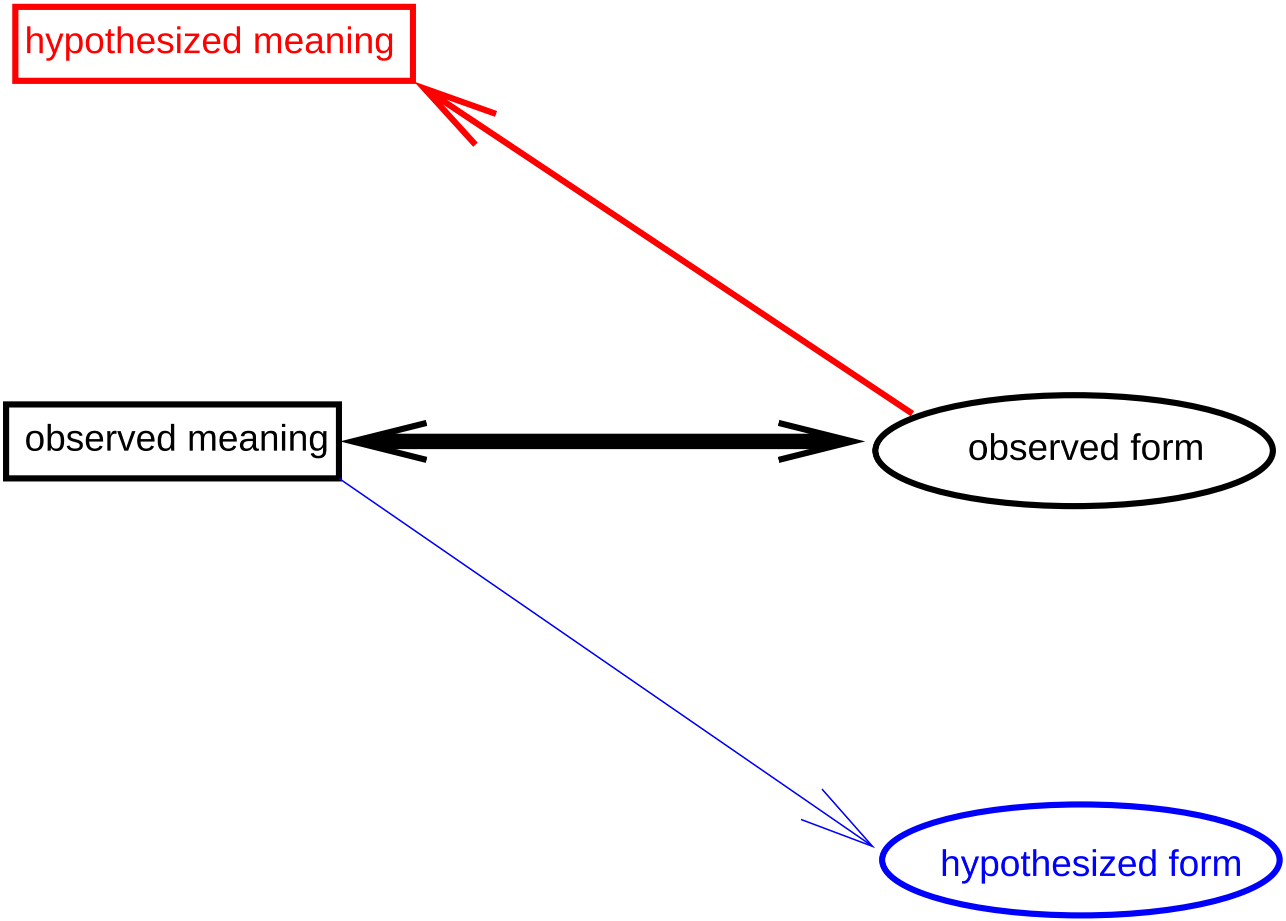
- unidirectional learning (Tesar and Smolensky, Boersma):
  - learning triggered by insight: *Oops, I hadn't said it like this!*
  - "luxury problem" (Zeevat, p.c.)
- more urgent trigger for learning:
  - learning trigger: *I don't understand you guys!*
  - requires comparison of observed with hypothesized interpretation
- together: **bidirectional learning**

