

Learning constraint sub-hierarchies. The Bidirectional Gradual Learning Algorithm*

Gerhard Jäger

University of Potsdam & ZAS Berlin

jaeger@ling.uni-potsdam.de

May 2003

1 Differential Case Marking

It is a common feature of many case marking languages that some but not all objects are case marked.¹ 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. Take Hebrew as an example. In this language, definite objects carry an accusative morpheme while indefinite objects are unmarked.

- (1) a. Ha-seret her?a *?et-ha-milxama*
THE-MOVIE SHOWED ACC-THE-WAR
b. Ha-seret her?a (**?et-*)*milxama*
THE-MOVIE SHOWED (*ACC-)WAR
(from Aissen 2000)

Similar patterns are found in many languages. Bossong (1985) calls this phenomenon “Differential Object Marking” (DOM). A common pattern is that all NPs from the top section of the *definiteness hierarchy* are case marked while those from the bottom section are not.

*This is the revised version of a paper I posted at ROA in September 2002. The original version contained some flaws that are corrected now. Especially section 10 has been completely rewritten.

¹ Here and throughout the paper, I consider the morphological form of the subject of an intransitive clause as unmarked, and case marking that deviates from it as marked.

- (2) personal pronoun > proper noun > definite full NP > indefinite specific NP > non-specific indefinite NP

Catalan, for instance, only marks personal pronouns as objects. In Pitjantjatjara (an Australian language), pronouns and proper nouns are case marked when they are objects while other NPs aren't. Hebrew draws the line between definite and indefinite NPs and Turkish between specific and non-specific ones.²

Likewise, the criterion for using or omitting a case morpheme for objects may come from the animacy hierarchy.

- (3) human > animate > inanimate

As with the definiteness hierarchy, there are languages which only mark objects from some upper segment of this scale. Finally, there are instances of DOM where case marking is restricted to an upper segment of the product of the two scales.³

Differential case marking also frequently occurs with subjects.⁴ In contradistinction to DOM, DSM ("Differential Subject Marking") means that only instances of some *lower* segment of the definiteness/animacy hierarchy are case marked. (The observation that the relevant scales for subjects and objects are inverses of each other is due to Silverstein 1976.)

DOM and DSM may co-occur within one language. This phenomenon is usually called *split ergativity*. (This term covers both case marking systems where the case marking segments for subjects and for objects are complementary and systems where they overlap.)

The person specification of NPs induces another hierarchy. Simplifying somewhat, it says that the local persons (1st and 2nd) outrank 3rd person.

- (4) 1st/2nd person > 3rd person

These patterns underlie split ergative case marking in languages like Dyirbal where the choice between the nominative/accusative system and the ergative/absolutive system is based on person. Table 1 (which is taken from Aissen 1999) shows the basic case marking pattern for Dyirbal.

² See Aissen (2000) for a more elaborate discussion, examples and references.

³ By this I mean the partial order over the Cartesian product of the domain of the two scales, where $\langle a_1, b_1 \rangle \geq \langle a_2, b_2 \rangle$ iff $a_1 \geq a_2$ and $b_1 \geq b_2$.

⁴ Here and henceforth, I use the term "subject" to refer both to the single argument of an intransitive verb and to the controller/agent argument of transitive verb. "Object" refers to the non-subject argument of a simple transitive verb. While this terminology expresses a bias towards accusative systems and against ergative systems, no real harm is done by this in the context of this paper because it does not deal with intransitive clauses.

	Unmarked	Marked
Local persons	Subject	Object
3rd person	Object	Subject (of transitive)
Case	Nominative/Absolutive	Accusative/Ergative

Tab. 1: Case marking system of Dyirbal

Briefly put, Dyirbal only marks non-harmonic arguments, i.e. local objects and 3rd person subjects. It thus represents a combination of DOM with DSM.

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, egocentricity, or animacy (as proposed in Silverstein 1976). Ranking the grammatical functions according to prominence leads to the binary scale

(5) 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 non-specific 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 (like non-specific subjects or pronominal objects). 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.

(6) a. Subj/pronoun > Subj/name > Subj/def > Subj/spec > Subj/non-spec
b. Obj/non-spec > Obj/spec > Obj/def > Obj/name > 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 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 (see Aissen & Bresnan 2002).

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.⁵ The restriction to upper segments appears to be a strong universal tendency.

2 OT Formalization

Prince & 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. For the person/grammatical function interaction discussed above, this looks schematically as follows (adapted from Bresnan et al. 2001):

(7)	Prominence scales	Harmonically aligned scales	OT constraint sub-hierarchies
	Subj > Obj	Subj/local > Subj/3rd	$*\text{Subj}/3\text{rd} \gg * \text{Subj}/\text{local}$
	local > 3rd	Obj/3rd > Obj/local	$*\text{Obj}/\text{local} \gg * \text{Obj}/3\text{rd}$

To translate harmony scales into OT, first every feature combination f is compiled into a constraint saying “Avoid f !” For instance, the combination “Subj/local” corresponds to the constraint “*Subj/local”, that is violated by every local person subject. The ordering in the harmony scale is translated into universal sub-hierarchies which are to be respected by any language particular total constraint ranking. If, according to the harmony scale, local person subjects are better than third person subjects, then being a third person subject is (universally) worse than being a local person subject. This is expressed by the constraint sub-hierarchy “*Subj/3rd \gg *Subj/local”.

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* from Smolensky (1995). If C_1 and C_2 are constraints,

⁵ Dixon (1994), p. 90 gives two examples: the Australian language Arrernte has an inverse split ergativity system for pronouns—only first person pronouns are marked as subjects, while all other pronouns are unmarked as subjects but marked as objects. Nganasan (from the Samoyedic group of the Uralic family) has inverse DOM, i.e. full nouns but not pronouns are case marked as objects.

$C_1 \& C_2$ is another constraint which is violated iff both C_1 and C_2 are violated. Crucially, $C_1 \& C_2$ may outrank other constraints C_i that in turn outrank both C_1 and C_2 . So the following constraint ranking is possible:

$$C_1 \& C_2 \gg C_3 \gg C_4 \gg C_1 \gg C_5 \gg C_2$$

Furthermore, two general constraints play a role:

- “* \emptyset ” is violated if a morphological feature is not marked
- “*STRUC” is violated by any morphological marking

Each constraint resulting from harmonic alignment is conjoined with * \emptyset , and the ranking of the conjoined constraints is isomorphic to the ranking induced by alignment. (Also the conjoined constraints outrank each of their conjuncts.) The alignment of the person hierarchy with the scale of grammatical functions thus for instance leads to the following universal constraint sub-hierarchies:

$$(8) \quad \begin{array}{l} * \emptyset \ \& \ * \text{Subj}/3\text{rd} \ \gg \ * \emptyset \ \& \ * \text{Subj}/\text{local} \\ * \emptyset \ \& \ * \text{Obj}/\text{local} \ \gg \ * \emptyset \ \& \ * \text{Obj}/3\text{rd} \end{array}$$

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. The Dyirbal pattern, for instance, would follow from the following constraint ranking.

$$(9) \quad * \emptyset \ \& \ * \text{Subj}/3\text{rd} \ \gg \ * \emptyset \ \& \ * \text{Obj}/\text{local} \ \gg \ * \text{STRUC} \ \gg \ * \emptyset \ \& \ * \text{Subj}/\text{local} \ \gg \ * \emptyset \ \& \ * \text{Obj}/3\text{rd}$$

3 Statistical bias

In Zeevat & 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 2 displays the relative frequencies of feature combinations in the corpus SAMTAL, a collection of everyday conversations in Swedish that was

	NP	+def	-def	+pron	-pron	+anim	-anim
Subj	3151	3098	53	2984	167	2948	203
Obj	3151	1830	1321	1512	1639	317	2834
χ^2		1496		1681		4399	
$p < 0.01\%$		yes		yes		yes	

Tab. 2: Frequencies in the SAMTAL corpus of spoken Swedish

$p(\text{subj} +\text{def}) = 62.9\%$	$p(\text{subj} -\text{def}) = 3.9\%$
$p(\text{obj} +\text{def}) = 37.1\%$	$p(\text{obj} -\text{def}) = 96.1\%$
$p(\text{subj} +\text{pron}) = 66.4\%$	$p(\text{subj} -\text{pron}) = 9.2\%$
$p(\text{obj} +\text{pron}) = 33.6\%$	$p(\text{obj} -\text{pron}) = 90.8\%$
$p(\text{subj} +\text{anim}) = 90.3\%$	$p(\text{subj} -\text{anim}) = 6.7\%$
$p(\text{obj} +\text{anim}) = 9.7\%$	$p(\text{obj} -\text{anim}) = 93.3\%$

Tab. 3: Conditional probabilities

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:

This statistical bias has little to do with the grammar of the language at hand. There is some minor influence because diathesis can be used to avoid disharmonic combinations (see Aissen 1999 and Bresnan et al. 2001 for discussion), but since the passive is generally rare and there is no categorical grammaticalized correlation between referentiality or animacy and diathesis

in Swedish, the general pattern is hardly affected by this factor.⁶ So despite the thin cross-linguistic evidence (though the same patterns have been found in the Wall Street Journal Corpus by Henk Zeevat, in the CallHome corpus of spoken Japanese by Fry 2001, and in the SUSANNE and CHRISTINE corpora of written and spoken English by myself) I henceforth assume the working hypothesis that these statistical biases are universal features of language use.

4 Bias and bidirectional optimization

Differential case marking amounts to a preference for case marking of disharmonic feature combinations over case marking of harmonic combinations. Taking the statistical patterns of language use into account, this means that there is a preference for case marking of rare combinations, while frequent forms are more likely to be unmarked. This is a sensible strategy because it minimizes the overall effort of the speaker while preserving the disambiguating effect of case marking.⁷ As pointed out in ZJ, Bidirectional Optimality Theory in the sense of Blutner (2001) provides a good theoretical framework to formalize this kind of pragmatic reasoning.

