Simulating language change with Functional OT

Gerhard Jäger
University of Potsdam

http://www.ling.uni-potsdam.de/˜jaeger

1 Introduction

The research reported here is a reaction to recent work by Judith Aissen on the typology of case marking systems within Optimality Theory (OT). Aissen (2000) explains certain linguistic universals by assuming universal sub-hierarchies of OT constraints. I found this intriguing but unsatisfactory because these rankings are obviously much better adapted to the statistical patterns of actual language use than their inverse. Paul Boersma's Gradual Learning Algorithm for Stochastic OT is well-suited to establish a connection between frequencies of utterance types and constraint rankings, but it has to operate in generative and interpretive direction simultaneously. Intuitively, Aissen's sub-hierarchies are easier to acquire for such a bidirectional learning algorithm than possible (but non-existent) alternatives. This can be made precise by employing Iterated Learning in the sense of Kirby and Hurford (2002)—certain constraint ranking patterns are universal because they are evolutionary invariant. After briefly reviewing the empirical phenomena to be dealt with, Aissen's OT account, and the assumptions on OT learning that I employ in my version of Iterated Learning, I will report a series of simulations that link Aissen's findings with results of quantitative studies. A fuller account can be found in Jäger (2002).

2 Differential Case Marking

It is a common feature of many case marking languages that some but not all objects are case marked. However, it is usually not entirely random which objects are marked and which aren't. Rather, case marking only applies to a morphologically or semantically well-defined class of NPs. Bossong (1985) calls this phenomenon "Differential Object Marking" (DOM). A common pattern is that all NPs from the top section of the *definiteness hierarchy* or the *animacy hierarchy* are case marked while those from the bottom section are not.

(1) a. personal pronoun > proper noun > definite full NP > indefinite NP
b. human > animate > inanimate

Differential case marking also frequently occurs with subjects. In contradistinction to DOM, DSM ("Differential Subject Marking") means that only instances of some *lower* segment of the definiteness/animacy hierarchy are case marked. DSM usually co-occurs with DOM within one language. This phenomenon is called *split ergativity*. These patterns of "Differential Case Marking" (DCM) can be represented as the result of aligning two scales—the scale of grammatical functions (subject vs. object) with some scale which classifies NPs according to substantive features like definiteness or animacy. Ranking the grammatical functions according to prominence leads to the binary scale

(2) Subj $>$ Obj

Harmonic alignment of two scales means that items which assume comparable positions in both scales are considered most harmonic. For alignment of the scale above with the definiteness hierarchy this means that pronominal subjects (+prominent/+prominent), as well as indefinite objects (-prominent/-prominent) are maximally harmonic, while the combination of a prominent position in one scale with a non-prominent position in the other scale is disharmonic. More precisely, harmonically aligning the hierarchy of syntactic roles with the definiteness hierarchy leads to two scales of feature combinations, one confined to subjects, and the other to objects. The subject scale is isomorphic to the definiteness hierarchy, while the ordering for objects is reversed.

(3) a. Subj/pronoun $\succ$ Subj/name $\succ$ Subj/def $\succ$ Subj/indef
b. Obj/indef $\succ$ Obj/def $\succ$ Obj/name $\succ$ Obj/pronoun

In this way DCM can be represented as a uniform phenomenon—case marking is always restricted to upper segments of these scales. This pattern becomes even more obvious if optional case marking is taken into account. As Aissen (2000) points out, if case marking is optional for some feature combination, it is optional or obligatory for every feature combination that is lower in the same hierarchy, and it is optional or prohibited for every point higher in the same hierarchy. Furthermore, if one looks at actual frequencies of case marking patterns in corpora, all available evidence suggests that the relative frequency of case marking always increases the farther down one gets in the hierarchy. What is interesting from a typological perspective is that there are very few attested cases of "inverse DCM"—languages that would restrict case marking to lower segments of the above scales. The restriction to upper segments appears to be a strong universal tendency.

Prince and Smolensky (1993) develop a simple method to translate harmony scales into OT constraints: for each element x of a scale we have a constraint *x ("Avoid x!"), and the ranking of these constraints is just the reversal of the harmony scale. The constraint sub-hierarchies obtained in this way are assumed to be universal, i.e. language specific total ranking respects them.