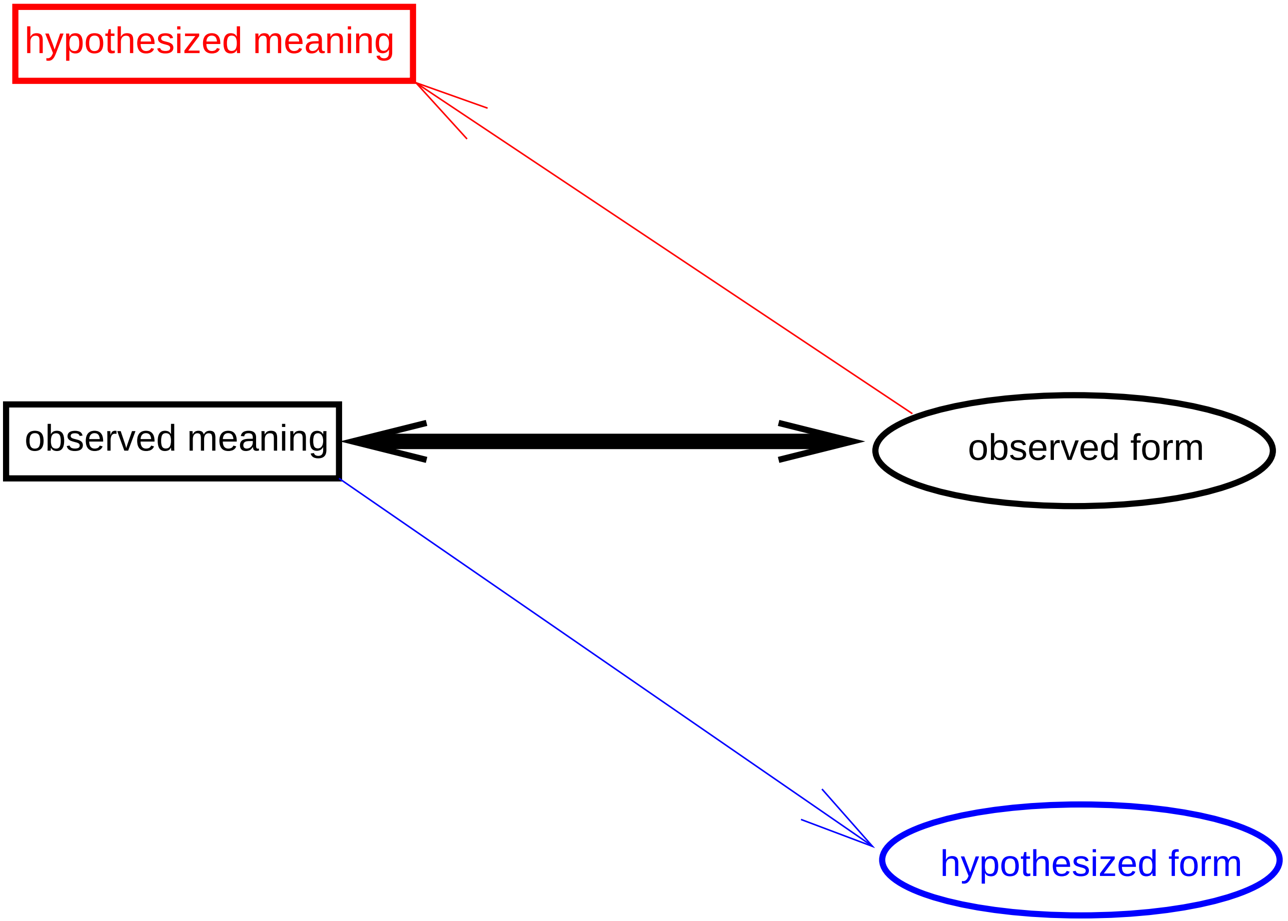

observed meaning
observed form


observed meaning
observed form
hypothesized form


hypothesized meaning
observed meaning
observed form
hypothesized form


hypothesized meaning
observed meaning
observed form
hypothesized form


hypothesized meaning
observed meaning
observed form
hypothesized form



hypothesized meaning
observed meaning
observed form
hypothesized form

- Bidirectional GLA (BiGLA):
  - Evaluation according to bidirectional optimization as above
  - Both speaker and hearer learn
  - Speaker compares different forms
  - Hearer compares different meanings

- **Initial state** All constraint values are set to 0.
- **Step 1: A datum** The algorithm is presented with a learning datum—a fully specified input-output pair $\langle f, m \rangle$.
- **Step 2: Generation**
  - For each constraint, a noise value is drawn from a normal distribution and added to its current ranking. This yields the *selection point*.
  - Constraints are ranked by descending order of the selection points. This yields a linear order of the constraints.
  - Based on this constraint ranking, the grammar generates two pairs $\langle f', m \rangle$ and $\langle f, m' \rangle$ that are both bidirectionally optimal.

- **Step 3.1: Comparison of forms** If $f = f'$, nothing happens. Otherwise, the algorithm compares the constraint violations of the learning datum $\langle f, m \rangle$ with the self-generated pair $\langle f', m \rangle$.
- **Step 3.2: Comparison of meanings** If $m = m'$, nothing happens. Otherwise, the algorithm compares the constraint violations of the learning datum $\langle f, m \rangle$ with the self-generated pair $\langle f, m' \rangle$.
- **Step 4: Adjustment**
  - All constraints that favor $\langle f, m \rangle$ over $\langle f', m \rangle$ are *increased* by the plasticity value.
  - All constraints that favor $\langle f', m \rangle$ over $\langle f, m \rangle$ are *decreased* by the plasticity value.
  - All constraints that favor $\langle f, m \rangle$ over $\langle f, m' \rangle$ are *increased* by the plasticity value.
  - All constraints that favor $\langle f, m' \rangle$ over $\langle f, m \rangle$ are *decreased* by the plasticity value.
- **Final state** Steps 1 – 4 are repeated until the constraint values stabilize.

The E/I-model of language evolution

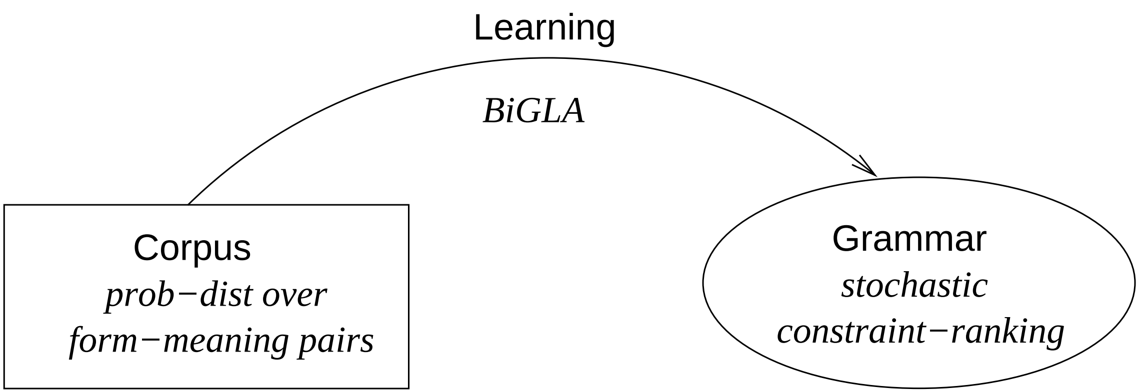
(cf. Kirby and Hurford 2001)

Corpus
prob−dist over
form−meaning pairs

$$\forall m : \quad \sum_f p(f, m) = \text{const}$$

The E/I-model of language evolution

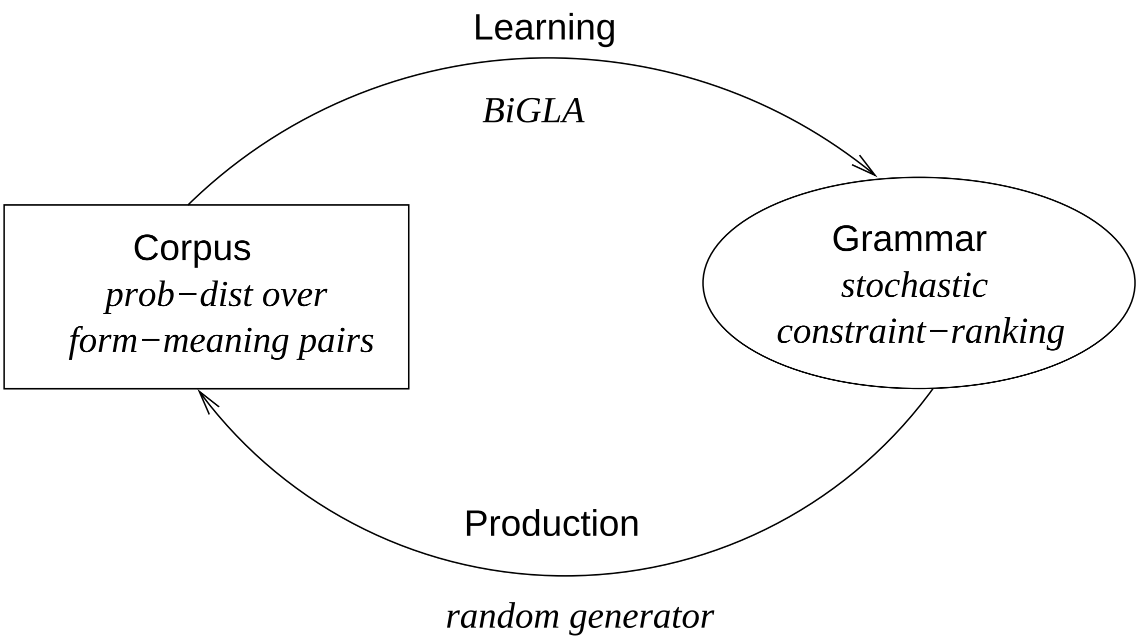
(cf. Kirby and Hurford 2001)



$$\forall m : \quad \sum_f p(f, m) = \text{const}$$

The E/I-model of language evolution

(cf. Kirby and Hurford 2001)



$$\forall m : \quad \sum_f p(f, m) = \text{const}$$

4.3. An experiment

- two meanings, a and b
- two forms, 1 and 2
- each form-meaning pair is admitted by GEN
- each form meaning pair is penalized by one constraint
- form 2 is more complex than form 2
- covered by constraint *2 (“Avoid 2!”)