According to Bidirectional OT (which is founded in work on formal pragmatics), a meaning-form pair is only optimal if it conforms to the preferences of both speaker and hearer in an optimal way. Speaker preferences and hearer preferences of course need not coincide. However, they do not contradict each other either, for the simple reason that the speaker has preferences between different ways to express a given meaning, while the hearer compares different interpretations of a given form. Applied to case marking, it is plausible to assume that the speaker has *ceteris paribus* a preference to avoid case marking. The hearer, on the other hand, has a preference for faithful interpretation (accusative NPs are preferredly interpreted as objects and ergative NPs as subjects). Furthermore, ZJ claim that there is a hearer preference to follow the statistical bias, i.e. to interpret definite or animate NPs as subjects and indefinite or inanimate NPs as objects.

These preferences can easily be interpreted as OT constraints. The speaker preference against case marking is just Aissen's constraint *STRUC. Pref-

⁶ In other languages, the impact of the grammar on these quantities might be considerable. To clearly separate the usage patterns from grammatical features of the language studied, one has to look at the correlation between animacy/definiteness and *semantic* roles. This has to be left for future research.

⁷ The resemblance to optimal coding in the sense of information theory is striking. Shannon (1948) showed that an optimal coding must assign long codes to rare events and short codes to frequent ones.

erence for faithfulness interpretation of case morphemes can be covered by a constraint FAITH (arguably there are different faithfulness constraints for different morphemes, but for the purposes of ZJ as well as of the present paper one big faithfulness constraint will do). Finally, ZJ assume a constraint BIAS that is fulfilled if an NP of a certain morphological category is interpreted as having the grammatical function that is most probable for this category.⁸

For FAITH to take any effect, it must be (universally) ranked higher than BIAS. The relative ranking of *STRUC is actually inessential. For the sake of illustration, we assume it to be ranked lowest. So the hierarchy of constraints is

$$\text{FAITH} \gg \text{BIAS} \gg \text{*STRUC}$$

In contradistinction to standard OT, Bidirectional OT takes both hearer preferences and speaker preferences into account. Hearer optimality means: for a given form, choose the meaning that has the least severe constraint violation pattern. For the constraint system at hand, this means: interpret an NP according to its case marking, and in the absence of case marking, follow the statistical bias. The speaker has to take this hearer strategy into account to get his message across. Only if two competing forms are both hearer optimal for a given meaning, the speaker is free to choose the preferred one (which means in the present setup: the one without case marking).

This view on bidirectional optimization can be formalized in the following way.⁹ I write $\langle m_1, f_1 \rangle < \langle m_2, f_2 \rangle$ iff the meaning-form pair $\langle m_1, f_1 \rangle$ is better than $\langle m_2, f_2 \rangle$ according to the constraints given above in the given ranking. Following standard practice, I assume a generator relation **GEN** between forms and meanings from which the optimal candidates are chosen. **GEN** supplies the morphological inventory of a language as well as some general, highly underspecified structural relation between forms and meanings.

Definition 1:

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



⁸ The terminology I use here differs somewhat from ZJ but is more in line with the bulk of the OT literature.

⁹ The notion of bidirectionality given in the definition differs from Blutner's, which treats speaker and hearer totally symmetrical. Also, I am again deviating somewhat from the original formulation in ZJ in a way that makes no difference for their general point.

- A meaning-form pair $\langle m, f \rangle$ is *optimal* iff it is hearer-optimal and there is no alternative form f' such that $\langle m, f' \rangle$ is hearer-optimal and $\langle m, f' \rangle < \langle m, f \rangle$.

Now suppose the **GEN** relation for a given language supplies both an accusative and an ergative morpheme. How would, say, an inanimate object be morphologically realized in an optimal way? To keep things simple, let us assume that the interpretation of an NP within a clause is uniquely determined by its grammatical function. (In a more elaborate system, grammatical functions only mediate between surface realization and semantic roles, but I will not go into that in the context of the present paper.) We get the following tableau:

(10)

		FAITH	BIAS	*STRUC
anim+ \emptyset 	Subj			
	Obj		*	
anim+ERG	Subj			*
	Obj	*	*	*
anim+ACC 	Subj	*		*
	Obj		*	*

To figure out which meaning-form pairs are hearer optimal, we have to compare the different meanings (subject vs. object) of the three potential morphological realizations: zero (i.e. identical to the subject marking in intransitive clauses), ergative or accusative. It is easy to see that the association of both zero marking and ergative marking with the subject role, and the association of accusative marking with the object role are hearer optimal. Speaker optimization chooses between the possible hearer-optimal realizations of a given meaning. For the subject interpretation, there is a choice between zero marking and ergative marking. Since the latter violates *STRUC and the former doesn't, and they do not differ with respect to other constraints, zero marking is optimal for the subject interpretation. For the object interpretation, there is only one hearer optimal realization—accusative marking—which is thus trivially optimal.

For inanimate NPs, the pattern is reversed. Here subjects must be case marked with the ergative morpheme, while objects are preferredly unmarked.

(11)

		FAITH	BIAS	*STRUC
inanim+ \emptyset	Subj		*	
	Obj			
inanim+ERG	Subj		*	*
	Obj	*		*
inanim+ACC	Subj	*	*	*
	Obj			*

ZJ's system thus predicts a split ergative system: case marking is restricted to disharmonic feature combinations—animate objects and inanimate subjects—while harmonic combinations are unmarked.

This mechanism only works though if the NP at hand is in fact ambiguous between subject and object interpretation. If it is disambiguated by means of external factors like word order, agreement, semantic plausibility etc. zero marking will always win. Let us call such a case marking system *pragmatic DCM*. However, the languages that were mentioned in the beginning require case marking of disharmonic combinations regardless of the particular contextual setting. Restricting attention to (in)animacy, this would mean that *all* animate objects and inanimate subjects must be case marked, no matter whether case marking is necessary for disambiguation in a particular context. I call such a system *structural DCM* henceforth. Bidirectional OT does not give an immediate explanation for such a pattern.

ZJ suggest that structural DCM emerges out of pragmatic DCM as the result of a grammaticalization process. If a language starts employing pragmatic DCM, the next generation of language learners are faced with two ways of making sense of the case marking pattern: pragmatic DCM or optional structural DCM. If both hypotheses are entertained, the overall probability for DCM increases (i.e. the probability for an animate subject to be zero-marked, for an inanimate subject to be case marked etc.). This in turn makes the hypothesis of structural DCM more plausible. After some generations of partial reanalysis, DCM is fully grammaticalized, i.e. pragmatic DCM has turned into structural DCM.

There are quite a few problems that are left open by the ZJ-approach. To start with, the explanation of structural DCM rests on a fairly sketchy account of grammaticalization. Also it predicts that in languages that have both ergative and accusative morphology at their disposal, a split ergative system should emerge where the split points for DSM and DOM are identical (as in Dyirbal, see above). While this is the common pattern for split ergativity, there are also languages where the segments for subject marking and for object marking overlap. Dixon (1994), p. 86 mentions the example of

Cashinawa (a language from Peru), where all pronouns have case marking in object position, and all third person NPs are case marked in subject position. In other words, third person pronouns occur in three forms: unmarked (as subjects of intransitive clauses), ergative case and accusative case. According to the ZJ-system, these pronouns should have a bias either towards a subject or an object interpretation, and thus only the interpretation unsupported by this bias should be marked (or, at any rate, this should have been the situation in the pragmatic DCM language from which the structural pattern of Cashinawa emerged). Also, the ZJ-system fails to explain the great cross-linguistic diversity of DCM systems. If DCM is directly rooted in a statistical bias which in turn has extra-linguistic sources, one would expect to find not just the same pattern but also the same split across languages.

There is also a conceptual problem with ZJ's approach. The constraint BIAS makes direct reference to the statistics of language use. While it might be plausible that grammatical rules and constraints are induced from frequencies, it seems unlikely that the internalized grammar of a speaker contains a counter that keeps track of the relative frequencies of feature associations, say. In other words, frequencies may help to explain why and how a certain grammar has been learned, but they are not part of this grammar.

In the remainder of this paper I will outline a theory that remedies the last problem. While the explanation of pragmatic DCM in terms of bidirectional optimization is preserved, the connection between the statistics of language use and the competence grammar of the speakers of a language is established via a learning algorithm, rather than feeding the statistical information directly into the grammar. This approach solves two puzzles: It explains why the constraint sub-hierarchies that Aissen assumes to be universal are so common without taking resort to UG, and it gives a formal account of the diachronic shift from pragmatic to structural DCM. The cross-linguistic diversity of the possible split points for DCM is not further discussed in this paper, but it is likely that this problem can be dealt with in this framework as well.

5 Stochastic Optimality Theory

Aissen (2000) and Aissen & 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. In colloquial Japanese, for example, 70% of the inanimate subjects, but only 65% of the animate subjects are case marked. Conversely, 54% of the animate, but only 47%

of the inanimate objects are marked (these figures are taken from Aissen & Bresnan 2002, who attribute them to Fry 2001). 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**). Ungrammaticality is thus the borderline case where the grammar assigns the probability 0. StOT deviates from standard OT in two ways:

- **Constraint ranking on a continuous scale:** Every constraint is assigned a real number. This number does not only determine the ranking of the constraints, but it is also a measure for the distance between them.
- **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. To make this point clear, suppose we have some constraint C which, according to the grammar, has the value 0.5.¹⁰ To evaluate whether a particular linguistic item in a corpus violates this constraint, we have to determine C's actual value. It is obtained from its grammar value by adding some amount z of unpredictable noise. z may be any real number, so the actual value of C can be any number as well. However, z is likely to have a small absolute value, so the actual value of C is likely to be in the vicinity of 0.5. Boersma assumes that z is distributed according to a normal distribution with mean $\mu = 0$ and standard deviation $\sigma = 2$.¹¹ So the actual value of C is also normally distributed, with mean 0.5 and standard deviation 2. This distribution is depicted in figure 1.

