# Probabilistic Learning Algorithms and Optimality Theory

*Frank Keller*
*Ash Asudeh*

This article provides a critical assessment of the Gradual Learning Algorithm (GLA) for probabilistic optimality-theoretic (OT) grammars proposed by Boersma and Hayes (2001). We discuss the limitations of a standard algorithm for OT learning and outline how the GLA attempts to overcome these limitations. We point out a number of serious shortcomings with the GLA: (a) A methodological problem is that the GLA has not been tested on unseen data, which is standard practice in computational language learning. (b) We provide counter-examples, that is, attested data sets that the GLA is not able to learn. (c) Essential algorithmic properties of the GLA (correctness and convergence) have not been proven formally. (d) By modeling frequency distributions in the grammar, the GLA conflates the notions of competence and performance. This leads to serious conceptual problems, as OT crucially relies on the competence/performance distinction.

*Keywords:* Optimality Theory, probabilistic grammars, language acquisition, corpus frequencies, degrees of grammaticality, competence/performance

## 1 Learnability and Optimality Theory: Problems and Solutions

A generative grammar is empirically inadequate (and some would say theoretically uninteresting) unless it is provably learnable. Of course, it is not necessary to provide such a proof for every theoretical grammar postulated. Rather, any generative linguistic framework must have an associated learning theory that states how grammars couched in this framework can be learned. One reason that Optimality Theory (OT; Prince and Smolensky 1993) has proven so influential in such a short time is that it was developed hand in hand with a learning algorithm for OT grammars: the Constraint Demotion Algorithm (CDA; Tesar and Smolensky 1996, 1998, 2000).[1]

Tesar and Smolensky claim that the CDA is able to learn every totally ordered constraint

We would like to thank Paul Boersma, Anette Frank, Jonas Kuhn, Maria Lapata, Chris Manning, Ida Toivonen, and two anonymous reviewers for their comments. Not all of these people will necessarily agree with the views expressed in this article, and all remaining errors are of course our own. Asudeh's work is funded in part by SSHRC Doctoral Fellowship 752-98-0424.

[1] We are aware that there are other proposals for OT learning in the literature (see, e.g., Pulleyblank and Turkel 2000, Hale and Reiss 1998). However, we will take the CDA as the standard of comparison for Boersma and Hayes's (2001) Gradual Learning Algorithm because the CDA is only minimally different and is also the most well known OT learning algorithm.

Linguistic Inquiry, Volume 33, Number 2, Spring 2002
225–244

hierarchy (i.e., OT grammar) provided it is supplied with suitable training data. Such an algorithmic claim has to be backed up by a rigorous demonstration that the algorithm works in the general case, which means that *proofs* of the algorithm's correctness and convergence have to be given. A learning algorithm is correct if it computes the correct grammar provided it is supplied with suitable training data. An algorithm converges if it yields a result on every training set (rather than oscillating indefinitely on certain sets).

Tesar and Smolensky (1998:257–265) provide proofs of the CDA's formal properties: they show that it always learns the correct grammar if given suitable training data and that it will converge on any consistent training set. This means that Tesar and Smolensky are able to provide a generative framework—OT—with an associated learning theory—the CDA. In other words, OT grammars *with totally ordered constraint hierarchies* are provably learnable. Let us call such OT grammars *Standard OT grammars*.

Although learnability is a necessary condition for a grammar's empirical adequacy, it is obviously not a sufficient condition: the grammar still has to get the linguistic facts right (i.e., it has to be descriptively adequate). There are two crucial properties of linguistic competence that Standard OT (SOT) grammars have trouble representing: one is free variation (i.e., optionality) and ambiguity, the other is gradient grammaticality (both will be discussed in more detail in section 2). These two representational problems of SOT are inherited by Tesar and Smolensky's learning theory, which cannot deal with free variation and ambiguity and is not designed to handle gradient grammaticality. In addition, the CDA lacks robustness. That is to say, it cannot deal with noisy data: errors in the training set can mean that the algorithm fails to learn the correct grammar (see section 2).

In order to deal with these deficiencies of SOT and its associated learning theory, Boersma and Hayes (2001) have proposed a modified version of OT, which we will call *Probabilistic Optimality Theory* (POT). POT comes with an associated learning algorithm, the Gradual Learning Algorithm (GLA), and is claimed to solve the problems that plague SOT: (a) it can model free variation and ambiguity, (b) it can account for gradient grammaticality, and (c) it is robust (i.e., it can learn from data that contain errors).[2]

In this article, however, we will present several problems with the GLA. More specifically, we will argue that

- since the GLA has not been tested on unseen data, it is unclear whether it is able to generalize;
- there are data sets that the GLA cannot learn;
- Boersma and Hayes (2001) offer no proof of correctness and convergence for the GLA; and
- the GLA model conflates grammaticality and corpus frequency in a way that is not compatible with standard assumptions about competence and performance.

[2] Previous incarnations of the POT framework are presented by Hayes (2000), Hayes and MacEachern (1998), and Boersma (1997, 1998, 2000), who also describes various antecedents of the GLA.

**Table 1**
Example for Ilokano metathesis variation

| /taʔo-en/ | $C_1$ | ... | $C_n$ |
|---|---|---|---|
| ☞ taʔ.wen | | | |
| ☞ taw.ʔen | | | |
| ⋮ | | | |

We will conclude that the GLA (at least in the form presented in Boersma and Hayes 2001) is seriously deficient and will have to be modified if the above-mentioned problems are to be resolved.

## 2 Free Variation, Ambiguity, Gradience, and Robustness

### *2.1 Free Variation and Ambiguity*

Free variation and ambiguity are formally the same in OT (Asudeh 2001). Each is a case of one input corresponding to multiple outputs, the former in the production direction and the latter in the comprehension direction. First let us consider free variation. As an example, take the Ilokano metathesis variation that Boersma and Hayes (2001) discuss, following Hayes and Abad (1989).[3] In Ilokano /ʔo/ can be realized as either [ʔw] or [wʔ], under certain conditions (Boersma and Hayes 2001:55–59). For example, /taʔo-en/ is realized as either [taʔ.wen] or [taw.ʔen]. It seems straightforward to represent this in an OT tableau, abstracting away from the actual constraints involved, as illustrated in table 1. We have one input to production, /taʔo-en/, and two outputs, the two winners [taʔ.wen] and [taw.ʔen].

Next let us consider ambiguity, taking Germanic final devoicing as an example. Roughly speaking, we can say that word-final obstruents are realized as [−voiced]. So /læb/ would be realized as [læp]. But /læp/ would also be realized as [læp]. The form [læp] is ambiguous, having two possible underlying forms. This is clearly formally the same problem as optionality in OT: we have one input to comprehension, and two outputs, the two winners /læp/ and /læb/.

It is obvious why SOT has trouble representing optionality and ambiguity (recall from section 1 that SOT as defined by Prince and Smolensky (1993) assumes strict ranking of all constraints). In the cases we have considered, there have been two winners, but each SOT competition has *one* optimal candidate corresponding to one winning output. SOT can, in principle, produce multiple outputs, but only if there are candidates with identical constraint violation profiles, a situation that is extremely rare for a grammar with a realistic number of constraints. However, the CDA was not designed to handle optionality and ambiguity (Tesar and Smolensky 1998:

[3] Ilokano is an Austronesian language, spoken principally in the Philippines, with roughly eight million speakers (data from Ethnologue, http://www.ethnologue.com).