	*a1	*a2	*b1	*b2	*2
$a1$	$*$				
$a2$		$*$			$*$
$b1$			$*$		
$b2$				$*$	$*$

- fix frequencies of the four candidates
- run BiGLA on this "training corpus"
- use the acquired grammar to generate sample of the acquired language
- keep the total frequencies of the two meanings constant

`http://www.ling.uni-potsdam.de/~jaeger/evolOT`

Emergence of Iconicity

$$\mathit{freq}(a) > \mathit{freq}(b)$$

$$\rightsquigarrow$$

$$p(1|a) \gg p(2|a)$$
$$p(2|b) \gg p(1|b)$$

5. Differential Case Marking

- three basic syntactic functions of NPs:
  - subject of intransitive verb (S)
  - subject of transitive verb (A)
  - direct object of transitive verb (O)
- case of S: *zero (= nominative/absolutive)*
- case of A: *zero or ergative*
- case O: *zero or accusative*
- choice *zero vs erg* and *zero vs acc* language specific
- Differential Case Marking (DCM): case is correlated with animacy, definiteness, specificity, person etc.

- universal tendencies (cf. Aissen 2000)

$$p(\text{erg}|\text{A,-anim}) > p(\text{erg}|\text{A,+anim})$$
$$p(\text{acc}|\text{O,+anim}) > p(\text{acc}|\text{O,-anim})$$

- similar correlations for definiteness etc.
- functional motivation (cf. Zeevat and Jäger 2002)
- rare forms are more likely to be case marked than frequent ones

$$\textit{freq}(A, +anim) > \textit{freq}(A, -anim)$$
$$\textit{freq}(O, -anim) > \textit{freq}(O, +anim)$$

DCM and OT

- Aissen proposes the following constraint system to deal with DCM:

1. *(su/a/Z): *Case mark animate subjects!*
2. *(su/i/Z): *Case mark inanimate subjects!*
3. *(ob/a/Z): *Case mark animate objects!*
4. *(ob/i/Z): *Case mark inanimate objects!*
5. *STRUC: *Avoid case marking!*

- universal case marking patterns correspond to universal constraint sub-hierarchies.:

$$\begin{gathered} *(\mathrm{su/i/z}) \gg *(\mathrm{su/a/z}) \\ *(\mathrm{ob/a/z}) \gg *(\mathrm{ob/i/z}) \end{gathered}$$

Functional OT

- Hypothesis: Aissen's sub-hierarchies are not innate, but result of functional pressure
- basic intuition: animate subjects are more frequent than inanimate ones $\rightsquigarrow$ animate subjects have stronger impact on learning

More experiments

- Suppose: training corpus with
  - only simple transitive clauses
  - relative frequencies of clause types wrt. animacy of subject and object are as in naturally occuring conversations
  - exactly 50 % of all NPs are (faithfully) case marked (ergative or accusative)
  - no statistic correlation between animacy and case marking
- clause type frequencies in SAMTAL (corpus of spoken Swedisch):

	subj/anim	subj/inanim
obj/anim	300	17
obj/inanim	2648	186

- additional constraints

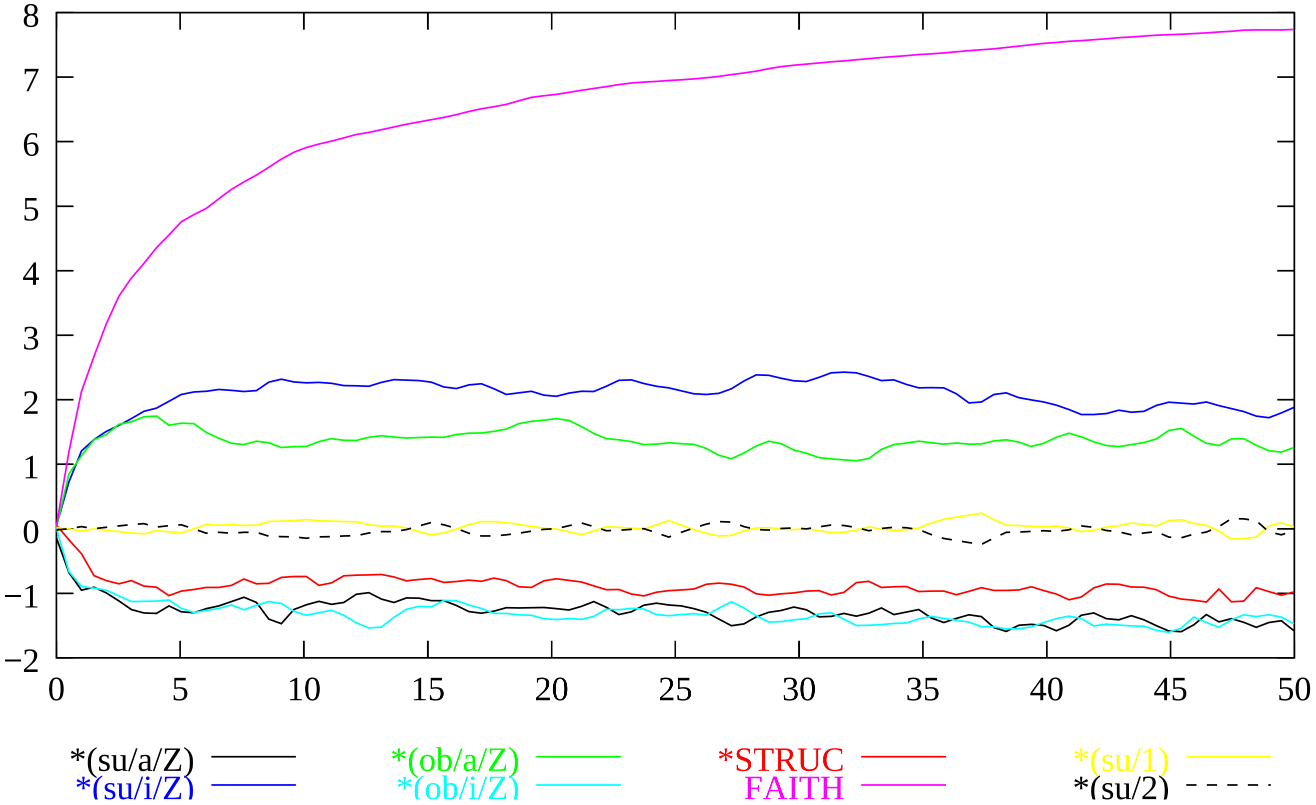
6. FAITH: *Interpret ergative as subject and accusative as object!*
7. *(su/2): *NP1 is subject and NP2 object.*
8. *(su/1): *NP2 is subject and NP1 object.*