¹⁰ Boersma (1998) and Boersma & Hayes (2001) prefer values around 100 while I find values around 0 easier to work with. Since only the distance between constraint values matters and not the values as such, this makes no real difference.

¹¹ In the graph of a normal distribution (see figure 1), the mean corresponds to the center where the value of the function is at its maximum, and the standard deviation is the distance between the mean and the points on both sides where the shape of the curve changes from concave to convex.

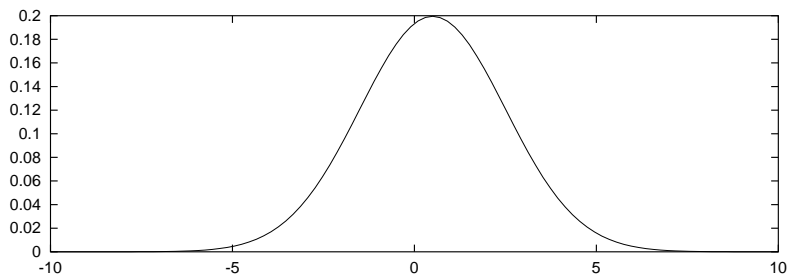


Fig. 1: Normal distribution

It has the Gaussian bell shape that is familiar from many stochastic phenomena. The probability that the value of C falls within a certain interval on the x -axis is proportional to the area between the x -axis and the bell curve over this interval. The entire area under the curve is 100%. So for instance the probability that the value of C is less than 0.5 is exactly 50%. While the curve never touches the x -axis in theory on either side, the probability that C is ranked below -9 or over 10 is so small (about 10^{-6}) that it can be neglected.

An OT system consists of several constraints, and the addition of a noise value is done for each constraint separately. Suppose the grammar assigns the constraints $C1$ and $C2$ the mean values -0.5 and 0.5 respectively. The corresponding function graph is depicted in figure 2.

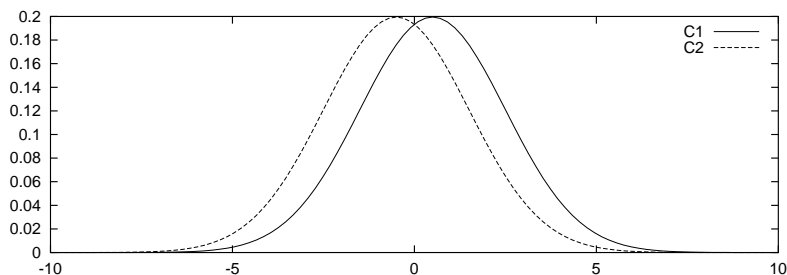


Fig. 2: Two constraints

Since the mean of $C1$ is higher than the mean of $C2$, most of the time $C1$ will end up having a higher value than $C2$. However, it is perfectly possible that $C2$ receives an unusually high and $C1$ an unusually low value, so that in the end $C2 > C1$.¹² The probability for this is about 36%. In any event,

¹² I use $C1$, $C2$ etc. both as names of constraints and as stochastic variables over the

after adding the noise values, the actual values of the constraints define a total ranking. This generalizes to systems with more than two constraints. 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. To take the example above, suppose there are two candidates, and the first violates only C1 and the second only C2. In 64% of all cases, $C1 > C2$, and thus the first candidate will be selected as optimal. However, in 36% of all evaluation events, $C2 > C1$ and thus the second candidate wins. 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, $c1 - c2$.¹³ It is depicted in figure 3.

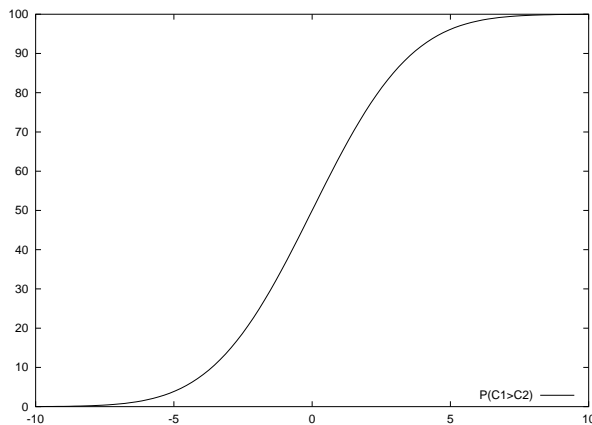


Fig. 3: probability of $C1 > C2$ as a function of $c1 - c2$ in %

If $c1 = c2$, both have the same chance to outrank the other, and accordingly $P(C1 > C2) = 50\%$. 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 2, say, the probability that C1 outranks C2 is already 76%. A difference of 5 units corresponds to a chance of 96% that $C1 > C2$. Candidates that are favored by C2 are a rare exception in a language described by such a grammar, but they are still possible. If the difference is larger

actual values of these constraints.

¹³ To be precise, the dependency is the distribution function of a normal distribution with mean = 0 and standard deviation = $2\sqrt{2}$; cf. Boersma (1998), p. 284.

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 C2, 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 constraint ranking is the borderline case of the stochastic evaluation. It obtains if the distances between the constraints are sufficiently large.

6 The Gradual Learning Algorithm

StOT is equipped with a learning algorithm that extracts a constraint ranking from a representative sample of a language—Boersma’s Gradual Learning Algorithm (GLA; see Boersma 1998; Boersma & Hayes 2001). A note of caution is in order here: the algorithm only learns a constraint ranking. Both **GEN** and the inventory of constraints have to be known in advance. Furthermore, the algorithm requires as input an analyzed corpus, i.e. a set of input-output pairs. (These are pairs of phonological and phonetic representations in the realm of phonology where this system was originally developed. In the present context this amounts to meaning-form pairs.) At every stage of the learning process, the algorithm has its own hypothetical StOT grammar. When it is confronted with an observation, it generates an output for the observed input according to its current grammar and compares it to the observed output. If the two outputs coincide, the observation is taken as confirmation of the hypothetical grammar and no action is taken. If there is a mismatch though, the constraints of the learner’s grammar are re-ranked in such a way that the observed output becomes more likely and the output that the learner produced on the basis of its hypothetical grammar becomes less likely. This process is repeated until further observations do not lead to significant changes of the learner’s grammar anymore. If the training corpus is a (sufficiently large) representative sample of a language that was generated by a StOT-grammar G (which is based on the same **GEN** and constraint set that the learner assumes), the grammar to which the algorithm converges describes a language that assigns the same probabilities to all candidates as G . So the learned grammar reproduces the statistical patterns from the training corpus, not just the categorical distinctions between grammatical and ungrammatical. Note that the algorithm is *error-driven*—the learner revises his hypothesized grammar only if there is a discrepancy between the observations and her own preferences.

Schematically, the algorithm goes through six different stages during the learning process:

- **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 $\langle i, o \rangle$.
- **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 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 4: 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.

There are several numerical parameters involved that influence the behavior of the GLA to a certain degree. I assume here that all constraints start with the initial value 0 (Boersma & Hayes 2001 use 100 here). The concrete value is totally inessential. The plasticity value is crucial for the impact of a single observation and thus for the speed of learning. A high plasticity accelerates learning at the expense of allowing a single observation to have a high impact. Conversely, a low plasticity makes the algorithm slower but more robust.¹⁴ In the context of the present paper, I will use a plasticity of 0.01.

¹⁴ Boersma (1998) assumes that the plasticity value decreases over time. This is in fact essential to ensure that the algorithm converges. Keeping the plasticity constant lets the algorithm oscillate around the grammar to be learned without getting closer. For all practical purposes, a small constant value for plasticity is good enough though.

7 GLA and grammaticalization

Cable (2002) makes an ingenious proposal regarding how the GLA can be used to explain the shift from pragmatic to structural DCM—a problem that was largely left open by ZJ. Suppose a language has reached a stage where pragmatic DCM is mandatory, while there is no structural DCM (yet). Let us also assume, for the sake of the argument, that the relative frequencies are as in the SAMTAL corpus (see figure 2), and that on the average one out of two NPs is unambiguous with respect to its grammatical role (for instance due to word order). We restrict attention to the contrast $+/-$ animacy. (Cable’s example is the contrast between local persons and third person, which makes no difference to the argument.) In this language, there is never case marking on animate subjects or inanimate objects because such NPs are either unambiguous to start with, or if they are ambiguous, BIAS assigns them the correct interpretation. Disharmonic combinations (inanimate subjects and animate objects) are marked whenever they are otherwise ambiguous, i.e. in 50% of all cases by assumption. Now suppose this language is fed into the GLA based on a **GEN** that supplies both ergative and accusative morphology. The constraints to be ranked are Aissen’s: $*\text{Subj}/\text{anim}\&*\emptyset$, $*\text{Subj}/\text{inanim}\&*\emptyset$, $*\text{Obj}/\text{anim}\&*\emptyset$, $*\text{Obj}/\text{inanim}\&*\emptyset$, and $*\text{STRUC}$. Since in the language under discussion 50% of all inanimate subjects are case marked, the GLA converges to a ranking where $*\text{Subj}/\text{inanim}\&*\emptyset$ and $*\text{STRUC}$ have the same rank (and thus their two possible rankings with respect to each other are equally likely, leading to a 50% preference in favor and a 50% preference against ergative marking of inanimate subjects). The same applies to animate objects. Animate subjects and inanimate objects are never case marked, so the constraints $*\text{Subj}/\text{anim}\&*\emptyset$ and $*\text{Obj}/\text{inanim}\&*\emptyset$ end up being ranked well below $*\text{STRUC}$ so that it is virtually impossible for them to outrank $*\text{STRUC}$. A ranking with these properties would be