249–251). This means that grammars that model optionality or ambiguity using multiple winners are not learnable with the CDA, as Boersma and Hayes (2001) demonstrate.

The simplest solution to the problem of free variation is to make the constraint hierarchy a partial order instead of a total order (Anttila 1997a,b): in this setting some constraints are tied for their rank in the ordering. The partial ordering can be resolved to varying total orders, and each of the orders produces a different winner. The POT/GLA framework constitutes a probabilistic implementation of this idea, as we will explain in more detail in section 3.

### 2.2 *Gradient Grammaticality*

There is a growing body of evidence showing that grammaticality is a gradient notion, rather than a categorical one (for a review see Schütze 1996). A number of experimental studies demonstrate that speakers can reliably make gradient well-formedness distinctions, in morphology and phonology (Hayes 1997, 2000, Hayes and MacEachern 1998, Keller and Alexopoulou 2001) and in syntax (Bard, Robertson, and Sorace 1996, Cowart 1997, Keller 2000a,b, Keller and Asudeh 2001, McDaniel and Cowart 1999, Sorace 1993a,b, 2000). Gradient well-formedness is clearly a feature of native speakers' knowledge of language and as such should be accounted for by linguistic theory.

SOT, however, is not designed to handle gradient well-formedness: for every input, there is exactly one winning candidate, which is grammatical; all other candidates are ungrammatical. This means that SOT can only model categorical well-formedness judgments (it shares this feature with most other generative theories; see, e.g., Bresnan 2001, Chomsky 1981, 1995, Pollard and Sag 1994). The CDA is designed as a learning algorithm for SOT and hence inherits this limitation; that is, it can only learn grammars that make categorical well-formedness distinctions.

There are two proposals for extensions of OT that can handle gradient grammaticality (Keller 2000b, Müller 1999). Both approaches are based on a distinction between two types of constraints, one that triggers categorical grammaticality, and one that triggers gradient well-formedness. However, neither of these approaches addresses the issues of free variation, ambiguity, and robustness.

### 2.3 *Robustness*

In developing the CDA, Tesar and Smolensky (1998) rely on an important idealization. They assume that the learning algorithm has access to training data that reflect the grammar perfectly (i.e., that are free of erroneous examples). The CDA is guaranteed to converge on the correct grammar only under this idealization.

A real-world language learner, however, has to cope with noise in the training data, such as slips of the tongue or distorted and incomplete utterances. As Boersma and Hayes (2001) show, the CDA does not work well in the face of noisy training data: a single erroneous training example can trigger drastic changes in the learner's grammar, possibly leading to a situation where the whole constraint hierarchy has to be relearned. The GLA is designed to overcome this limitation: it is robust against noise in the training data (i.e., a small proportion of erroneous examples will not affect its learning behavior).

## 3 Probabilistic Optimality Theory and the Gradual Learning Algorithm

Boersma and Hayes (2001) propose a probabilistic variant of Optimality Theory (POT) that is claimed to overcome the problems with SOT discussed in the previous section. It is designed to account for corpus frequencies (thus modeling free variation) and gradient acceptability judgments (thus accounting for degrees of grammaticality). Furthermore, POT is equipped with a learning algorithm that is robust (i.e., that can deal with noise in the training data). The POT framework has been applied in phonology (Boersma 1997, 1998, 2000, Boersma and Hayes 2001, Boersma and Levelt 2000, Hayes 2000, Hayes and MacEachern 1998), morphology (Boersma and Hayes 2001, Hayes 1997), and syntax (Asudeh 2001, Bresnan, Dingare, and Manning 2001, Dingare 2001, Koontz-Garboden 2001).

The POT model stipulates a continuous scale of *constraint strictness.* OT constraints are annotated with numerical strictness values; if a constraint $C_1$ has a higher strictness value than a constraint $C_2$, then $C_1$ outranks $C_2$. Boersma and Hayes (2001) assume *probabilistic constraint evaluation,* which means that at evaluation time a small amount of random noise is added to the strictness value of a constraint. As a consequence, *rerankings* of constraints are possible if the amount of noise added to the strictness values exceeds the distance between the constraints on the strictness scale.

For instance, assume that two constraints $C_1$ and $C_2$ are ranked $C_1 \gg C_2$, selecting the structure $S_1$ as optimal for a given input. Under Boersma and Hayes's (2001) approach, a reranking of $C_1$ and $C_2$ can occur at evaluation time, resulting in the opposite ranking $C_2 \gg C_1$. This reranking might result in an alternative optimal candidate $S_2$. The probability of the reranking that makes $S_2$ optimal depends on the distance between $C_1$ and $C_2$ on the strictness scale (and on the amount of noise added to the strictness values). The reranking probability is assumed to predict the corpus frequency of $S_2$ and thus account for free variation. The more probable the reranking $C_2 \gg C_1$, the higher the corpus frequency of $S_2$; if the rankings $C_1 \gg C_2$ and $C_2 \gg C_1$ are equally probable, then $S_1$ and $S_2$ have the same corpus frequency (i.e., we have a case of true optionality). Furthermore, Boersma and Hayes (2001) assume that corpus frequency and degree of grammaticality are directly related: ''intermediate well-formedness judgments often result from grammatically encodable patterns in the learning data that are rare, but not vanishingly so, the degree of ill-formedness being related monotonically to the rarity of the pattern'' (Boersma and Hayes 2001:73). This means that POT also provides a model of gradient grammaticality (see section 4.4 for a critique of this assumption).

The POT framework comes with its own learning theory in the form of the GLA (Boersma 1998, 2000, Boersma and Hayes 2001). This algorithm is a generalization of Tesar and Smolensky's (1998) CDA: it performs constraint promotion as well as demotion. Note that both the CDA and the GLA assume as training data a corpus of parsed examples; that is, they have access not only to the surface strings, but also to the underlying structures of the training examples.[4]

[4] This in itself is a problematic assumption, but we will grant it for the sake of argument. For criticisms, which have largely not been addressed in the OT community, see Turkel 1994 and Hale and Reiss 1998.

More specifically, the GLA works as follows. It starts with a grammar *G,* in which initially the constraints are ranked arbitrarily (i.e., they have random strictness values). If the GLA encounters a training example *S,* it computes the corresponding structure $S'$ currently generated by the grammar *G.* If *S* and $S'$ are not identical, then learning takes place; the constraint hierarchy of *G* has to be adjusted such that it makes *S* optimal, instead of $S'$. (The example *S* is attested in the training set; hence, it has to win over the unattested competitor $S'$.) In order to achieve this adjustment, the GLA first performs *mark cancellation;* that is, it disregards all constraint violations that are incurred both by *S* and by $S'$. On the remaining uncanceled marks, the algorithm performs the following steps to adjust constraint strictness: (a) it decreases (by a small amount) the strictness values of all constraints that are violated by *S* but not by $S'$; (b) it increases (by a small amount) the strictness values of all constraints that are violated by $S'$ but not by *S.*

This procedure will gradually adjust the strictness values of the constraints in *G,* resulting ultimately in the correct constraint hierarchy (given that enough training data are available). Just like the CDA, the GLA performs constraint reranking, but it does so gradually; one training example is not sufficient to change the ranking of a given constraint, as it only triggers small changes in constraint strictness. This means that the GLA is robust. A small number of incorrect training examples will not disturb the learning process; the effect of the noise is outweighed by the effect of the correct training examples, which can be assumed to form the majority of the training data.