Generally, the common pattern of DCM is that non-harmonic combinations must be morphologically marked while harmonic combinations are unmarked. To formalize this idea in OT, Aissen employs the formal operation of *constraint conjunction*. If C_1 and C_2 are constraints, $C_1 \& C_2$ is another constraint which is violated iff both C_1 and C_2 are violated. Furthermore, two general constraints play a role: "*∅" is violated if a morphological feature is not marked, and "*STRUC" is violated by any morphological marking. Each constraint resulting from harmonic alignment is conjoined with *∅, and the ranking of the conjoined constraints is isomorphic to the ranking induced by alignment. The alignment of the animacy hierarchy with the scale of grammatical functions thus for instance leads to the following universal sub-hierarchies:

(4) *∅ & *Subj/inanim $\gg$ *∅ & *Subj/anim
*∅ & *Obj/anim $\gg$ *∅ & *Obj/inanim

Interpolating the constraint *STRUC at any point in any linearization of these sub-hierarchies leads to a pattern where morphological marking indicates non-harmony. The choice of the threshold for morphological marking depends on the relative position of *STRUC.

3 Statistical bias

In Zeevat and Jäger (2002) (ZJ henceforth) we attempt to come up with a functional explanation for the DCM pattern that are analyzed by Aissen. The basis for this approach is the observation that harmonic combinations of substantive and formal features (like the combinations “subject+animate” or “object+inanimate”) are common in actual language use, while disharmonic combinations (like “subject+inanimate” or “object+animate”) are rather rare. This intuition has been confirmed by several corpus studies. Table 1 displays the relative frequencies of feature combinations in the corpus SAMTAL, a collection of everyday conversations in Swedish that was annotated by Oesten Dahl. (Only subjects and direct objects of transitive clauses are considered,)

There are statistically significant correlations between grammatical function and each of the substantive features definiteness, pronominalization and animacy. The correlations all go in the same direction: harmonic combinations are over-represented, while disharmonic combinations are under-represented. If attention is restricted to simple transitive clauses, the chance that an arbitrarily picked NP is a subject is (of course) exactly 50%—exactly as high as the chance that it is a direct object. However, if an NP is picked at random and it turns out to be definite, the likelihood that it is a subject increases to 62.9%. On the other hand, if it turns out to be indefinite, the probability that it is a subject is as low as 3.9%. Analogous patterns obtain for all combinations. I henceforth assume the working hypothesis that these statistical biases are universal features of language use.

	NP	+def	-def	+anim	-anim
Subj	3151	3098	53	2948	203
Obj	3151	1830	1321	317	2834

Table 1: Frequencies in the SAMTAL corpus of spoken Swedish

4 Bidirectional Stochastic Optimality Theory

Aissen (2000) and Aissen and Bresnan (2002) point out that there is not just a universal tendency towards DCM across languages, but that DCM can also be used to describe statistical tendencies within one language that has, in the traditional terminology, optional case marking. Structural DCM can actually be seen as the extreme borderline case where these probabilities are either 100% or 0%. Stochastic Optimality Theory (StOT henceforth) in the sense of Boersma (1998) is a theoretical framework that is well-suited to formalize this kind of intuition. As a stochastic grammar, a StOT-Grammar does not just distinguish between grammatical and ungrammatical signs, but it defines a probability distribution over some domain of potential signs (in the context of OT: **GEN**).

StOT deviates from standard OT in two ways:

- **Constraint ranking on a continuous scale:** Every constraint is assigned a real number rather than a position in an ordinal hierarchy.
- **Stochastic evaluation:** At each evaluation, the placement of a constraint is modified by adding a normally distributed noise value. The ordering of the constraint after adding this noise value determines the actual evaluation of the candidate set at hand.

So we have to distinguish between the value that the grammar assigns to a constraint, and its actual ranking during the evaluation of a particular candidate.

An OT system consists of several constraints, and the addition of a noise value is done for each constraint separately. After adding the noise values, the actual values of the constraints define a total ranking. This total ordering of constraints is then used to evaluate candidates in the standard OT fashion, i.e. the strongest constraint is used first as a decision criterion, if there is a draw resort is taken to the second highest constraint and so on.