- (12) a. $*\text{STRUC} = *\text{Subj}/\text{inanim}\&*\emptyset = *\text{Obj}/\text{anim}\&*\emptyset = 5$
b. $*\text{Subj}/\text{anim}\&*\emptyset = *\text{Obj}/\text{inanim}\&*\emptyset = -5$

The next generation uses this grammar but also employs pragmatic DCM for disambiguation. Hence it will also never use case marking for harmonic combinations—neither the grammar nor pragmatics gives a reason to do so. Now the grammar requires that 50% of all disharmonic NPs are case marked, but the correlation between case marking and ambiguity is lost. On the average half of the ambiguous and half of the unambiguous disharmonic NPs are marked for grammatical reasons. If a disharmonic NP is *per se* ambiguous and happens to be unmarked by the grammar, the pragmatic DCM strategy requires it to be marked nevertheless. Hence in the end this generation

will use case marking for 75% of all disharmonic NPs. The next generation will thus learn a grammar where $*\text{Subj}/\text{inanim}\&*\emptyset$ and $*\text{Obj}/\text{anim}\&*\emptyset$ are placed 2 units higher than $*\text{STRUC}$ to mimic this 75/25 proportion, while $*\text{Subj}/\text{anim}\&*\emptyset = *\text{Obj}/\text{inanim}\&*\emptyset$ are again way below $*\text{STRUC}$. The grammar that is learned by this generation looks like:

- (13) a. $*\text{Subj}/\text{inanim}\&*\emptyset = *\text{Obj}/\text{anim}\&*\emptyset = 7$
 b. $*\text{STRUC} = 5$
 c. $*\text{Subj}/\text{anim}\&*\emptyset = *\text{Obj}/\text{inanim}\&*\emptyset = -5$

By the same kind of reasoning, in the next generation this ratio will rise to 87.5% and so on. After 10 generations 99.9% of all disharmonic NPs but still none of the harmonic NPs are case marked. In other words, pragmatic DCM has turned into structural DCM.

8 Learning and bidirectionality

Cable’s approach solves one big problem that ZJ leave open: it describes a precise mechanism of grammaticalization of DCM, the shift from pragmatic towards structural DCM. The other big problem is still open: the whole mechanism is driven by pragmatic DCM which in turn is based on the constraint BIAS and thus mixes grammar with statistical tendencies. Also, Cable’s mechanism in a sense assumes that the learner is pragmatically ignorant—it is confronted with pragmatic DCM and mistakenly analyzes it as optional structural DCM. After completion of learning, however, the next generation re-invents pragmatic DCM on top of the acquired structural DCM. So the learner uses a different type of grammar than the adult speaker.

These shortcomings can be overcome by extending bidirectional optimization to the learning process. Assume that the training corpus is drawn from a language that was generated by a StOT-grammar based on bidirectional optimization in the sense of definition 1. (As argued in the discussion of ZJ above, this has the advantage that pragmatic DCM is integrated into the OT machinery.) Accordingly, the same bidirectional notion of optimality should be used by the learning algorithm in the second step (generation). Recall that the learner takes the observed input and generates an output for that input on the basis of her current hypothesized grammar. This output has to be optimal on the basis of the hypothesized grammar, and in the bidirectional version of the GLA, “optimal” means “bidirectionally optimal”.

There is a minor problem with this adjustment though. For the generation step of the GLA to succeed, it has to be guaranteed that there is some

optimal output for each observed input. This is always the case according to standard (unidirectional) optimization, but it need not be the case if one uses bidirectional optimization in the sense defined above.¹⁵ To remedy this, the definition of bidirectional optimality has to be modified somewhat. In its present form, a form is optimal for a given meaning if it is the best option among the hearer-optimal forms for this meaning. We have to add the clause that the optimal form should be the best hearer-optimal form *if there is any*. If no possible form for a given meaning is hearer optimal for this form, we ignore the requirement of hearer optimality. Formally this reads as

Definition 2:

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

You can think of the requirement of hearer-optimality as another constraint that outranks all other constraints. If it is possible to fulfill it, the optimal candidate must do so, but if it cannot be fulfilled it is simply ignored.¹⁶

Using this notion of optimality together with the GLA, learning involves interpretation as well as generation. This idea of bidirectional learning can be pushed even further by assuming that the learner assumes the hearer perspective and the speaker perspective simultaneously. In Boersma's version of the GLA, the learner observes a datum $\langle m, f \rangle$, generates a pair $\langle m, f' \rangle$ which is optimal according to her own grammar, and then compares f with f' . Bidirectional learning means that the learner also interprets f according

¹⁵ To take a simple example, suppose there are two inputs, i_1 and i_2 and one output, o . \mathbf{GEN} relates both inputs to the single output. There is only one constraint that is fulfilled by $\langle i_1, o \rangle$ but violated by $\langle i_2, o \rangle$. Hence $\langle i_1, o \rangle < \langle i_2, o \rangle$, and so $\langle i_1, o \rangle$ is hearer-optimal while $\langle i_2, o \rangle$ is not. There is no hearer-optimal, and thus no optimal, output for i_2 .

¹⁶ The idea to implement bidirectional optimality by using hearer-optimality as a constraint within a speaker oriented evaluation mechanism is inspired by Beaver (2000). There a version of hearer-optimality is a regular constraint that can even be outranked by other constraints. I'm a bit more conservative here by treating bidirectionality as a part of the evaluation component; so it can never be outranked, and it is not subject to stochastic perturbation.

to her own grammar and compares the result with the observation.¹⁷ Formally, the learner generates a pair $\langle m', f \rangle$ which is optimal according to her own grammar, and compares m with m' . The next steps—comparison and adjustment—are applied both to m/m' and f/f' . So the bidirectional version of the GLA—call it “Bidirectional Gradual Learning Algorithm” (BiGLA)—is as follows:

- **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 $\langle m, f \rangle$.
- **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 $\langle m, f' \rangle$ and $\langle m', f \rangle$ 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 $\langle m, f \rangle$ with the self-generated pair $\langle m, f' \rangle$.
- **Step 3.2: Comparison of meanings** If $m = m'$, nothing happens. Otherwise, the algorithm compares the constraint violations of the learning datum $\langle m, f \rangle$ with the self-generated pair $\langle m', f \rangle$.
- **Step 4: Adjustment**
 - All constraints that favor $\langle m, f \rangle$ over $\langle m, f' \rangle$ are *increased* by the plasticity value.
 - All constraints that favor $\langle m, f' \rangle$ over $\langle m, f \rangle$ are *decreased* by the plasticity value.
 - All constraints that favor $\langle m, f \rangle$ over $\langle m', f \rangle$ are *increased* by the plasticity value.

¹⁷ In chapter 15 of Boersma (1998), Boersma also considers a purely hearer-oriented version of GLA. There the learner only compares competing interpretations for the observed form. The idea of *bidirectional* learning, i.e. of simultaneous speaker-oriented and hearer-oriented learning is to my knowledge new though.

- All constraints that favor $\langle m', f \rangle$ over $\langle m, f \rangle$ are *decreased* by the plasticity value.
- **Final state** Steps 1 – 4 are repeated until the constraint values stabilize.

9 BiGLA and DCM

In this section I will argue that the BiGLA combines the advantages of the ZJ-approach to pragmatic DCM and of Cable’s theory of grammaticalization. To see this point, let us do a thought experiment. 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,
- always respects FAITH, and
- uses case marking in exactly 50% of all cases, but in a way that is totally uncorrelated to animacy. For each clause type, in 25% of all cases no case marking is used, in 25% the subject is ergative marked and the object is unmarked, in 25% the subject is unmarked and the object accusative marked, and in 25% both NPs are case marked.

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 both ergative and accusative morphology, which are both optional. The linking of case morphemes to grammatical

functions is governed by the FAITH constraint, so **GEN** imposes no restrictions in this respect. **GEN** thus admits nine types of morphological marking within a clause: both NP1 and NP2 can be ergative marked, accusative marked or unmarked. This gives nine different form patterns. If $+/-\text{anim}$ is taken into account, we get 36 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 72 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 4.

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

Tab. 4: Frequencies of clause types in SAMTAL

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.

Table 5 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 (E, A and Z abbreviate “ergative”, “accusative” 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 $*(\text{su}/\text{a}/\text{z})$, $*(\text{su}/\text{i}/\text{z})$, $*(\text{ob}/\text{a}/\text{z})$, and $*(\text{ob}/\text{i}/\text{z})$. They all enforce case marking. They are counteracted by $*\text{STRUC}$ 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 constraint FAITH takes care of the linking between case morphemes and grammatical roles. It is violated always if an ergative marked NP is interpreted as an object or an accusative NP as a subject. Finally I assume that the grammar does distinguish between interpreting NP1 or NP2 as a subject. In real languages there are many constraints involved here (pertaining to syn-

	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

Tab. 5: Training corpus

tax, 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 eight constraints:

1. *(su/a/z): *Avoid unmarked animate subjects!*
2. *(su/i/z): *Avoid unmarked inanimate subjects!*
3. *(ob/a/z): *Avoid unmarked animate objects!*
4. *(ob/i/z): *Avoid unmarked inanimate objects!*
5. *STRUC: *Avoid case marking!*
6. FAITH: *Avoid ergative marked objects and accusative marked subjects!*
7. SO: *NP1 is subject and NP2 object.*
8. OS: *NP2 is subject and NP1 object.*

All these constraints are set to the initial value 0 and the BiGLA is applied to a training corpus with the frequencies as in table 5. What is the learning effect of the different observations? Suppose the algorithm is confronted with a clause containing an ergative marked animate subject. In speaker mode, the algorithm produces its own form for the observed meaning (su/a),

which may be either ergative marking as well, or else accusative marking or zero marking. The constraint violation profiles of the three candidates at hand are given in (). (For simplicity, I leave out the last two constraints. The horizontal ordering of the constraints is arbitrary and should not be interpreted as a ranking.)