Crucially, Boersma and Hayes (2001) claim that the GLA converges on a *frequency-matching* grammar. If two forms $S_1$ and $S_2$ both occur in the training set, then the resulting grammar will also generate both forms. In particular, the probabilities that the grammar assigns to $S_1$ and $S_2$ will correspond to the frequencies of the two forms in the training data. This means that the GLA offers an account of free variation, and also of gradient grammaticality (under the assumption that corpus frequency and degree of grammaticality are directly related).

## 4 Problems with the Gradual Learning Algorithm

### *4.1 Testing on Unseen Data*

Boersma and Hayes (2001) test the POT/GLA model on three data sets: (a) frequency data for Ilokano reduplication and metathesis, (b) frequency data for Finnish genitive plurals, and (c) acceptability judgment data for the distribution of English light and dark /l/. For each of the data sets, a good model fit is achieved; that is, the algorithm learns a grammar that generates frequency distributions that closely match those in the training data (as shown by a low average error rate).

Achieving a good fit on the training data is a first step in testing a learning algorithm. The next step is to test the algorithm on unseen data. A learning algorithm is useful only if it achieves a low error rate both on the training data and on unseen test data. The parameters of the algorithm are determined using the training data, and then the algorithm is applied to the test data, while holding the parameters constant. Testing on unseen data makes it possible to assess the ability of the algorithm to generalize. Such tests are standard practice in machine learning (e.g., Mitchell 1997) and computational linguistics (e.g., Manning and Schütze 1999). Also, in the literature on models of human language acquisition, testing on unseen data is routinely carried out to validate a proposed learning algorithm (e.g., Gillis, Daelemans, and Durieux 2000, Westermann 1998).[5]

However, no tests on unseen data are reported for the GLA by Boersma and Hayes (2001). The absence of such tests leaves open the possibility that the algorithm *overfits* the data (i.e., that it achieves a good fit on the training set, but is unable to generalize to unseen data). This problem of overfitting is potentially quite serious. In Boersma and Hayes's (2001) model of light versus dark /l/, six free parameters (viz., the strictness values of the six constraints in the model) are used to fit seven data points (viz., the seven mean acceptability ratings that are being modeled). Overfitting seems very likely in this situation.

In the following paragraphs we will briefly discuss how the problem of overfitting could be addressed in the context of a POT-based learning algorithm. First, we will briefly review a set of standard crossvalidation techniques from the machine learning literature (Mitchell 1997).

*Held-out data.* This approach involves randomly splitting the data set into two sets: the training set that is used to estimate the parameters of the model, and the test set that is used to test the model. Then the model fit is computed on both the test set and the training set; a good model fit on the test set indicates that the model is able to generalize to unseen data (i.e., does not overfit the training data). The disadvantage of the held-out data approach is that a fairly large data set has to be used. The test set should be about 10% of the overall data set; if the data set is too small, no meaningful results can be achieved when testing the model.

k-*fold crossvalidation.* This approach is a generalization of the held-out data approach. The data set is randomly partitioned into $k$ subsets. The model is tested on one of these subsets, after having been trained on the remaining $k - 1$ subsets. This procedure is repeated $k$ times such that each subset serves once as the test set and $k - 1$ times as part of the training set. On the basis of the training and testing results, average values for the model fit can be computed. The $k$-fold crossvalidation approach has the advantage of also being applicable to fairly small data sets, as in effect the whole data set is used for testing. In addition, average values are obtained for the model fit on the training and the test data; that is, confidence intervals can be computed. Typically, a value of $k = 10$ is used in the literature.

*Leave one out.* This method is an instance of $k$-fold crossvalidation where $k$ is set to the size of the data set. This means that the model is trained on all items of the training set, leaving out only one item, on which the model is then tested. This procedure is repeated $k$ times and the average model fit is computed. The advantage of the leave-one-out approach is that it is even more suitable for small data sets than standard $k$-fold crossvalidation. An obvious disadvantage is that a large number of training and test runs have to be carried out.

Which of these three tests for overfitting will be chosen for a given learning task largely depends on the amount of data available. The data sets on which Boersma and Hayes (2001) test the GLA are all fairly small: the Ilokano reduplication data set consists of 29 data points; the Finnish plural data set, 44 data points; and the data set for the distribution of English /l/, 7 data

[5] Note that testing on unseen data is unnecessary for Tesar and Smolensky's CDA. As this algorithm presupposes idealized training data (see section 2), the error rate on both training and testing data will be zero.

points (see Boersma and Hayes 2001:(22), (30), and (35), respectively). This means that the only test for overfitting that can be expected to yield reliable results on these data is the leave-one-out procedure. In this approach the GLA would be trained on all data points but one, and the resulting grammar would be tested on its ability to correctly predict this missing data point. This procedure would then be repeated for all data points, and the average model fit computed.

In principle, the number of data points available for training and testing could be increased by testing on tokens (i.e., on corpus instances of a given training example) instead of on types. However, this option is available only for the Finnish plural data, as this is the only phenomenon discussed by Boersma and Hayes (2001) for which actual corpus data are available. For the Ilokano and English data Boersma and Hayes (2001) have to resort to simulating corpus evidence. In the first case they assume that all optional forms are equally distributed in the corpus; in the second case they assume an exponential relationship between degrees of acceptability and corpus frequencies.

## *4.2 Counterexamples*

In this section we will provide two types of counterexamples illustrating that there are acceptability or frequency patterns that the GLA is not able to learn. We will also refer to experimental results and frequency data that instantiate these patterns, showing that they are not just hypothetical counterexamples, but constitute a serious problem for the GLA. These data cover both phonology and syntax, and they include acceptability as well as frequency data.

The first counterexample involves harmonic bounding. Assume two structures $S_1$ and $S_2$ in the same candidate set, which both incur a violation of the constraint $C_1$. The structure $S_2$ incurs an additional violation of the constraint $C_2$, and $S_1$ and $S_2$ incur no other violations (or incur the same violations). Now assume a third structure $S_3$ that only incurs a violation of the constraint $C_3$. Assume further that $S_2$ is less grammatical (or less frequent) than $S_1$. Let $S_3$ be less grammatical (or less frequent) than $S_2$.