The probability for C1 $>$ C2 depends on the difference between their mean values that are assigned by the grammar. Let us denote the mean values of C1 and C2 as c1 and c2 respectively. Then the probability that C1 outranks C2 is a monotonic function of the difference between their mean values, c1$-$c2. If c1 $=$ c2, both have the same chance to outrank the other. This corresponds to a scenario where there is free variation between the candidates favored by C1 and those favored by C2. If C1 is higher ranked than C2, there is a preference for the C1-candidates. If the difference is larger than 12 units, the probability that C2 outranks C1 is less than 10^{-5}, which means that it is impossible for all practical purposes. In such a grammar C1 always outranks C1, and candidates that fulfill C2 at the expense of violating C1 can be regarded simply as ungrammatical (provided there are alternative candidates fulfilling C1, that is). So the classical pattern of a categorical ranking is the borderline case of the stochastic evaluation. It obtains if the distances between the constraints are sufficiently large.

Paul Boersma's Gradual Learning Algorithm (GLA) is an algorithm for learning a Stochastic OT grammar. It maps a set of utterance tokens to a grammar that describes the language from which this corpus is drawn. As a stochastic grammar, the acquired grammar makes not just predictions about grammaticality and ungrammaticality, but it assign probability distributions over each non-empty set of potential utterances. If learning is successful, they converge towards the relative frequencies of utterance types in the training corpus.

GLA operates on a predefined generator relation GEN that determines what qualifies as possible inputs and outputs, and which input-output pairs are admitted by the grammatical architecture. Furthermore it is assumed that a set CON of constraints is given, i.e. a set of functions which each assign a natural number (the "number of violations") to each element of GEN.

At each stage of the learning process, GLA assumes a certain constraint ranking. As an elementary learning step, GLA is confronted with an element of the training corpus, i.e. an input-output pair. The current grammar of the algorithm defines a probability distribution over possible outputs for the observed input, and the algorithm draws its own output for this input at random according to this distribution. If the result of this sampling does not coincide with the observation, the current grammar of the algorithm is slightly modified such that the observation becomes more likely and the hypothesis of the algorithm becomes less likely. This procedure is repeated for each item from the training corpus.

Optimality Theoretic evaluation is speaker oriented. It operates on a set of possible realizations of a hidden state (in the present context: possible forms corresponding to a given meaning). However, the statistical bias discussed above is only operative in the hearer direction. A hearer that is confronted with two structurally ambiguous NPs in construal with a transitive verb has to decide which one is subject and which is object. For animate NPs for instance the odds are that it is a subject while inanimates are most likely objects. To match these biases in a StOT grammar, Optimality Theory—and OT learning—has to operate in hearer direction as well.

The **Bidirectional Gradual Learning Algorithm** ("BiGLA" henceforth) differs from the original version in two respects. First, during the generation step the algorithm generates an optimal output for the observed input on the basis of a certain constraint ranking. It is tacitly assumed that "optimal" here means "incurring the least severe pattern of constraint violations" in standard OT fashion. In BiGLA it is instead assumed that the optimal output is selected from the set of outputs from which the input is *recoverable*. The input is recoverable from the output if among all inputs that lead to this output, the input in question incurs the least severe constraint violation profile (i.e. we apply interpretive optimization). If there are several outputs from which the input is recoverable, the optimal one is selected. If recoverability is impossible, the unidirectionally optimal output is selected.