(14)

	* $(su/a/z)$	* $(su/i/z)$	* $(ob/a/z)$	* $(ob/i/z)$	*STRUC	FAITH
su/a/erg					*	
su/a/acc					*	*
su/a/z	*					

If the learner's form coincides with the observation, nothing happens. Otherwise, all constraints that favor the observation over the learner's output will be promoted, and all constraints that favor the learner's hypothesis will be demoted.

If the learner chooses accusative as its own hypothesis, there is only one constraint that distinguishes between observation and hypothesis, namely FAITH. It favors the observation over the hypothesis and is thus promoted in this scenario. If the learner chooses zero marking, $*(su/a/z)$ favors the observation and is thus promoted, while *STRUC favors the hypothesis and is demoted. The effect of observing other case marked NPs is analogous. So the net effect of observing case marked NPs under the speaker perspective is

- promotion of $*(su/a/z)$, $*(su/i/z)$, $*(ob/a/z)$, $*(ob/i/z)$, and FAITH
- demotion of *STRUC

Observing unmarked NPs has by and large the opposite effect. If an animate subject with zero marking is observed, a mismatch can occur if the learner produces accusative marking or ergative marking. Both will cause a promotion of *STRUC and a demotion of $*(su/a/z)$. In the former case, we will additionally get a promotion of FAITH. The same applies *mutatis mutandis* for other unmarked NPs. So in total observing unmarked NPs in speaker mode has the following total learning effect:

- promotion of *STRUC and FAITH
- demotion of $*(su/a/z)$, $*(su/i/z)$, $*(ob/a/z)$ and $*(ob/i/z)$

Since there is the same number of marked and unmarked NPs in the training corpus, we expect these competing forces to cancel out each other, with the exception of FAITH, which is always promoted. So the unidirectional GLA would come up with a grammar where FAITH is high and all other constraints remain around the initial value 0.¹⁸ Such a grammar would reproduce the distribution from the training corpus, i.e. 50% case marking respecting FAITH but uncorrelated to animacy.

However, the BiGLA also learns in hearer mode, and here the effect is different. First consider what happens if a case marked NP is observed, for instance an ergative marked animate subject. The possible interpretations are animate subject and animate object. The pattern of constraint violations of the relevant candidates is given in (15):

(15)

	*(su/a/z)	*(su/i/z)	*(ob/a/z)	*(ob/i/z)	*STRUC	FAITH
su/a/erg					*	
ob/a/erg					*	*

The latter but not the former violates FAITH. Due to the effect of speaker learning, FAITH quickly becomes the strongest constraint, so the learner will rarely, if ever, come up with a non-faithful interpretation for an observed form. Hence mismatches between observations and the learner’s interpretation are rare. Case marked NPs thus have almost no learning effect in hearer mode.

This is dramatically different for unmarked NPs. Suppose the learner is confronted with an unmarked animate *subject*.

(16)

	*(su/a/z)	*(su/i/z)	*(ob/a/z)	*(ob/i/z)	*STRUC	FAITH
su/a/z	*					
ob/a/z			*			

Now both interpretations, as subject and as object, are consistent with FAITH. So an object interpretation and thus a mismatch is possible. This will lead to a promotion of *(ob/a/z) and a demotion of *(su/a/z). Observing an animate *object*, a mismatch has the opposite effect—promotion of *(su/a/z) and demotion of *(ob/a/z). There are about nine times as many

¹⁸ Due to the symmetry of the training corpus with respect to linking, OS and SO are promoted and demoted by the same amount and both remain close to 0.

animate subjects as animate objects in the training corpus though. So the net effect of observing unmarked animate NPs is a promotion of $*(ob/a/z)$ and a demotion of $*(su/a/z)$. For inanimate NPs this is exactly the other way round. Here the objects outnumber the subjects roughly by the factor 14. Hence in total $*(su/i/z)$ will be promoted and $*(ob/i/z)$ demoted. To summarize, the net effect of learning in hearer mode is

- promotion of $*(su/i/z)$ and $*(ob/a/z)$
- demotion of $*(su/a/z)$ and $*(ob/i/z)$

So bidirectional learning has the effect that the asymmetries in the frequencies of NP types in the training corpus lead to an asymmetric ranking of the corresponding constraints in the learned grammar. Note that Aissen’s sub-hierarchies are in fact induced from the statistics of language use here: $*(su/i/z) \gg *(su/a/z)$, and $*(ob/a/z) \gg *(ob/i/z)$.

A computer simulation revealed that the above considerations are largely correct (except that there is a net demotion of $*STRUC$). The BiGLA was fed with 50,000 observations which were drawn at random from a distribution as in table 5. The constraint rankings that were acquired are give in table 6.

$*(su/a/z)$:	−1.32
$*(su/i/z)$:	2.89
$*(ob/a/z)$:	0.92
$*(ob/i/z)$:	−1.07
$*STRUC$:	−1.05
FAITH:	7.94
OS:	−0.03
SO:	0.03

Tab. 6: Grammar that was acquired by the BiGLA from the corpus with random case marking

The development of the rankings of the constraints are plotted in figure 4. The x-axis gives the number of observations (in thousands) and the y-axis the ranking of the constraints.

In this grammar, FAITH is by far the strongest constraint. Hence the language described by this grammar never uses case marking in an unfaithful way. Further, the disharmonic constraints $*(su/i/z)$ and $*(ob/a/z)$ are ranked well above $*STRUC$. So case marking of disharmonic NPs is strongly preferred (the distance between the relevant competing constraints is about

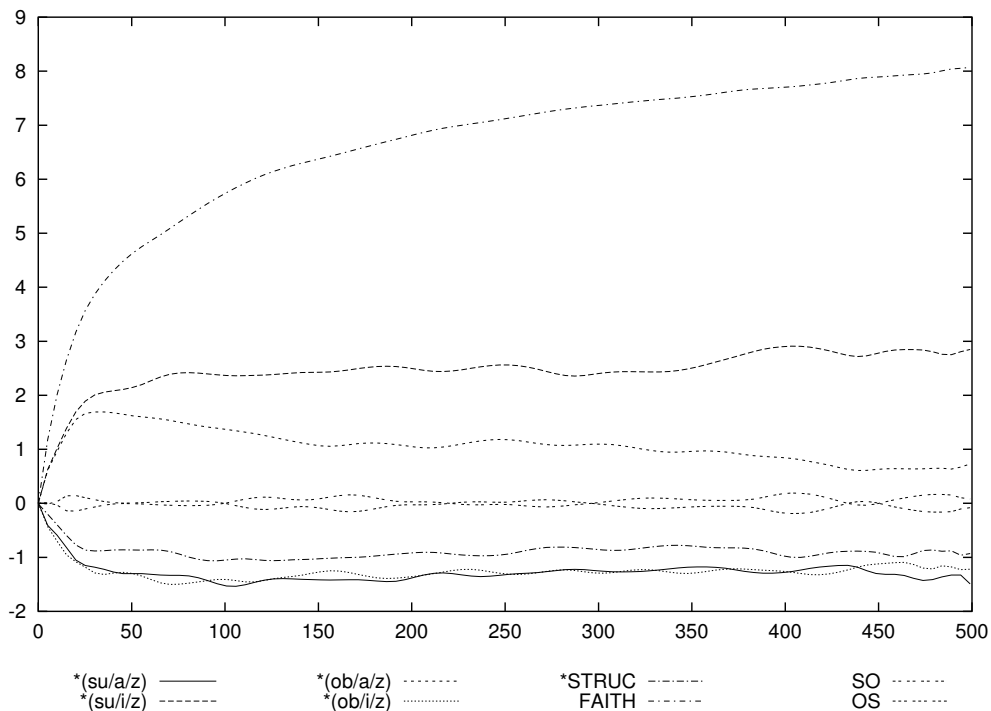


Fig. 4: Learning curves

4.0 and 2.0 units respectively, which corresponds to a strong preference, but not a categorical rule). The harmonic constraints $*(su/a/z)$ and $*(ob/i/z)$ have about the same ranking as $*STRUC$ —case marking of harmonic NPs is thus totally optional.

These considerations apply if an NP is unambiguous. For an ambiguous unmarked NP, the harmonic interpretation is always preferred because $*(ob/a/z) \gg *(su/a/z)$ and $*(su/i/z) \gg *(ob/i/z)$. To achieve bidirectional optimality, this tendency has to be counteracted by using case marking for disharmonic NPs, while harmonic NPs also receive the correct interpretation without case marking. Hence on top of the preference for structural DCM, there is an even stronger tendency for pragmatic DCM.

The chart below gives the relative frequencies of NP types in a corpus that was generated by maintaining the proportions of meanings from the SAMTAL corpus but using the grammar from table 6.

Let us consider all cells where the object is accusative marked and the subject is thus not in danger of being understood as an object. Ergative marking is redundant. It is nevertheless used in 60.6% of all cases. These 60.6% are not equally distributed over animate and inanimate subjects. 95.7% of

	E-E	E-A	E-Z	A-E	A-A	A-Z	Z-E	Z-A	Z-Z
su/a-ob/a	0.0	1.59	0.43	0.0	0.0	0.0	0.0	2.17	0.57
su/a-ob/i	0.0	12.09	7.05	0.0	0.0	0.0	0.0	8.68	13.65
su/i-ob/a	0.0	0.16	0.17	0.0	0.0	0.0	0.0	0.03	0.0
su/i-ob/i	0.0	1.41	1.29	0.0	0.0	0.0	0.0	0.48	0.21
ob/a-su/a	0.0	0.0	0.0	2.08	0.0	1.92	0.29	0.0	0.48
ob/a-su/i	0.0	0.0	0.0	0.29	0.0	0.0	0.79	0.0	0.0
ob/i-su/a	0.0	0.0	0.0	13.49	0.0	8.68	7.12	0.0	12.98
ob/i-su/i	0.0	0.0	0.0	1.32	0.0	0.63	1.46	0.0	0.11

Tab. 7: Corpus that was generated by the acquired grammar

all (unambiguous) inanimate subjects, but only 58.3% of all (unambiguous) animate subjects carry ergative case. The same pattern can be observed for objects. Redundant accusative marking is used in 65.2% of all cases. However, 83.0% of the animate objects, but only 63.3% of the inanimate objects are accusative marked (if they co-occur with an ergative marked subject). So we in fact see a clear preference for structural DCM.