This configuration is illustrated in table 2. The GLA is not able to learn such a data set: there is no reranking under which $S_2$ is optimal, as $S_2$ incurs the same violations as $S_1$, plus an additional violation of $C_2$. Hence, $S_1$ will always win over $S_2$, no matter which constraint re-rankings we assume. Under a GLA approach the degree of grammaticality (or frequency) of a

**Table 2**
Data set that the GLA cannot learn (hypothetical frequencies or acceptability scores)

| /input/ | $C_3$ | $C_1$ | $C_2$ | Frequency/Acceptability |
|---|---|---|---|---|
| $S_1$ | | * | | 3 |
| $S_2$ | | * | * | 2 |
| $S_3$ | * | | | 1 |

**Table 3**
Data set that the GLA cannot learn (log-transformed mean acceptability scores for word order in German; Keller 2000b, experiment 10)

| /S, O, V/ | VERB | NOM | PRO | Acceptability |
|---|---|---|---|---|
| O[pro,acc] S[nom] V | | * | | .2412 |
| O[acc] S[pro,nom] V | | * | * | −.0887 |
| V S[pro,nom] O[acc] | * | | | −.1861 |

structure depends on how likely it is for this structure to be optimal. $S_2$ can never be optimal; it is a ‘‘perpetual loser’’ and therefore is predicted to be categorically ungrammatical (or of frequency zero). $S_3$, on the other hand, is not a perpetual loser, as there are rerankings that make it optimal (e.g., $C_1 \gg C_3$ and $C_2 \gg C_3$). This means that a situation where $S_3$ is less grammatical (or less frequent) than $S_2$ cannot be modeled by the GLA.

Configurations such as this one can be found in the experimental literature on gradient grammaticality. An example is provided by Keller’s (2000a,b) study of word order variation in German.[6] Table 3 lists experimentally elicited acceptability scores for subordinate clauses, varying the relative order of the subject NP (S, nominative case), the object NP (O, accusative case), and the verb (V). One of the NPs is pronominalized, as indicated by the feature [pro].

The data in table 3 can be accounted for by a simple set of linear precedence constraints: VERB specifies that the verb has to be in final position, NOM specifies that nominative NPs have to precede nonnominative NPs, and PRO specifies that pronouns have to precede full NPs. Another linear precedence constraint is DAT, requiring that dative NPs precede accusative NPs (this constraint will become relevant later on). This set of constraints provides an intuitive, straightforward account of word order preferences in the German subordinate clause. It is largely uncontroversial in the theoretical literature, which is evidenced by the fact that a number of authors assume essentially the same set of constraints (Choi 1996, Jacobs 1988, Keller 2000a,b, Müller 1999, Uszkoreit 1987).

Under this account the structures in table 3 incur one violation of NOM, a combined violation of NOM and PRO, and one violation of VERB, respectively. The relative acceptability values match the ones in the counterexample in table 2. This means that we have a case of an experimentally attested acceptability pattern that cannot be learned by the GLA.[7] Given the uncontroversial status

[6] Although Boersma and Hayes (2001) do not explicitly claim that the GLA is applicable to syntax, there is no reason to believe that it should not be. The GLA is a learning algorithm for OT, which is not in itself a theory of phonology or morphology. Given that syntactic analyses can be couched in OT (for recent examples see Legendre, Grimshaw, and Vikner 2001, Sells 2001), the GLA should be able to learn syntactic OT grammars. In addition, there has been recent work in syntax that specifically uses Boersma and Hayes’s (2001) POT/GLA model (Asudeh 2001, Bresnan, Dingare, and Manning 2001, Dingare 2001, Koontz-Garboden 2001).

[7] Note that table 3 assumes that all the structures are in the same candidate set (i.e., they compete with each other).

**Table 4**
Data set with cumulative constraint violations (hypothetical frequencies or acceptability scores)

| /input/ | $C_1$ | $C_2$ | Frequency/Acceptability |
|---|---|---|---|
| $S_1$ | | * | 4 |
| $S_2$ | | ** | 3 |
| $S_3$ | | *** | 2 |
| $S_4$ | * | | 1 |

of the word order constraints in this example, we would certainly expect the GLA to be able to learn the corresponding acceptability scores.[8]

A related problem with the GLA concerns effects from cumulative constraint violations (cumulative violations are a special case of harmonic bounding). Consider the constraint set in table 4. Here the winning candidate is $S_1$, incurring a single violation of $C_2$. If a reranking $C_2 \gg C_1$ occurs, then $S_4$, incurring a single violation of $C_1$, will win. However, there is no reranking that can make $S_2$ or $S_3$ optimal, as these candidates have the same violation profile as $S_1$, but incur multiple violations of $C_2$. The structures $S_2$ and $S_3$ are ''perpetual losers'' and are expected to be categorically ungrammatical (or of frequency zero). This means that the GLA predicts that there should be no cumulative effects from multiple constraint violations: all structures that incur more than one violation of a given constraint will be equally ungrammatical (provided they are minimal pairs—i.e., they share the same constraint profile on all other constraints).

Cumulative effects are attested in actual linguistic data. They are not just theoretical constructs, and they therefore pose a real problem for the GLA. We illustrate this point with reference to Guy and Boberg's (1997) frequency data for coronal stop deletion in English (see Guy 1997 for a detailed discussion). The assumption is that the deletion of a coronal stop is governed by the generalized Obligatory Contour Principle (OCP), which can be formulated as *[αF] [αF]: feature sharing with the preceding segment is disallowed. Guy and Boberg (1997) show that the frequency with which the deletion of a coronal stop occurs depends on the number of features that are shared with the preceding segment (see table 5). In other words, they observe a cumulative effect triggered by the generalized OCP: the more OCP violations a structure incurs, the lower the frequency of retention of the coronal stop. This situation can be easily mapped on the cumula-

This is of course an assumption that could be challenged on theoretical grounds. However, in the POT/GLA framework, differences in degree of grammaticality or frequency can *only* be predicted for structures that are in the same candidate set. This means that the data in table 3 are problematic for POT/GLA, even if we drop this assumption.

[8] It is important to note that by means of examples such as the one in table 3, we can only refute the *conjunction* of a given linguistic analysis and a given learning algorithm. Even though the constraint set assumed in our word order example is uncontroversial in the literature, it seems conceivable that an alternative analysis of the word order data could be provided. If this analysis avoids harmonic bounding, then it could make the data learnable for the GLA.

**Table 5**
Preceding segment effect on coronal stop deletion in English (Guy and Boberg 1997, cited in Guy 1997)

| | | Deletion | |
|---|---|---|---|
| Preceding segment | | *N* | % |
| All features shared with target | | | |
| /t,d/ | [+cor, −son, −cont] | – | (categorical absence) |
| Two features shared with target | | | |
| /s,z,ʃ,ʒ/ | [+cor, −son] | 276 | 49 |
| /p,b,k,g/ | [+son, −cont] | 136 | 37 |
| /n/ | [+cor, −cont] | 337 | 46 |
| One feature shared with target | | | |
| /f,v/ | [+son] | 45 | 29 |
| /l/ | [+cor] | 182 | 32 |
| /m,ŋ/ | [+cont] | 9 | 11 |
| No feature shared with target | | | |
| /r/ | – | 86 | 7 |
| vowels | – | – | (nearly categorical retention) |

tive example that we discussed earlier: compare table 4 and table 6 (note that we have converted relative deletion frequencies to relative retention frequencies to illustrate our point). This means that the GLA is not able to learn Guy and Boberg's (1997) frequency data.[9]

Cumulative effects occur not only in frequency data such as those presented by Guy and Boberg (1997), but also in acceptability data, as demonstrated by Keller (2000b) for word order variation in German. Table 7 lists experimentally elicited acceptability scores for permutations of subject (S), object (O), and indirect object (I) in subordinate clauses with ditransitive verbs. This acceptability pattern can be accounted for straightforwardly using the constraints NOM (nominative precedes nonnominative) and DAT (dative precedes accusative) (Choi 1996, Jacobs 1988, Keller 2000a,b, Müller 1999, Uszkoreit 1987).