- relative frequencies in training corpus (in %)

	E-E	E-A	E-Z	A-E	A-A	A-Z	Z-E	Z-A	Z-Z
su/a-ob/a	0.0	1.19	1.19	0.0	0.0	0.0	0.0	1.19	1.19
su/a-ob/i	0.0	10.50	10.50	0.0	0.0	0.0	0.0	10.50	10.50
su/i-ob/a	0.0	0.07	0.07	0.0	0.0	0.0	0.0	0.07	0.07
su/i-ob/i	0.0	0.74	0.74	0.0	0.0	0.0	0.0	0.74	0.74
ob/a-su/a	0.0	0.0	0.0	1.19	0.0	1.19	1.19	0.0	1.19
ob/a-su/i	0.0	0.0	0.0	0.07	0.0	0.07	0.07	0.0	0.07
ob/i-su/a	0.0	0.0	0.0	10.50	0.0	10.50	10.50	0.0	10.50
ob/i-su/i	0.0	0.0	0.0	0.74	0.0	0.74	0.74	0.0	0.74

E ... ergative
A ... accusative
Z ... zero marking
a ... animate
i ... inanimate
X-Y ... NP1 has features X and NP2 features Y

The learning process



- acquired grammar:

*(su/a/Z):	-1.58
*(su/i/Z):	1.88
*(ob/a/Z):	1.26
*(ob/i/Z):	-1.47
*STRUC:	-0.98
FAITH:	7.74
*(su/1):	0.03
*(su/2):	-0.03

- Emergence of Aissen's sub-hierarchies

$$*(\mathrm{su/i/Z}) \gg *(\mathrm{su/a/Z})$$
$$*(\mathrm{ob/a/Z}) \gg *(\mathrm{ob/i/Z})$$

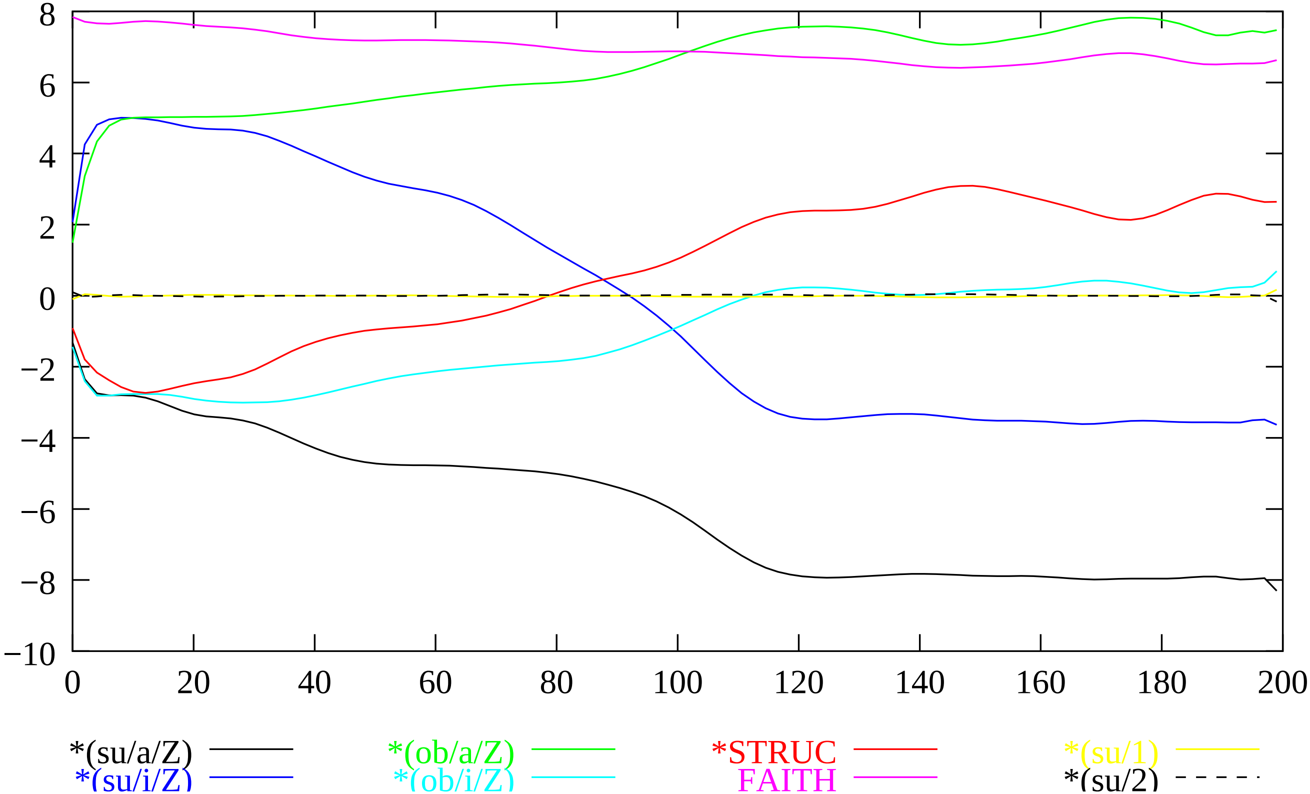
- can be used to generate new sample corpus
- probality distribution over meanings from SAMTAL are maintained

	E-E	E-A	E-Z	A-E	A-A	A-Z	Z-E	Z-A	Z-Z
su/a-ob/a	0.0	1.84	0.19	0.0	0.0	0.0	0.0	2.23	0.40
su/a-ob/i	0.0	11.09	7.35	0.0	0.0	0.0	0.0	8.52	15.04
su/i-ob/a	0.0	0.21	0.05	0.0	0.0	0.0	0.0	0.02	0.0
su/i-ob/i	0.0	1.22	1.47	0.0	0.0	0.0	0.0	0.11	0.16
ob/a-su/a	0.0	0.0	0.0	2.0	0.0	2.12	0.25	0.0	0.39
ob/a-su/i	0.0	0.0	0.0	0.18	0.0	0.3	0.07	0.0	0.0
ob/i-su/a	0.0	0.0	0.0	11.15	0.0	8.40	7.69	0.0	14.76
ob/i-su/i	0.0	0.0	0.0	1.17	0.0	0.09	1.47	0.0	0.23