This effect is more dramatic if we consider potentially ambiguous NPs. In total, 38.9% of all subjects that co-occur with an unmarked object are ergative marked. For animate subjects, this figure is 34.8%, but for inanimate subjects it is 90.4%. As for the objects, 43.6% of objects in a clause with an unmarked subject are accusative marked. For animate objects, this figure rises to 79.8%, while for inanimate objects it is only 39.4%. Of course case marking of subjects and objects influence each other: for the most harmonic meaning (animate subject and inanimate object) 31.5% of all clauses use no case marking at all, while for the least harmonic meaning (inanimate subject and animate object) case marking is 100% obligatory, only the choice between subject marking, object marking or both is optional. So in sum we see that pragmatic DCM is also present on top of structural DCM.

10 The next generation

The sample corpus that was generated with the acquired grammar can of course be used as input to a second run of the BiGLA. This procedure may be repeated over several “generations”. In this way the BiGLA can be used to simulate diachronic development. The successive constraint rankings that emerge in this way are plotted in figure 5. 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. While there are no rough

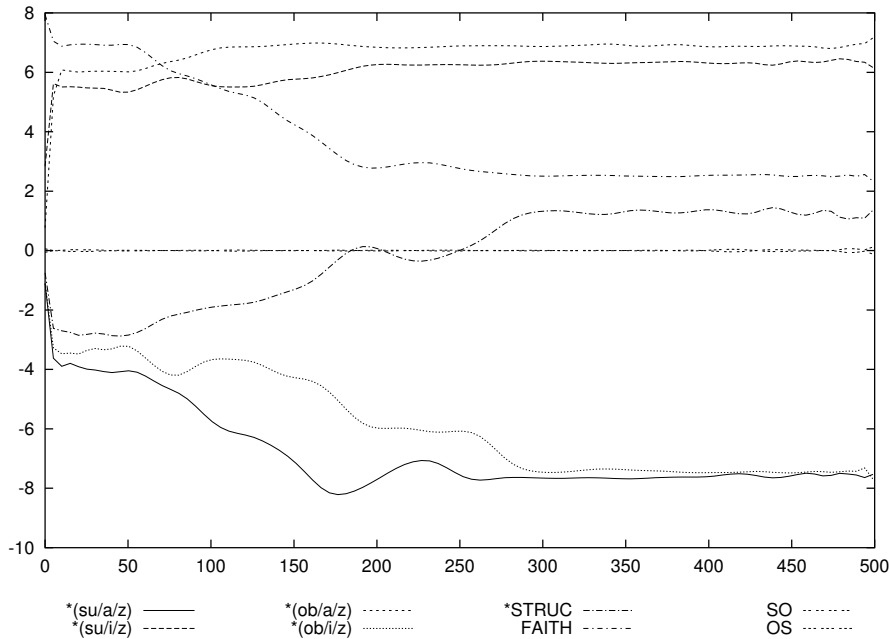


Fig. 5: Diachronic development

changes from one generations to the next, the grammar as a whole gradually changes its characteristics over time. The Aissen sub-hierarchies— $*(su/i/z) \gg *(su/a/z)$ and $*(ob/a/z) \gg *(ob/i/z)$ —are always respected 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 11–60). Then the system becomes unstable. The two constraints pertaining to the disharmonic combinations— $*(su/i/z)$ and $*(ob/a/z)$ —remain high. However, the two “harmonic” constraints $*(su/a/z)$ and $*(ob/i/z)$ are gradually lowered while $*STRUC$ rises. During this process, $*STRUC$ assumes a position strictly below the disharmonic but strictly above the harmonic case marking constraints. This amounts to a gradual loss of case marking of harmonic NPs, while marking of disharmonic NPs remains obligatory. At around generation 280 this process is completed, and in the remaining

220 generations the system remains stable in a state where case marking is obligatory for disharmonic and prohibited for harmonic NPs. An almost¹⁹ categorical split ergative system has emerged.

The development of the probabilities for of structural (i.e. redundant) case marking of an NP of a given semantic type are given in the first graphics of figure 6. There the gradual loss of case morphology at harmonic NPs is easy to discern.

Needless to say that the diachronic development that is predicted by the BiGLA (together with **GEN**, 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 devices have been reanalyzed as case morphemes), the overall development is similar to the previous set up. The ranking that BiGLA induces from the initial corpus places *STRUC extremely high (at 35.25), 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 -21.33 , *(su/i/z) at 4.38, *(ob/a/z) at 1.26 and *(ob/i/z) at -21.07 . During the following 50 generations *STRUC is constantly lowered until it assumes a position half-way between the harmonic and the disharmonic constraints. The ranking that thus emerges is qualitatively identical to the steady state that was reached after 280 generations in the previous experiment. On the corpus side, this means that the probability of a disharmonic NP to be case marked gradually rises from 0% to 100% within 50 generations, while harmonic NPs remain obligatorily unmarked. Again, the emerging split ergative system is a steady state. The change of the case marking probabilities over time is depicted in the second graphics of figure 6.

So if the initial training corpus does not display a correlation between animacy and case marking, the iteration of bidirectional learning with the said constraint set and lexicon shows an inherent tendency towards split ergative systems.

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

¹⁹ In a corpus that was generated by the grammar of the 500th generation, more than 95% of all NPs follow the split ergative pattern.

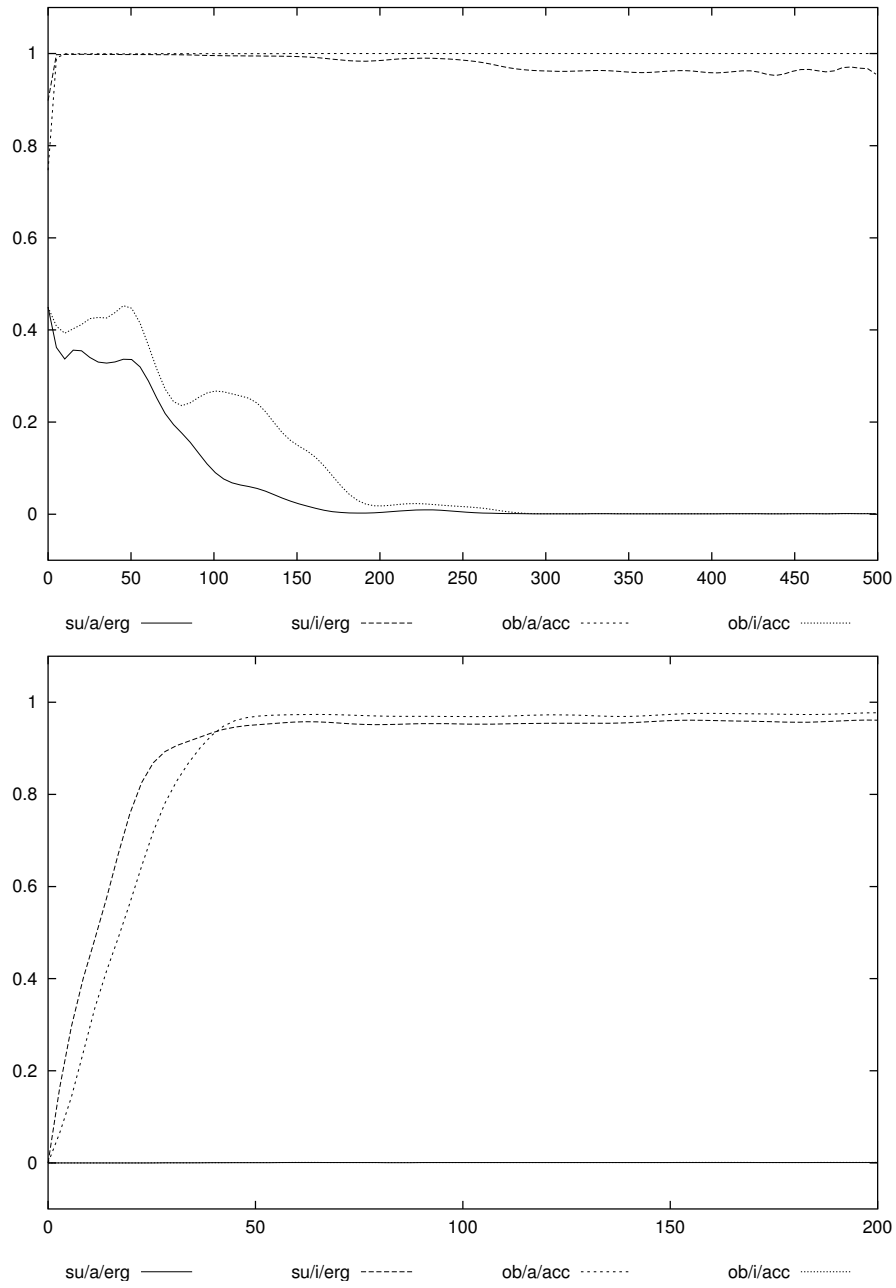


Fig. 6: Probabilities of case marking

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 and of the case marking probabilities is given in the figures 7 and 8. The BiGLA in fact learns the inverse pat-