The word order data in table 7 combine the properties of the counterexamples in tables 2 and 4. On the one hand, we find cumulative effects (as in table 4): the structure I[dat] O[acc] S[nom] V incurs a double violation of NOM and is less acceptable than the structure I[dat] S[nom] O[acc] V, which incurs only one violation of NOM. On the other hand, the data in table 7 provide

[9] Again, it is possible to challenge the assumption that the cases in table 6 should all be in the same competition (see also footnote 7). However, even if they are not, the POT/GLA model makes the wrong prediction, as it would predict that all outputs are equally grammatical, if they are the sole winners of their competitions. This is contrary to the data presented by Guy and Boberg (1997). In other words, the POT/GLA model can predict the differing frequencies of the various pre-coronal segments only if they are in the same competition; if they are in the same competition, then the grammar is not learnable. See Guy 1997 for a more detailed discussion of these data and their implications for various OT-based models of corpus frequencies.

**Table 6**
Data set with cumulative constraint violations (relative frequencies for coronal stop retention, Guy and Boberg 1997)

| Segment preceding coronal (in input) | *[αF] [αF] | Frequency |
|---|---|---|
| /t,d/ | *** | 0 |
| /s,z,ʃ,ʒ/ | ** | 51 |
| /p,b,k,g/ | ** | 63 |
| /n/ | ** | 54 |
| /f,v/ | * | 71 |
| /l/ | * | 68 |
| /m,ŋ/ | * | 89 |
| /r/ | | 93 |
| vowels | | 100 |

another example of the problems with harmonic bounding that the GLA faces. The structure O[acc] S[nom] I[dat] V incurs a combined violation of Nom and Dat, which means that it will always lose against I[dat] S[nom] O[acc] V or S[nom] O[acc] I[dat] V, the structures that incur only one violation of Nom and Dat, respectively. This means that O[acc] S[nom] I[dat] V is a ''perpetual loser'': it can never be optimal and thus is predicted to be maximally ungrammatical by POT. However, as table 7 shows, there are several structures in this candidate set that are more ungrammatical than O[acc] S[nom] I[dat] V.

Neither the cumulativity effect nor the harmonic bounding effect can be accommodated by

**Table 7**
Data set with cumulative constraint violations (log-transformed mean acceptability scores for word order in German; Keller 2000b, experiment 6)

| /S, O, I, V/ | Nom | Dat | Acceptability |
|---|---|---|---|
| O[acc] I[dat] S[nom] V | ** | * | −.2736 |
| I[dat] O[acc] S[nom] V | ** | | −.2667 |
| O[acc] S[nom] I[dat] V | * | * | −.2038 |
| I[dat] S[nom] O[acc] V | * | | −.0716 |
| S[nom] O[acc] I[dat] V | | * | .0994 |
| S[nom] I[dat] O[acc] V | | | .2083 |

Boersma and Hayes's (2001) model, which means that the GLA is unable to learn the data set in table 7.

## 4.3 Formal Properties

Boersma and Hayes (2001) fail to provide a formal proof of correctness for the GLA, which means that it is not clear that the GLA always generates a correct set of strictness values if supplied with adequate training data. It is not trivial to show the correctness of the GLA, as it is part of a class of possible learning algorithms for POT, not all of which are suitable for learning frequency data. An example is the Minimal Gradual Learning Algorithm, a variant of the GLA originally proposed by Boersma (1997), which Boersma (1998) later showed to be incorrect.

Boersma (2000:517–518) briefly discusses the correctness of the GLA and refers to Boersma 1998. In Boersma 1998, however, only a sketch of a proof is given and the author concedes that "[w]e have made plausible, though not yet rigorously proved, that the maximal symmetrized gradual learning algorithm [the GLA] is capable of learning any stochastically evaluating OT grammar" (Boersma 1998:345). Hence, a rigorous proof of the correctness of the GLA has yet to be provided.[10]

Another problem is that the convergence properties of the GLA are unknown. This leaves open the possibility that there are data sets on which the GLA will not converge or not produce a meaningful set of constraint rankings. Convergence is a crucial property of a learning algorithm that should be investigated formally. Boersma and Hayes (2001) fail to provide the relevant proof.

In section 4.2 we presented counterexamples that the GLA cannot learn. In addition, the GLA never stops trying to learn these examples; that is, it fails to converge on data sets such as the ones in tables 2 and 4. We will illustrate this point with reference to cumulative constraint violations. It is sufficient to consider the training examples $S_1$, $S_2$, and $S_4$ in table 4.

Assume that the learner encounters the example $S_1$. The probabilistic evaluation component will produce either the constraint ordering $C_1 \gg C_2$ or the ordering $C_2 \gg C_1$. If the ordering is $C_1 \gg C_2$, then no changes in strictness will occur, as the training example $S_1$ is already optimal. If the ordering is $C_2 \gg C_1$, then the GLA will compare $S_1$ with the winning competitor $S_4$ and will decrease the strictness of $C_2$ (violated by the training example $S_1$) and increase the strictness of $C_1$ (violated by the competitor $S_4$). No change of strictness is triggered by $S_2$, as $S_1$ wins over $S_2$. The learning behavior for all three training examples is summarized in table 8, where the notation $C_n+$ denotes an increase in the strictness of $C_n$, and $C_n-$ denotes a decrease.

Table 8 makes clear why the GLA fails to converge on a training set that contains the examples $S_1$, $S_2$, and $S_4$. Assume that we start off with equal strictness values for $C_1$ and $C_2$. As $S_1$ is the most frequent training example, the situation that occurs most often is (b). Situation (e) occurs less often, owing to the lower frequency of $S_4$. This means that the strictness values of

[10] Note also that the sketch of a proof in Boersma 1998 and Boersma 2000 makes two simplifying assumptions: (a) candidate sets are finite, and (b) constraints can only be violated once. This means that the sketch does not extend straightforwardly to a full proof.

**Table 8**
Learning behavior of the GLA on the data set in table 4 (examples $S_1$, $S_2$, $S_4$)

| Example | Frequency | Probability evaluation | Change in strictness |
|---|---|---|---|
| $S_1$ | 4 | (a) $C_1 \gg C_2$<br>(b) $C_2 \gg C_1$ | no change<br>$C_1+$, $C_2-$ |
| $S_2$ | 3 | (c) $C_1 \gg C_2$<br>(d) $C_2 \gg C_1$ | $C_2-$<br>$C_1+$, $C_2-$ |
| $S_4$ | 1 | (e) $C_1 \gg C_2$<br>(f) $C_2 \gg C_1$ | $C_1-$, $C_2+$<br>no change |

$C_1$ and $C_2$ drift apart, leading to the ranking $C_1 \gg C_2$. In the limit the GLA will find the optimal distance between the strictness values of $C_1$ and $C_2$; that is, the distance that corresponds to the relative frequency of $S_1$ and $S_4$.