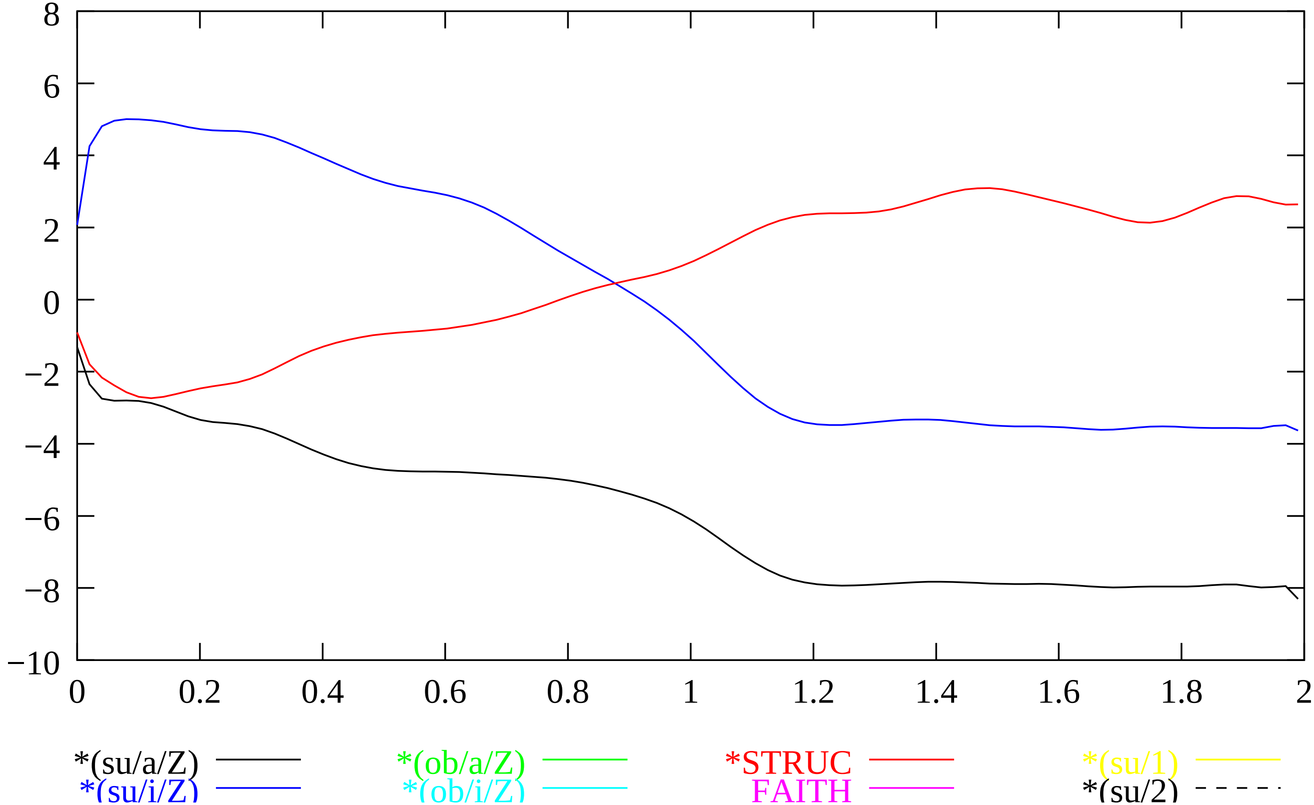
6. The next generation

- can be repeated:
  - resulting sample corpus is used as training corpus for next run of BiGLA
  - acquired grammar is used to generate next sample corpus
  - relative frequencies of inputs (meanings) are kept constant
  - conditional probabilities p(form | meaning) may change from generation to generation

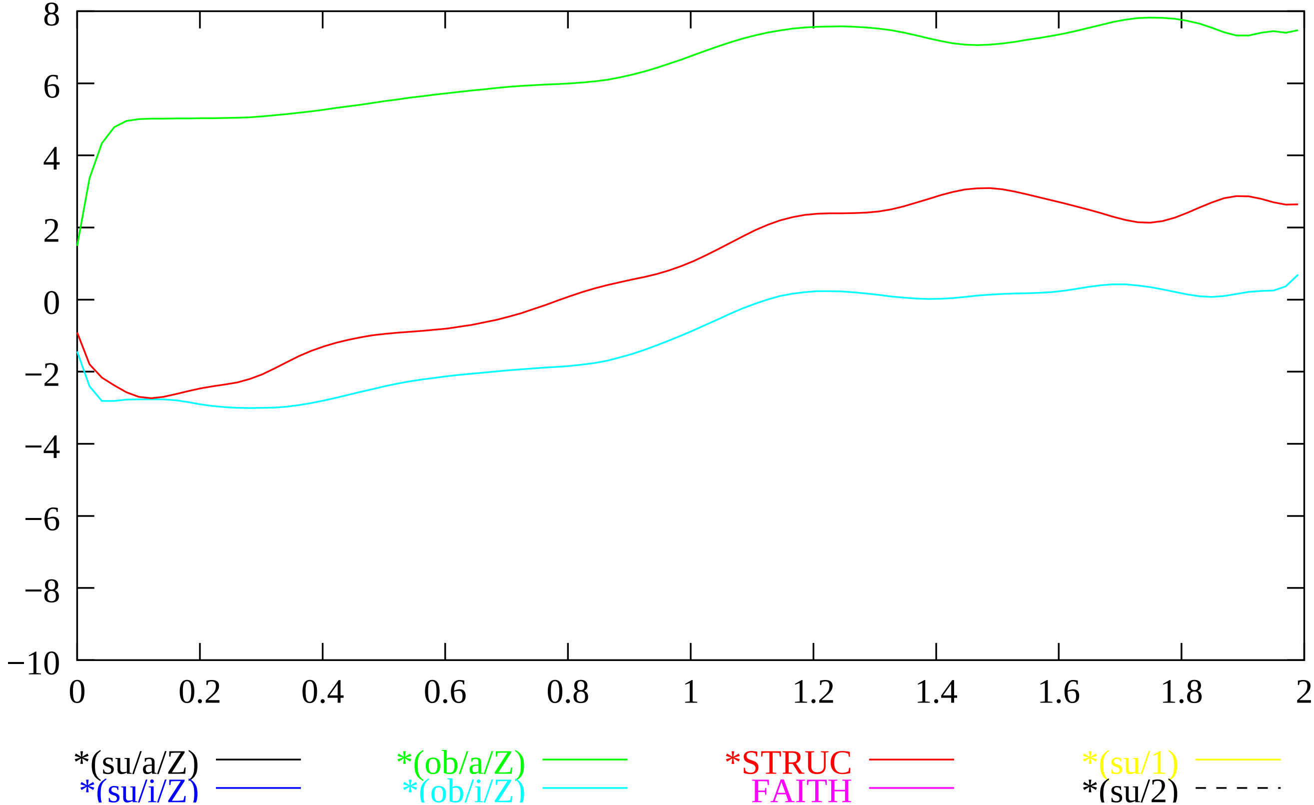
- starting with corpus given above; 200 generations
- long phase of split ergativity, followed by transition toward accusative system with DOM