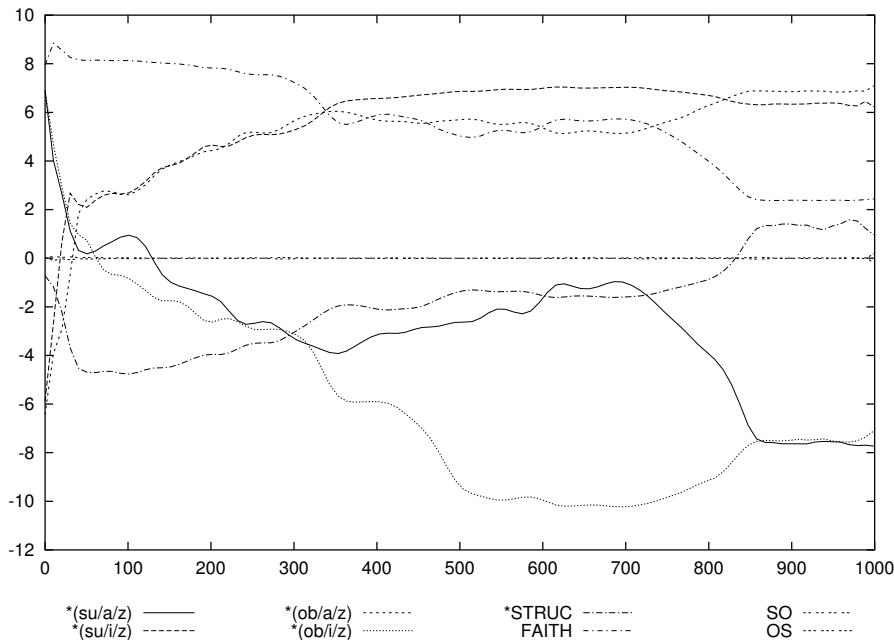


Fig. 7: The future of anti-DCM: constraint rankings

tern, 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. After fifteen generations the Aissen sub-hierarchies emerge and remain stable for the remainder of the simulation (which ran over 1000 generations). Nonetheless, the case marking patterns changed dramatically after that. For about 100 generations after the emergence of the Aissen hierarchies, case marking is virtually obligatory for all NPs. This corresponds to a ranking where $*STRUC$ is ranked very low. This phase is followed by a smooth raising of $*STRUC$, accompanied by a simultaneous lowering of $*(su/a/z)$ and $*(ob/i/z)$, until all three constraints are roughly at the same level. This means that case marking of harmonic

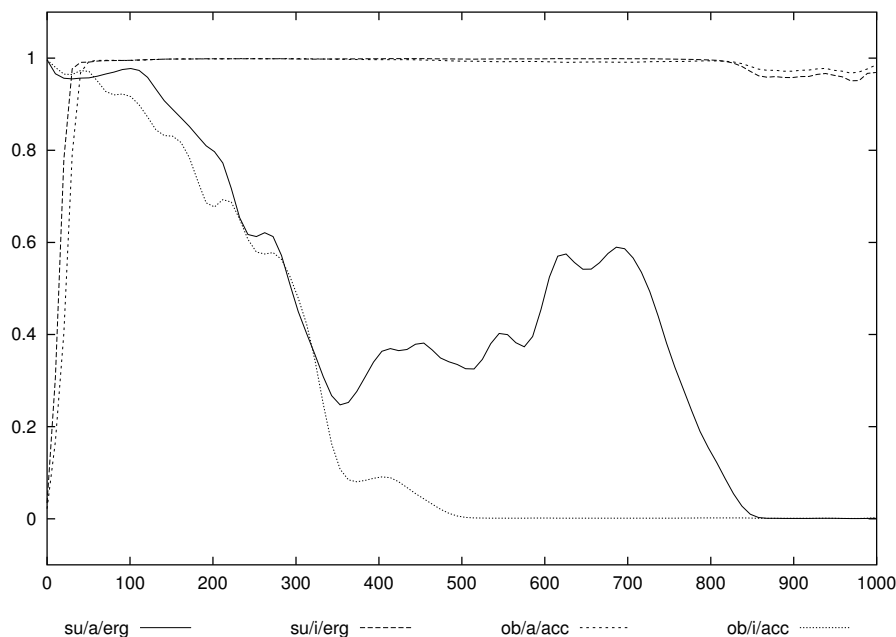


Fig. 8: The future of anti-DCM: case marking probabilities

NPs becomes optional while marking of disharmonic NPs remains obligatory. During the subsequent 500 generations, the symmetry between subjects and objects is broken. Accusative marking of inanimate NPs is totally lost, while ergative marking of animate NPs stays optional. After a final crisis where $*(su/a/z)$ is lowered and hence ergative marking of animates is lost, the system also enters the steady state of split ergativity.

A large number of further simulations indicated that split ergativity is in fact the only stable state under the side conditions used here, i.e. the constraint set, the generator and the relative probabilities of the possible interpretations. 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 generator relation that was used in the above experiments represents a typologically marked language type because each NP has both an ergative and an accusative form next to the unmarked form. Such tripartite systems exist but are very rare. In most languages, each NP has at most two morphological forms for the syntactic core functions. In most split ergative languages, some NPs have a special ergative and other NPs a special accusative form next to the unmarked one, but no NP has both. So another plausible approximation to a lexicon would stipulate only two morphological forms for each NP, unmarked and marked, and leave the interpretation of the marked form as ergative or accusative to the constraint ranking.²⁰

In this setup, each transitive clause type has four morphological variants because both NPs can be either marked or unmarked each. We still have eight possible meanings. A training corpus with 50% probability of case marking for each NP type (using the SAMTAL distribution of meanings) thus looks as in table 8. Here “M” stands for “marked.”

	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

Tab. 8: Training corpus

In the previous setup, the interpretation of the case morphemes was taken care of by the constraints FAITH. Since here we only have one case morpheme, this constraint has to be split up into two, one favoring an accusative and one an ergative interpretation of this morpheme.

6.1 : $m \Rightarrow su$: *Marked NPs are subjects.*

6.1 : $m \Rightarrow ob$: *Marked NPs are objects.*

²⁰ For the purposes of this paper, I equate the generator relation with the lexicon and hence do not assume the generator to be universal. A more refined model of learning has thus to include the acquisition of the generator as well. For the time being, I ignore this issue for the sake of simplicity.

The development of the constraint rankings under this setup is given in the first graphics of figure 9.

Here it takes about 400 generations before a steady state is reached. The stable ranking is virtually categorical with three strata, namely

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

This ranking corresponds to a split ergative pattern. Systematic experimentation showed that as in the previous setup, split ergativity is in fact the only steady state, regardless of the nature of the initial training corpus.

However, 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 where both arguments are animate (300) has the same order of magnitude as the number of clauses with two inanimate arguments (186). If we look for instance at definiteness instead, this is different. Here the absolute frequencies are as in table 9.

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

Tab. 9: Frequencies of clause types with respect to definiteness

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 experiment. It is given in the second graphics of figure 9.

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.):

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

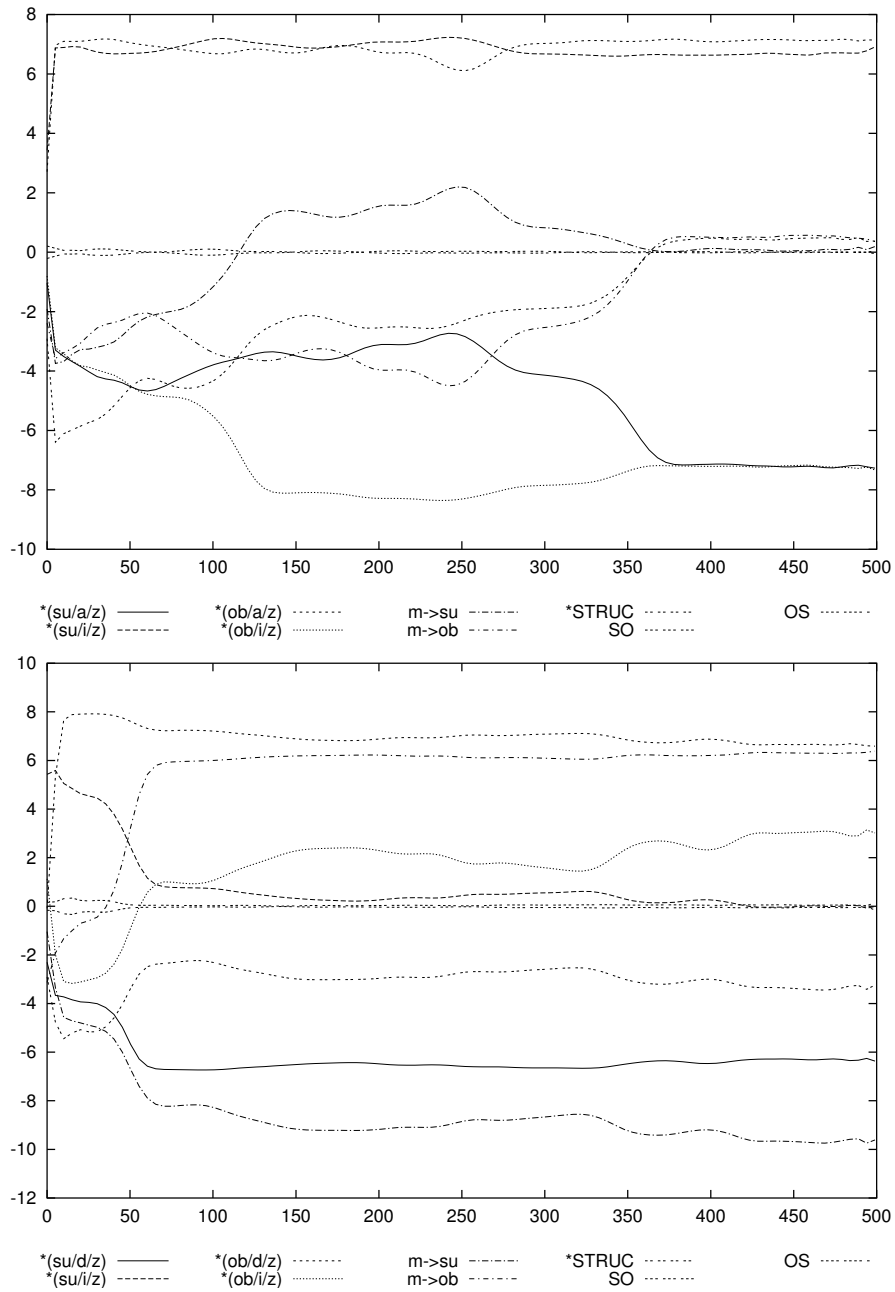


Fig. 9: Simulation using only two forms per NP: animacy and definiteness

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 $m \Rightarrow \text{obj} \gg m \Rightarrow \text{su}$ 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 constraint ranking which describes such a language, namely

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

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

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

As can be seen from figure 6, this corresponds to a language without *structural* case marking. (Structural case marking only evolves in the second generation.) However, 16.9% of the NPs in a sample corpus drawn from this language carry case marking nevertheless. In other words, this language has *pragmatic* case marking.