This modification can be called **bidirectional evaluation**. Besides BiGLA involves **bidirectional learning**. This means that BiGLA both generates the optimal output for the observed input, and the optimal input for the observed output. "Comparison" and "adjustment" apply both to inputs and outputs as well. The pseudo-code for BiGLA is:

```
Initial state All constraint values are set to the initial value.

for ($i := 0; i < NumberOfObservations; i := i + 1$) {
        Observation A training datum is drawn at random from the training corpus, i.e. a fully specified
        input-output pair $\langle i, o \rangle$.

        Generation
            ○ For each constraint, a noise value is drawn from a normal distribution $N$ and added to its
              current ranking. This yields the selection point.
            ○ Constraints are linearly ranked by descending order of the selection points.
            ○ Based on this ranking, the grammar generates an optimal output $o'$ for the input $i$ and an
              optimal input $i'$ for the output $o$ using bidirectional evaluation.

        Comparison If $i = i'$ and $o = o'$, nothing happens. Otherwise, the algorithm compares the
        constraint violations of the learning datum $\langle i, o \rangle$ with the self-generated pairs $\langle i, o' \rangle$ and $\langle i', o \rangle$.

        Adjustment
            ○ All constraints that favor $\langle i, o \rangle$ over $\langle i, o' \rangle$ or $\langle i', o \rangle$ are promoted by some small predefined
              numerical amount ("plasticity").
            ○ All constraints that favor $\langle i, o' \rangle$ or $\langle i', o \rangle$ over $\langle i, o \rangle$ are demoted by the plasticity value.
}
```

5 BiGLA and DCM

Suppose the BiGLA is confronted with a language that has the same frequency distribution of the possible combinations of subject vs. object with animate vs. inanimate as the spoken Swedish from the SAMTAL corpus and uses case marking in exactly 50% of all cases, but in a way that is totally uncorrelated to animacy. We only consider simple transitive clauses, and we assume that this toy language has no other means for disambiguation besides case marking. So a learning datum will always be a combination of two NPs with a transitive verb. (I also assume that there are no verb specific preferences for certain readings of morphological markings.) Let us call the first NP "NP1" and the second one "NP2".

To see how BiGLA reacts to this language, we have to specify **GEN** and a set of constraints. Strictly speaking, animacy plays a double function in this experiment: it is of course an aspect of the meaning of an NP, but I also assume that this specification for +anim or −anim can be read off directly from the form of an NP. So +anim and −anim are treated as formal features, and **GEN** only relates animate meanings to +anim forms and inanimate meanings to −anim forms. There are thus eight possible semantic clause types to be distinguished because NP1 can be subject and NP2 object or vice versa, and both subject and object can be either animate or inanimate.

Let us assume that **GEN** supplies just one case morpheme, which is optional. The linking of this morpheme to a grammatical function is governed by the constraints, so **GEN** imposes no restrictions in this respect. **GEN** thus admits four types of morphological marking within a clause: both NP1 and NP2 can be case marked or unmarked. If $+/-$anim is taken into account, we get 16 different forms in total. However, **GEN** is organized in such a way that the animacy specification of the forms is completely determined by the meaning. So we end up with altogether 32 meaning-form combinations that are consistent with this **GEN**.

As mentioned above, we extract the frequencies of the possible meanings from the SAMTAL corpus. The absolute numbers are given in table 2.
Not surprisingly, the combination where both subject and object are harmonic is by far the most frequent pattern, and the combination of two disharmonic NPs is very rare.

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

Table 2: Frequencies of clause types

Table 3 gives a frequency distribution (in per cent of all clauses in the corpus) over this **GEN** which respects the relative frequencies of the different meanings from SAMTAL and treats the linking of NP1 or NP2 to the subject role as equally likely. The notation “case1-case2” indicates that NP1 is marked with case1 and NP2 with case2 (M and Z abbreviate “marked” and “zero” respectively). Likewise, the notation “su/a-ob/i” means that NP1 is interpreted as animate subject and NP2 as inanimate object etc.

As for the constraint inventory, I basically assume the system from Aissen (2000) (restricted to the animate/inanimate contrast). This means we have four marking constraints. Using the same notation as in the table above, we can write them as *(su/a/z), *(su/i/z), *(ob/a/z), and *(ob/i/z). They all enforce case marking. They are counteracted by *STRUC which is violated by a clause as often as there are case morphemes present in a clause. (The evaluation of the constraints is done per clause, not just per NP.) The case morpheme may be interpreted either as ergative or as accusative case, and both interpretations are favored by one constraint each, m⇒su for ergative and m⇒ob for accusative. Finally I assume that the grammar does distinguish between interpreting NP1 or NP2 as a subject.