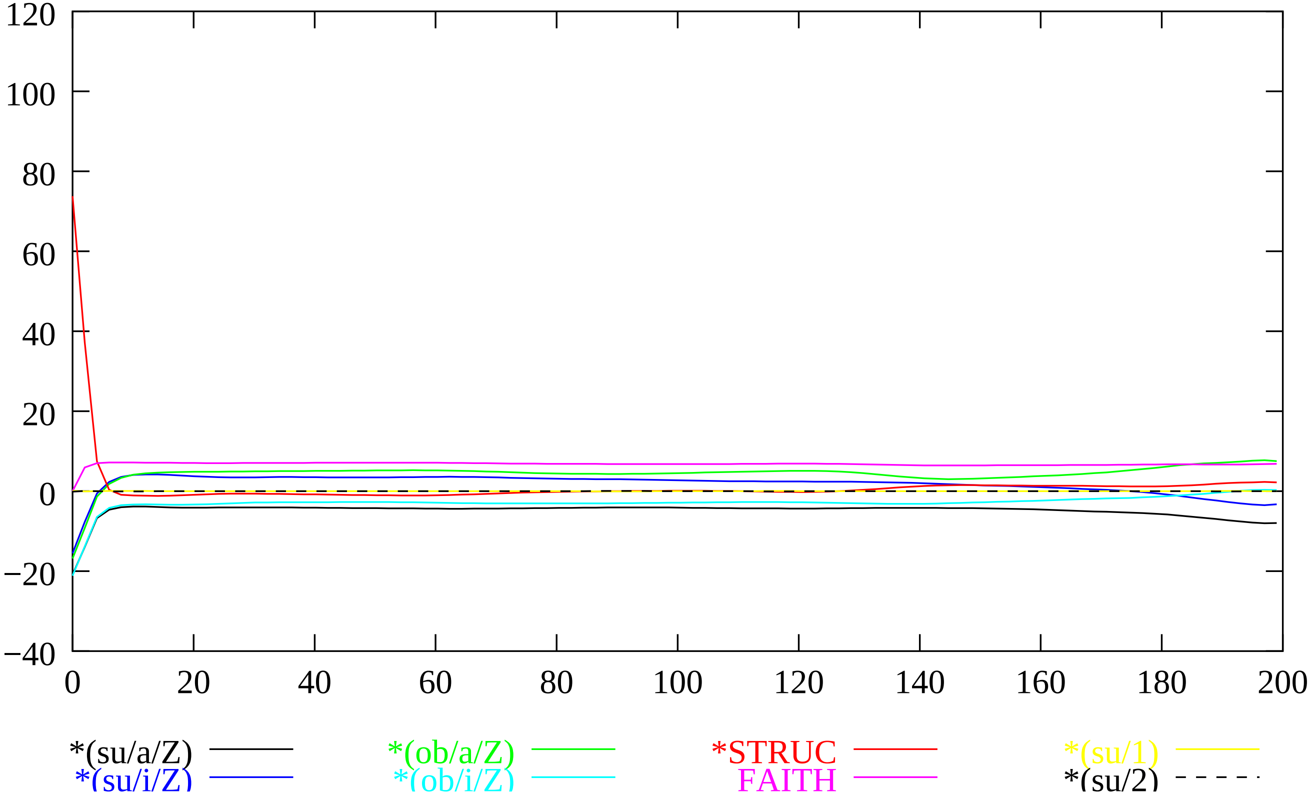
- first subhierarchy *(su/i/Z) ≫ *(su/a/Z)



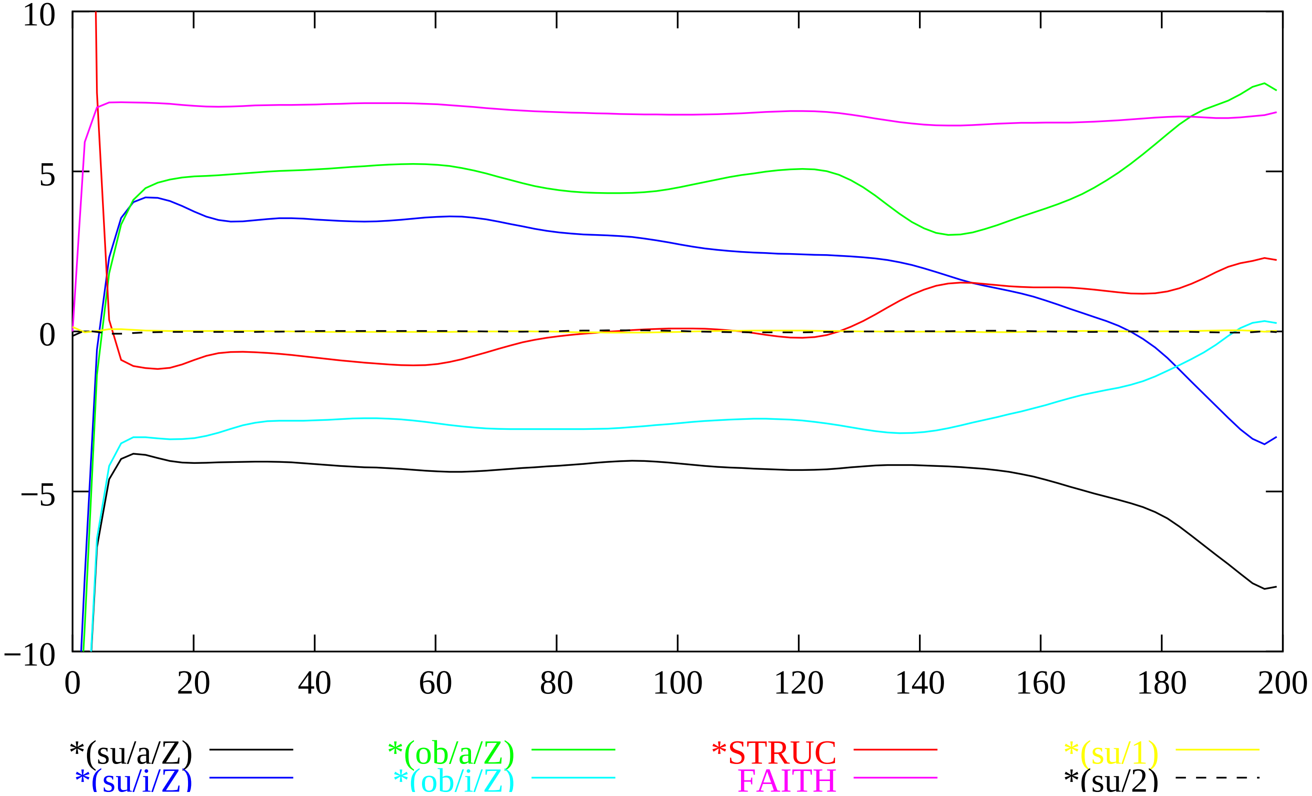
- second subhierarchy *(ob/a/Z) ≫ *(ob/i/Z)