The third group consists of languages that are both in accordance with UG and learnable (in the sense that the BiGLA reproduces a language with a similar characteristics), 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 8). There the language type switches from inverse split ergativity to obligatory case marking within less than twenty generations.

There are different degrees of instability. In the third experiment reported above, a pattern with categorical DOM and optional DSM would last as long as 400 generations before it changed to categorical split ergativity.

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 two and nominative-accusative in the third scenario.²¹

Schematically expressed, this predicts the following hierarchy of language types according to their likelihood:

1. *diachronically stable and target of diachronic change*: split ergative (first two scenarios), nominative-accusative (third scenario)
2. *diachronically moderately stable*: optional DSM paired with categorical DOM (first scenario)
3. *diachronically very unstable*: inverse DCM
4. *unlearnable*: no case marking, random case marking
5. *not UG-conform*: zero or two case markers per clause

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.

²¹ It is of course possible to construct artificial scenarios with several equilibria due to perfect symmetry.

11 Conclusion and open questions

In this paper I proposed a revised version of Boersma's Gradual Learning Algorithm. It incorporates the concept of bidirectional optimization in two ways. First it uses a notion of optimality of an input-output pair that takes both the hearer perspective and the speaker perspective into account. Second, learning is thought of as bidirectional as well. The learner gradually adjusts both its production and its interpretation preferences to the observations.

The working of this Bidirectional Gradual Learning Algorithm was applied to Aissen's theory of differential case marking. It could be shown that the constraint sub-hierarchies that Aissen simply assumes to be universal emerge automatically via learning if the training corpus contains substantially more harmonic meanings than disharmonic ones. This connection between harmony and frequency has been pointed out and used in ZJ's approach before. The present system diverges from ZJ in assuming that learning mediates between statistical biases in the language use and grammatical biases as expressed by the Aissen hierarchies, while ZJ simply identify these biases. Several computer simulations confirmed the correlation between the statistical patterns of usage in a training corpus and the characteristics of the grammars induced from these corpora by the BiGLA.

In these experiments, just the correlation between grammatical functions with the binary contrast animate/inanimate in simple transitive active clauses was studied. Further investigations will have to use more informed models. In particular the effect of using a more articulated and perhaps two-dimensional substantive hierarchy (the combination of the definiteness hierarchy with the animacy hierarchy) as well as the effect of diathesis should be studied.

There are also several theoretical questions pertaining to the BiGLA to be addressed. The most important one is the problem of convergence of learning. By definition, a learning algorithm for a stochastic language should converge to a grammar for the learned language provided the training corpus is a representative sample of the language. The BiGLA obviously does not have this property; otherwise every language type would be stable. So is it adequate to call the BiGLA a learning algorithm to start with?

There are several points involved here. First, since the BiGLA is based on a version of bidirectional StOT, it is only supposed to learn languages that are described by a grammar from this class. That non-UG-conforming languages are not learned is thus no problem. However, there are languages that correspond to some constraint ranking, but yet the BiGLA returns the grammar of a language that slightly or massively differs from the training language. The unidirectional GLA does not have this property. However, the conver-

gence condition just requires that a learning algorithm maps *representative* samples of a language to a grammar for that language. In unidirectional StOT, this means that the different possible outputs for a given input are distributed according to the conditional probabilities that the grammar assigns to them. The relative frequencies of the inputs (=meanings) has no impact on the learning result. This is different with bidirectional learning. Here also the relative frequencies of the different meanings of a given form in the training corpus have to converge towards their grammatically determined conditional probabilities to ensure convergence of learning. Another way to state this point is this: a StOT-grammar defines a probability distribution over meaning-form pairs, and a representative sample of a language has to mirror these probabilities in frequencies. In our experimental setup, however, the marginal probabilities of the different meanings were determined by extra-grammatical factors (the relative frequencies from SAMTAL). So the conditional probabilities of the different forms for a given meaning were matched by relative frequencies, but not the probabilities of the different meanings. Hence the BiGLA only converges towards a grammar of the training language if the SAMTAL-probabilities of meanings coincide with the probabilities assigned by the grammar. This is only the case if the least marked meanings are the most frequent ones. (This is the theoretical base for the correlation between frequencies and language types that is inherent in the BiGLA.)

Still, the grammar for the language without case marking mentioned above is in equilibrium in this sense, and yet it is not learnable by the BiGLA. How is this possible? The problem here is that during the learning process the hypothesized grammar fits the training corpus better and better, but it is not guaranteed that the difference to a real grammar becomes arbitrarily small. There are several remedies possible here, but perhaps this failure to converge with certain languages is not such a severe disadvantage after all. It should be noted that the language without case marking is extremely dysfunctional. On average only 50% of all utterances are interpreted correctly by the hearer. The language that the BiGLA acquires is better adapted to usage—due to pragmatic case marking more than half of all utterances get their message across. So there is also a tendency towards functionality inherent in the BiGLA, and it meets the convergence condition for a stochastic learning algorithm only for languages that are functionally adapted in a certain way. The exact content of this condition is a subject for further research.

Both tendencies that are “built into” the BiGLA—frequent meanings should be unmarked meanings, and functional languages are better than dysfunctional ones—have been identified as important linguistic factors time and again by functional linguists (see for instance the discussions in Haspelmath

1999 and Haspelmath 2002). I expect that formal learning theory and functional linguistics can profit from each other a great deal, and I hope that the present paper illustrates the fertility of such an alliance.

Acknowledgments

This paper emerged from a series of discussions I had with Judith Aissen and Joan Bresnan. I also profited a lot from the feedback I got from Reinhard Blutner, Paul Boersma, Lena Maslova and Henk Zeevat. Last but not least, I thank Jason Mattausch for correcting my English.

References

- Aissen, J. (1999). Markedness and subject choice in Optimality Theory. *Natural Language and Linguistic Theory*, 17, 673–711.
- Aissen, J. (2000). Differential object marking: Iconicity vs. markedness. Manuscript, UCSC.
- Aissen, J. & Bresnan, J. (2002). OT syntax and typology. course material from the Summer School on Formal and Functional Linguistics. University of Düsseldorf.
- Beaver, D. (2000). The optimization of discourse. manuscript, Stanford.
- Blutner, R. (2001). Some aspects of optimality in natural language interpretation. *Journal of Semantics*, 17(3), 189–216.
- Boersma, P. (1998). *Functional Phonology*. Ph.D. thesis, University of Amsterdam.
- Boersma, P. & Hayes, B. (2001). Empirical tests of the gradual learning algorithm. *Linguistic Inquiry*, 32(1), 45–86.
- Bossong, G. (1985). *Differentielle Objektmarkierung in den neuirischen Sprachen*. Tübingen: Günther Narr Verlag.
- Bresnan, J., Dingare, S., & Manning, C. (2001). Soft constraints mirror hard constraints: Voice and person in English and Lummi. In M. Butt & T. H. King (Eds.), *Proceedings of the LFG01 Conference*. Stanford: CSLI Publications. To appear.
- Cable, S. (2002). Hard constraints mirror soft constraints! Bias, Stochastic Optimality Theory, and split-ergativity. manuscript, University of Amsterdam.
- Dixon, R. M. W. (1994). *Ergativity*. Cambridge: Cambridge University Press.
- Fry, J. (2001). *Ellipsis and wa-marking in Japanese conversation*. Ph.D. thesis, Stanford University.

- Haspelmath, M. (1999). Optimality and diachronic adaptation. *Zeitschrift für Sprachwissenschaft*, 18(2), 180–205.
- Haspelmath, M. (2002). Explaining the ditransitive person-role constraint: A usage-based explanation. manuscript, MPI für evolutionäre Anthropologie, Leipzig.
- Prince, A. & Smolensky, P. (1993). Optimality theory: Constraint interaction in generative grammar. Technical Report TR-2, Rutgers University Cognitive Science Center, New Brunswick, NJ.
- Shannon, C. (1948). A mathematical theory of communication. *Bell Sys. Tech. Journal*, 27, 379–432, 623–656.
- Silverstein, M. (1976). Hierarchy of features and ergativity. In R. M. W. Dixon (Ed.), *Grammatical Categories in Australian Languages*, (pp. 112–171). Canberra: Australian Institute of Aboriginal Studies.
- Smolensky, P. (1995). On the internal structure of the constraint component Con of UG. ROA 86. Handout of talk given at UCLA.
- Zeevat, H. & Jäger, G. (2002). A reinterpretation of syntactic alignment. In D. de Jongh & H. Zeevat (Eds.), *Proceedings of the Fourth International Tbilisi Symposium on Language, Logic and Computation*. University of Amsterdam.