	M-M	M-Z	Z-M	Z-Z
su/a-ob/a	1.19	1.19	1.19	1.19
su/a-ob/i	10.50	10.50	10.50	10.50
su/i-ob/a	0.07	0.07	0.07	0.07
su/i-ob/i	0.74	0.74	0.74	0.74
ob/a-su/a	1.19	1.19	1.19	1.19
ob/a-su/i	0.07	0.07	0.07	0.07
ob/i-su/a	10.50	10.50	10.50	10.50
ob/i-su/i	0.74	0.74	0.74	0.74

Table 3: Training corpus

In real languages there are many constraints involved here (pertaining to syntax, prosody and information structure). In the context of our experiment, I skip over these details by assuming just two more constraints, SO and OS. They are violated if NP2 is subject and if NP1 is subject respectively. Since all constraints start off with the initial value 0, there is no *a priori* preference for a certain linking—these two constraints simply equip UG with means to distinguish between the two possible linking patterns. Altogether we thus get the nine constraints in table 4.

*(su/a/z):	*Avoid unmarked animate subjects!*
*(su/i/z):	*Avoid unmarked inanimate subjects!*
*(ob/a/z):	*Avoid unmarked animate objects!*
*(ob/i/z):	*Avoid unmarked inanimate objects!*
m⇒su:	*Marked NPs are subjects.*
m⇒ob:	*Marked NPs are objects.*
*STRUC:	*Avoid case marking!*
SO:	*NP1 is subject and NP2 object.*
OS:	*NP2 is subject and NP1 object.*

Table 4: Constraint inventory

6 Iterated learning and DCM

Depending on the OT system that is used, the training corpus and the chosen parameters, the stochastic language that is defined by the acquired grammar may deviate to a greater or lesser degree from the training language. Especially for BiGLA this deviation can be considerable. (It is perhaps misplaced to call BiGLA a "learning" algorithm; it rather describes a certain adaptation mechanism.) If a sample corpus is drawn from this language and used for another run of BiGLA, the grammar that is acquired this time may differ from the previously learned language as well.

Such a repeated cycle of grammar acquisition and language production has been dubbed the *Iterated Learning Model* of language evolution by Kirby and Hurford (2002).

The production half-cycle involves the usage of a random generator to produce a sample corpus from a stochastic grammar. In the simulations, I assumed that this sample corpus has the same absolute size than the initial corpus. Furthermore I assume that the absolute frequencies of the different *inputs* are kept constant in each cycle. What may change from cycle ("generation") to cycle are the relative frequencies of the different outputs for each input.

A given stochastic grammar G defines a probability distribution $p_G(\cdot|i)$ over the possible outputs o for each input i. Using a random generator, this probability distribution can be used to generate a new corpus that represents the acquired grammar. One cycle of learning and production represents one generation in the evolutionary process that is simulated. This cycle my be repeated arbitrarily many times, i.e. over an arbitrary number of generations.[1]

In the first simulation to be reported here I used the constraint inventory, generator relation and corpus frequencies given above as initial input for iterated learning. The successive constraint rankings that emerge in this way are plotted in figure 1. The learning procedure was repeated 500 times, and the generations are mapped to the x-axis, while the y-axis again gives the constraint rankings.

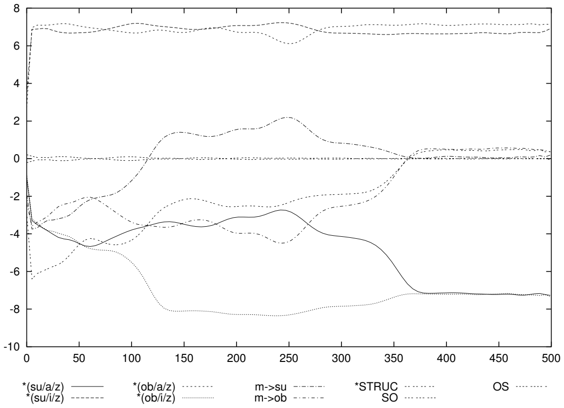

Figure 1: Diachronic development

While there are no rough changes from one generations to the next, the grammar as a whole gradually changes its characteristics over time. Aissen' sub-hierarchies—*(su/i/z) $\gg$ *(su/a/z) and *(ob/a/z) $\gg$ *(ob/i/z)—are invariant though.