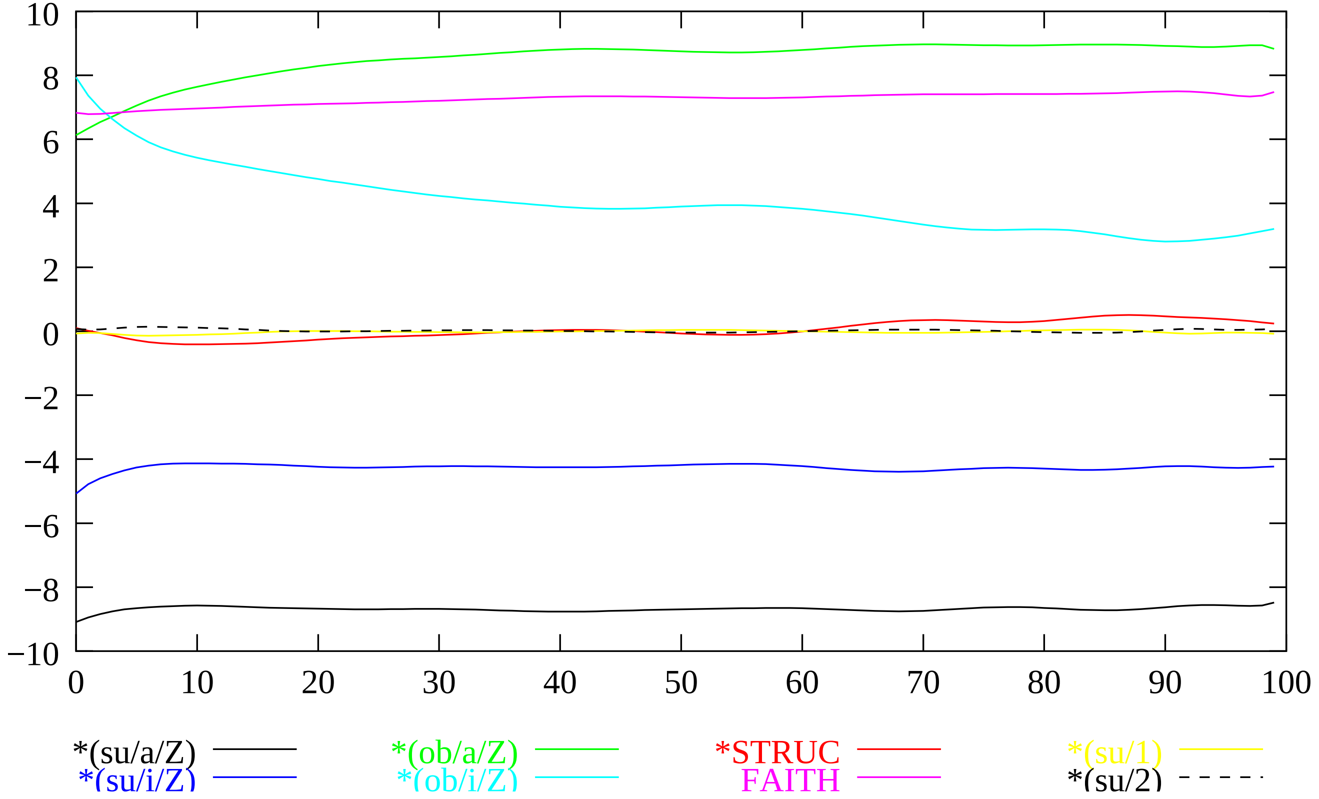
- Spread: initial corpus has no case morphemes (but GEN admits them)



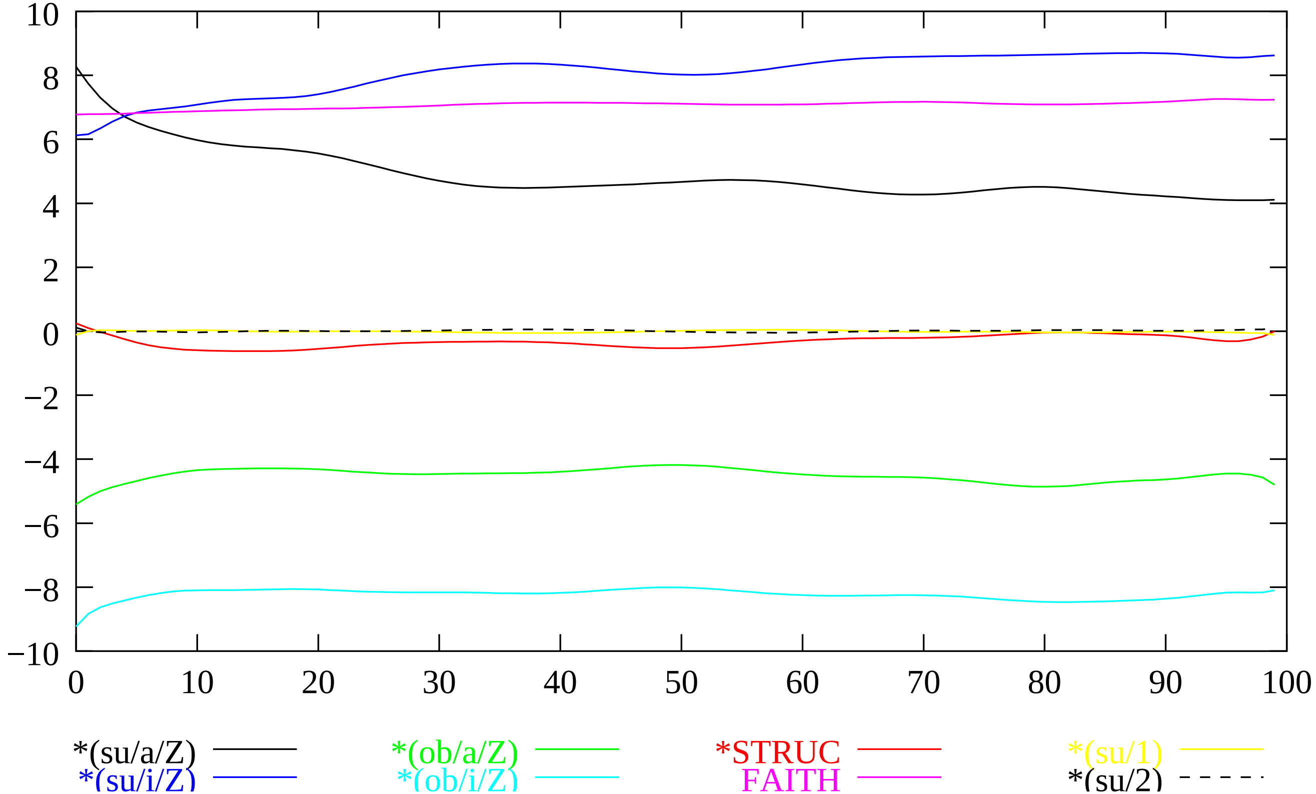
- Zooming in:
- similar diachronic tendencies as above