At the same time, however, the training example $S_2$ will continue to decrease the strictness value of $C_2$, no matter whether the probabilistic evaluation leads to the ranking $C_1 \gg C_2$ (situation (c)) or to the ranking $C_2 \gg C_1$ (situation (d)).[11] This decrease cannot be compensated for by the training example $S_4$, which increases the strictness of $C_2$, but occurs less frequently than $S_2$. The consequence is a continuous downdrift of $C_2$, which also triggers a downdrift of $C_1$, as the training examples $S_1$ and $S_4$ cause the GLA to try to find the optimal distance between $C_1$ and $C_2$, on the basis of the relative frequencies of $S_1$ and $S_4$. This means that the GLA will keep on reducing the strictness values of $C_1$ and $C_2$, no matter how long training continues.

The failure of the GLA to converge on training sets like the one in table 4 (and the one in table 2) can also be verified empirically using Praat, a software package that implements the GLA (Boersma 1999). When confronted with a training set that contains the configuration in table 4, Praat will produce a continuous downdrift of the strictness values of $C_1$ and $C_2$, as described above, confirming the GLA's failure to converge on such data sets.

### *4.4 Gradience and Frequency*

Boersma and Hayes (2001:73) assume that ''intermediate well-formedness judgments often result from grammatically encodable patterns in the learning data that are rare, but not vanishingly so, the degree of ill-formedness being related monotonically to the rarity of the pattern.'' Their assumption of a direct relationship between well-formedness and frequency is further witnessed by equations they provide relating the two (Boersma and Hayes 2001:82). However, the assumption that gradient grammaticality and corpus frequency are monotonically related and therefore

[11] A note on situation (c): The GLA performs mark cancellation before it adjusts strictness values (see section 3). This means that one $C_2$ violation incurred both by $S_1$ and by $S_2$ will be canceled, leaving one $C_2$ violation at $S_2$ and none at $S_1$. This situation then leads to demotion of $C_2$, as it is violated by the loser $S_2$, but not by the winner $S_1$.

can be treated in the same probabilistic model is far from uncontroversial. This topic has received considerable coverage in the computational linguistics and corpus linguistics literature.[12]

For instance, Keller (2000b) argues that the degree of grammaticality of a structure and its frequency of occurrence in a corpus are two distinct concepts and cannot both be modeled in the same probabilistic framework (as Boersma and Hayes propose). This argument is based on data sparseness: because a language consists of an infinite set of structures, there will always be structures that are grammatical, but occur infrequently (or fail to occur at all) in a finite corpus. This means that a probabilistic model that is trained on corpus frequencies cannot also be expected to account for gradient grammaticality: the absence of a given structure from a corpus cannot serve as evidence that it is ungrammatical.

A related point is put forward by Abney (1996), who states that ''[w]e must also distinguish degrees of grammaticality, and indeed, global goodness, from the probability of producing a sentence. Measures of goodness and probability are mathematically similar enhancements to algebraic grammars, but goodness alone does not determine probability. For example, for an infinite language, probability must ultimately decrease with length, though arbitrarily long sentences may be perfectly good'' (Abney 1996:14).[13] A related point is made by Culy (1998), who argues that the frequency distribution of a construction does not bear on the question of whether it is grammatical or not.

Evidence for Abney's (1996) and Culy's (1998) claims can be found in the psycholinguistic literature. A number of corpus studies have investigated verb subcategorization frequencies, that is, the frequency with which a verb occurs with a given subcategorization frame in a corpus (Lapata, Keller, and Schulte im Walde 2001, Merlo 1994, Roland and Jurafsky 1998). As an example, consider the verb *realize,* which allows both an NP and a sentence frame.

(1) a. The athlete realized her goals.
  b. The athlete realized her goals were out of reach.

It can be shown that the subcategorization frequencies of a verb influence how the verb is processed. In the case of locally ambiguous input (such as (1) up to *her goals*), the human sentence processor will prefer the reading that matches the verb frame with the highest corpus frequency. In example (1) this would mean that the processor prefers the S reading for *realize,* given that *realize* occurs more frequently with the S frame (as indicated by Lapata, Keller, and Schulte im Walde's (2001) frame frequency data for the British National Corpus).[14]

[12] While it is true that Boersma and Hayes (2001) claim only that intermediate well-formedness ''often'' results from rare grammatically encodable events, meaning that there can presumably be other factors giving rise to gradient grammaticality, their solution for these putatively often-arising cases is a monotonic relationship between gradience and frequency; it is with the latter claim that we take issue.

[13] An example for Abney's (1996) point about length and probability is provided by recursive rules in a probabilistic context-free grammar. If the length of a sentence is increased by adding material using a recursive rule (e.g., by adding an adjective using the rule $N' \rightarrow Adj\ N'$), then the probability of the sentence will necessarily decrease: in a probabilistic context-free grammar, the probability of a sentence is computed as the product of the probabilities of all the rules applied in generating the sentence.

[14] In the example at hand, the disambiguation preference of the human parser is also influenced by other factors, including the plausibility of the postverbal NP as an object of the verb (Pickering, Traxler, and Crocker 2000) and the tendency of the verb to omit the complementizer *that* (Trueswell, Tanenhaus, and Kello 1993).

While this type of frequency information has been shown to influence the online behavior of the human sentence processor, it is not standardly assumed that it has an effect on grammaticality. Few linguists will want to assume that a verb is less grammatical with a certain subcategorization frame just because this frame is less frequent in the corpus. In our example this assumption would mean that sentences involving *realize* with an NP complement are less grammatical than sentences involving *realize* with an S complement, clearly a counterintuitive result.

In our view the right way of conceptualizing the difference between frequency and gradient grammaticality follows from basic assumptions about competence and performance advocated by Chomsky (1965, 1981, 1995) and many others (for a review see Schütze 1996). The frequency of occurrence of a structure has to do with how the speaker processes this structure and is therefore a performance phenomenon. The degree of grammaticality of a structure, on the other hand, has to do with the speaker's knowledge of language and is therefore part of linguistic competence.

The model that Boersma and Hayes (2001) propose departs from these standard assumptions, a fact that the authors fail to comment on. The key difference in their approach lies in modeling frequency in a competence grammar: their model assumes that in cases of optionality the grammar not only delivers the options, but also predicts their frequency of occurrence.[15] However, if the grammar is a specification of linguistic competence, then there will be many performance factors affecting the *observed* occurrences of a structure generated by the grammar. These include processing factors (e.g., constraints on speech perception and articulation), general cognitive factors (e.g., memory limitations and fatigue), and extralinguistic factors (e.g., speech style and politeness). In fact, given the competence/performance distinction, a grammar that predicts corpus frequencies is almost *guaranteed* to be incorrect, because the frequencies produced by the grammar (although they match those in the corpus) will be affected by performance considerations and will fail to match the corpus frequencies once these performance factors are taken into account.

But suppose we were to simply give up the competence/performance distinction and put all relevant performance factors in the grammar. Then the grammar could predict actual frequencies, because there are no further factors affecting its outputs. Thus, all constraints on perception, articulation, memory, fatigue, style, and politeness interact with grammatical constraints. What would this mean for the claims of OT with respect to factorial typology, lexicon optimization, lack of rule conspiracies, and so on?

For factorial typology, for instance, we would arrive at predictions that are clearly counterintuitive. Surely speakers with distinct native languages have cognitive abilities in common and these cannot be reranked to yield their different languages. It is probably safe to assume that the difference between Swedish and Norwegian does not arise because of memory differences between the speakers of Swedish and Norwegian, for example.