We may distinguish four phases. During the first phase (generations 1–10), the constraints *(su/i/z) and *(ob/a/z) stay closely together, and they increase their distance from *STRUC. This amounts to an ever stronger tendency for case marking of disharmonic NPs. Simultaneously, *(su/a/z) and *(ob/i/z) stay close to *STRUC, i.e. we have optional case marking of harmonic NPs. This corresponds to a split ergative system with optional marking of harmonic and obligatory marking of disharmonic NPs. This characteristics remains relatively stable during the second phase (roughly generations 11–100). Then the system becomes unstable. After another thirty generations, it enters a relatively stable state where case marking of inanimate objects is completely lost while case marking of animate subjects is still optional. Case marking of disharmonic NPs remains obligatory. This situation remains constant (with some minor variation) for about 200 generations, when case marking of animate subjects is lost as well. The constraint ranking now reached is

$$\{\text{*(su/i/z), *(ob/a/z)}\} \gg \{\text{m}\Rightarrow\text{su, m}\Rightarrow\text{ob, *STRUC, SO, OS}\} \gg \{\text{*(su/a/z), *(ob/i/z)}\}$$

Needless to say that the diachronic development that is predicted by the BiGLA (together with **GEN**,

[1] The software package *evolOT* implements this version of the Iterated Learning Model. It is freely available from http://www.ling.uni-potsdam.de/˜jaeger/evolOT.

the constraint set, and the probability distribution over meanings from SAMTAL) depends on the pattern of case marking that was used in the first training corpus. A full understanding of the dynamics of this system and the influence of the initial conditions requires extensive further research. In the remainder of this section I will report the results of some experiments that give an idea of the overall tendencies though.

If the first training corpus contains no case marking at all (a somewhat unrealistic scenario, given that the **GEN** supplies case morphemes—perhaps this models the development of a language immediately after some other device has been reanalyzed as case morpheme), the overall development is similar to the previous setup. The ranking that BiGLA induces from the initial corpus places *STRUC extremely high (at 55.79), while the constraints that favor case marking are placed much lower, thus reflecting the absence of case marking. Still, the Aissen sub-hierarchies are respected, with *(su/a/z) at -33.04, *(su/i/z) at 5.03, *(ob/a/z) at 1.04 and *(ob/i/z) at -29.03. However, case marking of disharmonic NPs is gradually acquired within a few generations, and after thirty generations the system already enters the steady state of split ergativity (see figure 2).

It was mentioned in the beginning that DCM is a strong universal tendency. There are very few languages with an inverse DCM pattern. This is predicted by the assumption of Aissen's universal sub-hierarchies: there cannot be a language that marks animate subjects with higher probability than inanimate ones, say. It is revealing to run the BiGLA on a training corpus with such an (allegedly impossible) pattern. I did a simulation with a training corpus where all and only the harmonic NPs were case marked. The development of the constraint ranking is given in figure 3.

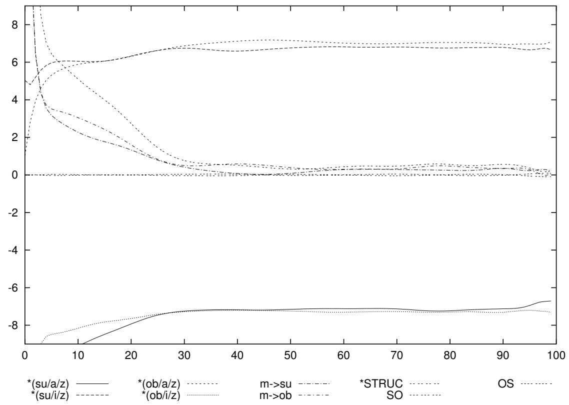

Figure 2: No case marking in the initial state

The BiGLA in fact learns the inverse pattern, i.e. it comes up with a grammar where the Aissen sub-hierarchies are reversed: *(su/a/z) $\gg$ *(su/i/z) and *(ob/i/z) $\gg$ *(ob/a/z). Accordingly, the language that is learned in the first generation marks almost all harmonic NP but nearly no disharmonic ones. So UG admits such a language, and it is also learnable. However, it is extremely unstable. Already after twelve generations the Aissen sub-hierarchies emerge and remain stable for the remainder of the simulation. Nonetheless, the case marking patterns changed dramatically after that. For about 100 generations after the emergence of the Aissen hierarchies, case marking is obligatory for disharmonic and optional for harmonic NPs. After that, the system dramatically changes its character and enters a state of a pure ergative system, i.e. all subjects and no objects are case marked. Around generation 1000 (not included in the graphics) the system switched into a split ergative state, as in the first two experiments.