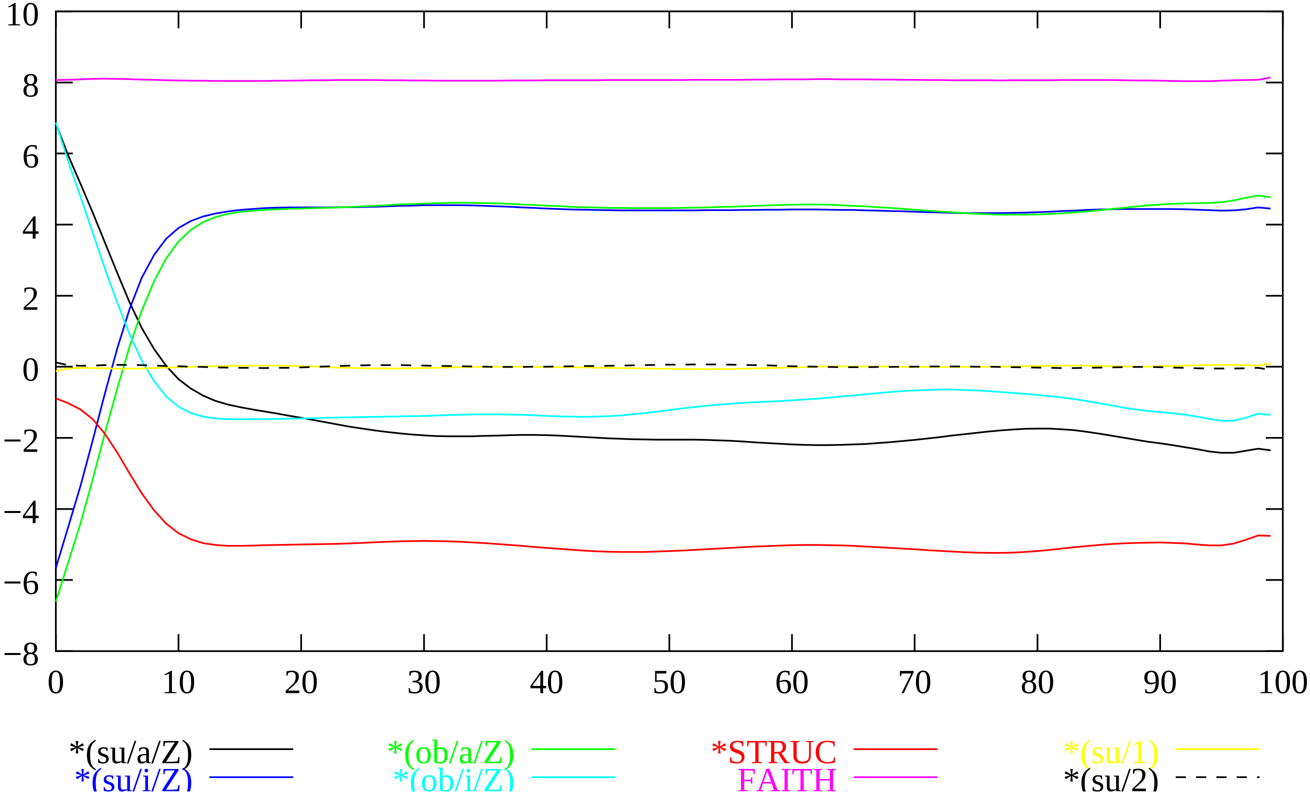
- Pure systems are diachronically stable
- starting with nominative-accusative



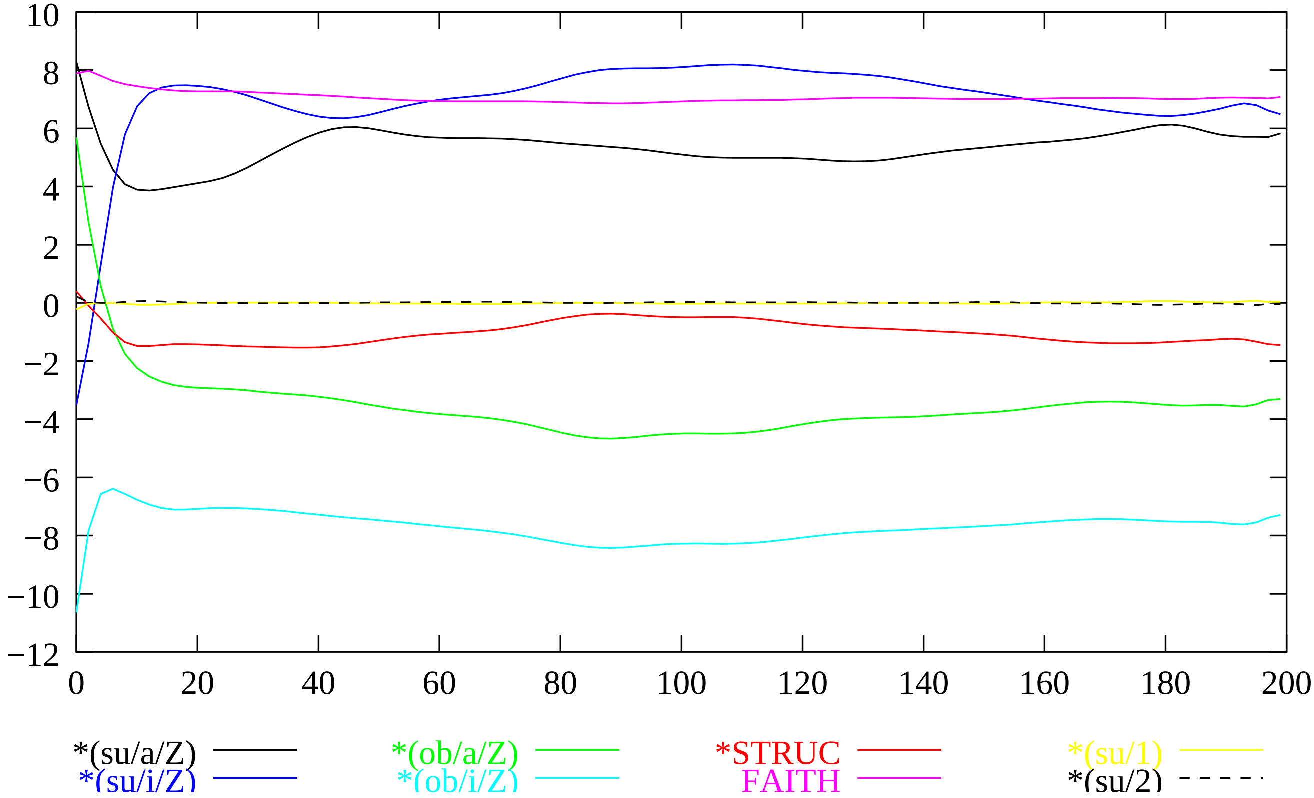
• starting with absolutive-ergative



- Violations of the Aissen-universals are possible, but extremely unstable
- starting with anti-DCM:



- starting with obligatory case marking of animate NPs (and no case marking on inanimate ones):



7. Conclusion

- Bidirectional GLA is sensitive to probabilities of meanings in training corpus
- establishes connection between statistical patterns of language use and competence grammar
- imperfect learning: acquired language might differ slightly from training language
- diachronic drift
- stable vs. unstable grammars
- can be applied to typology and historical linguistics

Contents