Or consider lexicon optimization: the underlying form (i.e., input to GEN) that is lexically stored for a given morpheme is the one that is most harmonic across grammatical contexts (Prince

[15] This assumption is shared by Anttila (1997a,b) and Bresnan, Dingare, and Manning (2001).

and Smolensky 1993). Suppose that there are some performance constraints in the constraint hierarchy. Alternatively, suppose that some performance factors are modeled by constraint reranking (Boersma 2000). In either case there will be more distinct outputs to consider (e.g., the drunk output is likely different from the polite output). Since lexicon optimization considers inputs for the *same* output, and there are more different outputs to consider, this will lead to a spurious proliferation of lexical items. In effect, there would not only be performance-related outputs, there would also be performance-related inputs, stored lexically.

These examples from factorial typology and lexicon optimization show that OT in particular *needs* the competence/performance distinction just to make sense. It is therefore not possible for the GLA model to give this distinction up entirely, and thus its claims about predicting frequencies are erroneous.

## 5 Conclusion

The picture we end up with is the following. We have two versions of Optimality Theory—Standard OT (SOT) and Probabilistic OT (POT)—and learning algorithms for the kinds of grammars that each specifies—the CDA and the GLA, respectively. SOT with its CDA has proofs of correctness and convergence. But this model has no account of optionality, ambiguity, or gradient grammaticality: the grammars cannot represent these phenomena satisfactorily and the learning algorithm cannot learn minimally modified OT grammars (based on partial constraint hierarchies; see Anttila 1997a,b) that can represent these phenomena. Also, the CDA is not robust; that is, it cannot deal with errors in the training data.

POT and the GLA offer a treatment of optionality and ambiguity, as demonstrated for phonology and morphology by Boersma and Hayes (2001) and others, and for syntax by Asudeh (2001), Bresnan, Dingare, and Manning (2001), Dingare (2001), and Koontz-Garboden (2001).[16] In addition, the GLA is a robust learning algorithm, thus offering a crucial advantage over the CDA. However, claims for its empirical adequacy are premature, as its learning behavior has not been verified using tests on unseen data (see section 4.1). Also, there are no formal proofs of the correctness and convergence of the CDA (see section 4.3). In fact, in section 4.2 we presented counterexamples that the GLA cannot learn, showing that it is incorrect (it cannot learn an example it should learn) and fails to converge (it also never stops trying).

While the POT/GLA model offers a promising approach to optionality and ambiguity in OT, its treatment of gradient grammaticality is conceptually flawed, as are its predictions of corpus frequencies. This was demonstrated in section 4.4 on the basis of standard assumptions about competence and performance.

[16] Asudeh's (2001) analysis is couched in POT with harmonic alignment of prominence scales (Aissen 1999, Prince and Smolensky 1993). Strictly speaking, the GLA would have to be extended to cope with harmonic alignment in a manner that comports with the theoretical understanding of this mechanism. Therefore, Asudeh (2001) offers a treatment of optionality and ambiguity using POT, but without adopting the GLA.

## References


Abney, Steven. 1996. Statistical methods and linguistics. In *The balancing act: Combining symbolic and statistical approaches to language,* ed. by Judith Klavans and Philip Resnik, 1–26. Cambridge, Mass.: MIT Press.

Aissen, Judith. 1999. Markedness and subject choice in Optimality Theory. *Natural Language & Linguistic Theory* 17:673–711.

Anttila, Arto. 1997a. Deriving variation from grammar: A study of Finnish genitives. In *Variation, change, and phonological theory,* ed. by Frans Hinskens, Roeland van Hout, and W. Leo Wetzels, 35–68. Amsterdam: John Benjamins.

Anttila, Arto. 1997b. Variation in Finnish phonology and morphology. Doctoral dissertation, Stanford University, Stanford, Calif.

Asudeh, Ash. 2001. Linking, optionality, and ambiguity in Marathi. In *Formal and empirical issues in optimality-theoretic syntax,* ed. by Peter Sells, 257–312. Stanford, Calif.: CSLI Publications.

Bard, Ellen Gurman, Dan Robertson, and Antonella Sorace. 1996. Magnitude estimation of linguistic acceptability. *Language* 72:32–68.

Boersma, Paul. 1997. How we learn variation, optionality, and probability. In *Proceedings of the Institute of Phonetic Sciences 21,* ed. by R. J. J. H. van Son, 43–58. Amsterdam: University of Amsterdam, Institute of Phonetic Sciences.

Boersma, Paul. 1998. *Functional phonology: Formalizing the interactions between articulatory and perceptual drives.* The Hague: Holland Academic Graphics.

Boersma, Paul. 1999. Optimality-theoretic learning in the Praat program. In *Proceedings of the Institute of Phonetic Sciences 23,* 17–35. Amsterdam: University of Amsterdam, Institute of Phonetic Sciences.

Boersma, Paul. 2000. Learning a grammar in Functional Phonology. In *Optimality Theory: Phonology, syntax, and acquisition,* ed. by Joost Dekkers, Frank van der Leeuw, and Jeroen van de Weijer, 465–523. Oxford: Oxford University Press.

Boersma, Paul, and Bruce Hayes. 2001. Empirical tests of the Gradual Learning Algorithm. *Linguistic Inquiry* 32:45–86.

Boersma, Paul, and Clara Levelt. 2000. Gradual constraint-ranking learning algorithm predicts acquisition order. In *Proceedings of the 30th Child Language Research Forum,* ed. by Eve V. Clark, 229–237. Stanford, Calif.: CSLI Publications.

Bresnan, Joan. 2001. *Lexical-Functional Syntax.* Oxford: Blackwell.

Bresnan, Joan, Shipra Dingare, and Christopher D. Manning. 2001. Soft constraints mirror hard constraints: Voice and person in English and Lummi. In *Proceedings of the LFG 2001 Conference,* ed. by Miriam Butt and Tracy Holloway King. Stanford, Calif.: CSLI Publications Online. http://cslipublications.stanford.edu/hand/miscpubsonline.html.

Choi, Hye-Won. 1996. Optimizing structure in context: Scrambling and information structure. Doctoral dissertation, Stanford University, Stanford, Calif.

Chomsky, Noam. 1965. *Aspects of the theory of syntax.* Cambridge, Mass.: MIT Press.

Chomsky, Noam. 1981. *Lectures on government and binding.* Dordrecht: Foris.

Chomsky, Noam. 1995. *The Minimalist Program.* Cambridge, Mass.: MIT Press.

Cowart, Wayne. 1997. *Experimental syntax: Applying objective methods to sentence judgments.* Thousand Oaks, Calif.: Sage Publications.

Culy, Christopher. 1998. Statistical distribution and the grammatical/ungrammatical distinction. *Grammars* 1:1–19.

Dingare, Shipra. 2001. The effect of feature hierarchies on frequencies of passivization in English. Master's thesis, Stanford University, Stanford, Calif.

Gillis, Steven, Walter Daelemans, and Gert Durieux. 2000. A comparison of natural and machine learning of stress. In *Models of language acquisition: Inductive and deductive approaches,* ed. by Peter Broeder and Jaap Murre, 76–99. Oxford: Oxford University Press.