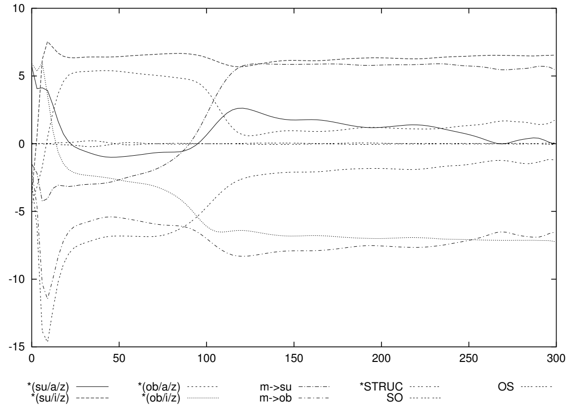

Figure 3: The future of anti-DCM

While these simulations establish a connection between the statistical patterns of language use and the independently motivated constraint hierarchies postulated by Aissen, the experimental results are at odds with the actual typological tendencies. Languages with split ergativity are a minority among the languages of the world. The majority of languages follows a nominative-accusative pattern, often combined with DOM. It is a matter of dispute whether pure (morphological) ergative languages exist at all, and in any case they are very rare. How do these facts relate to the predictions of iterated learning? I will conclude this section with some speculations about the typology of case marking patterns within the paradigm of iterated learning using BiGLA.

The dynamics of the system is very sensitive to the relative frequencies of the different meanings. The emergence of Aissen's sub-hierarchies is due to the fact that there are much more clauses of the type "animate subject – inanimate object" than the inverse type. The clauses where both arguments are of the same animacy are irrelevant here. Their relative frequency is decisive for the precise nature of the steady states though. In the SAMTAL corpus, the number of clauses were both arguments are animate (300) has the same order of magnitude as the number of clauses with two inanimate arguments (186). If we look at definiteness instead, this is different. Here the frequencies are as in table 5. There are about sixty times as many clauses with two definite arguments as clauses with two indefinite NPs. Feeding a training corpus with these relative frequencies and 50% probability of case marking for each NP type into iterated BiGLA gives a qualitatively different trajectory than in the previous experiments. It is given in figure 4.

	subj/def	subj/indef
obj/def	1806	24
obj/indef	1292	29

Table 5: Frequencies of clause types with respect to definiteness

Here the system reaches a steady state after about 70 generations. The emerging ranking is the following (where "*(ob/d/z)" stands for "Avoid unmarked definite objects!" etc.):

$$\{*(\text{obj/d/z}), \text{m}\Rightarrow\text{obj}\} \gg *(\text{obj/i/z}) \gg \{*(\text{subj/i/z}), \text{SO}, \text{OS}\} \gg *\text{STRUC} \gg *(\text{su/d/z}) \gg \text{m}\Rightarrow\text{su}$$

This grammar seems to describe a language with obligatory object marking and DSM. However, recall that **GEN** only supplies one case morpheme here, and the sub-hierarchy $\text{m}\Rightarrow\text{obj} \gg \text{m}\Rightarrow\text{su}$ ensures that this morpheme is unequivocally interpreted as accusative. Thus ergative marking is impossible and the constraint ranking describes a language with obligatory object marking and no subject marking.

To sum up the findings from this section, we may distinguish several types of case marking patterns according to their likelihood. Most unlikely are languages that violate UG, i.e. where there is no constraint ranking that describes such a language. If we assume a UG as above (i.e the **GEN** and set of constraints discussed in the previous section), there can't be a language where either both subject and object or neither are case marked. (Feeding such a corpus into BiGLA leads to a language where about 60% of all clauses contain exactly one case marker.) Note that it is extremely unlikely but not impossible to find a corpus with this characteristics, because this language is a subset of many UG-compatible languages. Such a corpus would be highly un-representative though.

The next group consists of languages that correspond to some constraint ranking but are not learnable in the sense that exposing the BiGLA to a sample from such a language leads to a grammar of a substantially different language. The language without any case marking would fall into this category (provided **GEN** supplies case marking devices). There is a ranking which describes such a language, namely

$$*\text{STRUC} \gg \{\text{OS}, \text{SO}\} \gg \{*(\text{su/a/z}), *(\text{su/i/z}), *(\text{ob/a/z}), *(\text{ob/i/z}), \text{m}\Rightarrow\text{su}, \text{m}\Rightarrow\text{ob}\}$$

However, if the BiGLA is exposed to a sample from this language, it comes up with a substantially different ranking, namely

$$*\text{STRUC} \gg \{\text{m}\Rightarrow\text{su}, \text{m}\Rightarrow\text{ob}\} \gg *(\text{su/i/z}) \gg *(\text{ob/a/z}) \gg \{\text{OS}, \text{SO}\} \gg *(\text{su/a/z}) \gg *(\text{ob/i/z})$$

11.1% of the NPs in a sample corpus drawn from this language do carry case marking.

The third group consists of languages that are both in accordance with UG and learnable, but diachronically instable. This means that the BiGLA acquires a language that is similar but not entirely identical to the training language, and that the deviation between training language and acquired language always goes into the same direction. Diachronically this leads to a change of language type after some generations. This can be observed most dramatically with languages with inverse DCM (compare figure 3). There the language type switches from inverse split ergativity to optional split ergativity within less than twenty generations.

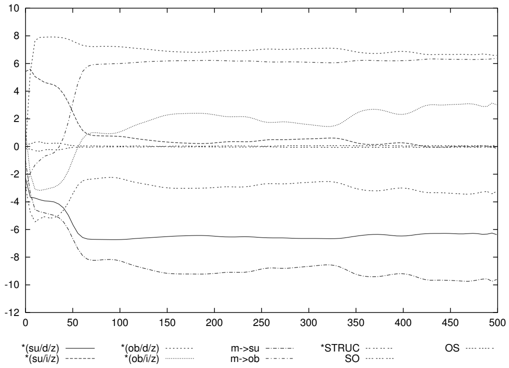

Figure 4: Stabilization at accusative system

The most likely language types are those that are diachronically stable and are additionally the target of diachronic change in many cases. The experiments conducted so far indicate that there is exactly one such steady state for each experimental setup—split ergativity in the first and nominative-accusative in the second scenario.

Given the extremely coarse modeling of the factors that determine case marking in our experiments and the fact that the experiments all depend on a probability distribution over meanings that is based on just one corpus study, these results have to be interpreted with extreme caution. They fit the actual patterns of typological variation fairly well though, so it seems worthwhile to pursue this line of investigation further.

References

Aissen, J. (2000). Differential object marking: Iconicity vs. markedness. Manuscript, UCSC.

Aissen, J. and J. Bresnan (2002). OT syntax and typology. course material from the Summer School on Formal and Functional Linguistics. University of Düsseldorf.

Boersma, P. (1998). *Functional Phonology*. Ph.D. thesis, University of Amsterdam.

Bossong, G. (1985). *Differentielle Objektmarkierung in den neuiranischen Sprachen*. Günther Narr Verlag, Tübingen.

Jäger, G. (2002). Learning constraint sub-hierarchies. The Bidirectional Gradual Learning Algorithm. manuscript, University of Potsdam.

Kirby, S. and J. R. Hurford (2002). The emergence of linguistic structure: An overview of the Iterated Learning Model. In A. Cangelosi and D. Parisi, eds., *Simulating the Evolution of Language*, pp. 121–147. Springer, London.

Prince, A. and P. Smolensky (1993). Optimality theory: Constraint interaction in generative grammar. Technical Report TR-2, Rutgers University Cognitive Science Center, New Brunswick, NJ.

Zeevat, H. and G. Jäger (2002). A reinterpretation of syntactic alignment. In D. de Jongh, M. Nielsenova, and H. Zeevat, eds., *Proceedings of the Fourth International Tbilisi Symposium on Language, Logic and Computation*. University of Amsterdam.