Guy, Gregory R. 1997. Violable is variable: Optimality Theory and linguistic variation. *Language Variation and Change* 9:333–347.

Guy, Gregory R., and Charles Boberg. 1997. Inherent variability and the Obligatory Contour Principle. *Language Variation and Change* 9:149–164.

Hale, Mark, and Charles Reiss. 1998. Formal and empirical arguments concerning phonological acquisition. *Linguistic Inquiry* 29:656–683.

Hayes, Bruce. 1997. Gradient well-formedness in Optimality Theory. Handout, Department of Linguistics, University of California, Los Angeles.

Hayes, Bruce. 2000. Gradient well-formedness in Optimality Theory. In *Optimality Theory: Phonology, syntax, and acquisition,* ed. by Joost Dekkers, Frank van der Leeuw, and Jeroen van de Weijer, 88–120. Oxford: Oxford University Press.

Hayes, Bruce, and May Abad. 1989. Reduplication and syllabification in Ilokano. *Lingua* 77:331–374.

Hayes, Bruce, and Margaret MacEachern. 1998. Folk verse form in English. *Language* 74:473–507.

Jacobs, Joachim. 1988. Probleme der freien Wortstellung im Deutschen. In *Sprache und Pragmatik,* ed. by Inger Rosengren, 8–37. Working Papers 5. Lund: Lund University, Department of German.

Keller, Frank. 2000a. Evaluating competition-based models of word order. In *Proceedings of the 22nd Annual Conference of the Cognitive Science Society,* ed. by Lila R. Gleitman and Aravind K. Joshi, 747–752. Mahwah, N.J.: Lawrence Erlbaum.

Keller, Frank. 2000b. Gradience in grammar: Experimental and computational aspects of degrees of grammaticality. Doctoral dissertation, University of Edinburgh.

Keller, Frank, and Theodora Alexopoulou. 2001. Phonology competes with syntax: Experimental evidence for the interaction of word order and accent placement in the realization of information structure. *Cognition* 79:301–372.

Keller, Frank, and Ash Asudeh. 2001. Constraints on linguistic coreference: Structural vs. pragmatic factors. In *Proceedings of the 23rd Annual Conference of the Cognitive Science Society,* ed. by Johanna D. Moore and Keith Stenning, 483–488. Mahwah, N.J.: Lawrence Erlbaum.

Koontz-Garboden, Andrew. 2001. A stochastic OT approach to word order variation in Korlai Portuguese. Paper presented at the 37th Regional Meeting of the Chicago Linguistic Society.

Lapata, Maria, Frank Keller, and Sabine Schulte im Walde. 2001. Verb frame frequency as a predictor of verb bias. *Journal of Psycholinguistic Research* 30:419–435.

Legendre, Géraldine, Jane Grimshaw, and Sten Vikner, eds. 2001. *Optimality-theoretic syntax.* Cambridge, Mass.: MIT Press.

Manning, Christopher D., and Hinrich Schütze. 1999. *Foundations of statistical natural language processing.* Cambridge, Mass.: MIT Press.

McDaniel, Dana, and Wayne Cowart. 1999. Experimental evidence of a minimalist account of English resumptive pronouns. *Cognition* 70:B15–B24.

Merlo, Paola. 1994. A corpus-based analysis of verb continuation frequencies for syntactic processing. *Journal of Psycholinguistic Research* 23:435–457.

Mitchell, Tom M. 1997. *Machine learning.* New York: McGraw-Hill.

Müller, Gereon. 1999. Optimality, markedness, and word order in German. *Linguistics* 37:777–818.

Pickering, Martin J., Matthew J. Traxler, and Matthew W. Crocker. 2000. Ambiguity resolution in sentence processing: Evidence against frequency-based accounts. *Journal of Memory and Language* 43: 447–475.

Pollard, Carl, and Ivan A. Sag. 1994. *Head-Driven Phrase Structure Grammar.* Stanford, Calif.: CSLI Publications, and Chicago: University of Chicago Press.

Prince, Alan, and Paul Smolensky. 1993. Optimality Theory: Constraint interaction in generative grammar. Technical report 2. New Brunswick, N.J.: Rutgers University, Center for Cognitive Science.

Pulleyblank, Douglas, and William J. Turkel. 2000. Learning phonology: Genetic algorithms and Yoruba tongue-root harmony. In *Optimality Theory: Phonology, syntax, and acquisition,* ed. by Joost Dekkers, Frank van der Leeuw, and Jeroen van de Weijer, 554–591. Oxford: Oxford University Press.

Roland, Douglas, and Daniel Jurafsky. 1998. How verb subcategorization frequencies are affected by corpus choice. In *Proceedings of the 17th International Conference on Computational Linguistics and 36th Annual Meeting of the Association for Computational Linguistics,* 1122–1128. Montréal.

Schütze, Carson T. 1996. *The empirical base of linguistics: Grammaticality judgments and linguistic methodology.* Chicago: University of Chicago Press.

Sells, Peter, ed. 2001. *Formal and empirical issues in optimality-theoretic syntax.* Stanford, Calif.: CSLI Publications.

Sorace, Antonella. 1993a. Incomplete vs. divergent representations of unaccusativity in non-native grammars of Italian. *Second Language Research* 9:22–47.

Sorace, Antonella. 1993b. Unaccusativity and auxiliary choice in non-native grammars of Italian and French: Asymmetries and predictable indeterminacy. *Journal of French Language Studies* 3:71–93.

Sorace, Antonella. 2000. Gradients in auxiliary selection with intransitive verbs. *Language* 76:859–890.

Tesar, Bruce, and Paul Smolensky. 1996. Learnability in Optimality Theory (long version). Technical report JHU-CogSci-96-4. Baltimore, Md.: Johns Hopkins University, Department of Cognitive Science.

Tesar, Bruce, and Paul Smolensky. 1998. Learnability in Optimality Theory. *Linguistic Inquiry* 29:229–268.

Tesar, Bruce, and Paul Smolensky. 2000. *Learnability in Optimality Theory.* Cambridge, Mass.: MIT Press.

Trueswell, John C., Michael K. Tanenhaus, and Christopher Kello. 1993. Verb-specific constraints in sentence processing: Separating effects of lexical preference from garden-paths. *Journal of Experimental Psychology: Learning, Memory, and Cognition* 19:528–553.

Turkel, William J. 1994. The acquisition of optimality theoretic systems. Ms., University of British Columbia, Vancouver. Rutgers Optimality Archive 11. http://roa.rutgers.edu

Uszkoreit, Hans. 1987. *Word order and constituent structure in German.* Stanford, Calif.: CSLI Publications.

Westermann, Gert. 1998. Emergent modularity and U-shaped learning in a constructivist neural network learning the English past tense. In *Proceedings of the 20th Annual Conference of the Cognitive Science Society,* ed. by Morton A. Gernsbacher and Sharon J. Derry, 1130–1135. Mahwah, N.J.: Lawrence Erlbaum.

*(Keller)*
*Department of Computational Linguistics*
*Saarland University*
*PO Box 15 11 50*
*66041 Saarbrücken*
*Germany*

*keller@coli.uni-sb.de*

*(Asudeh)*
*Linguistics Department*
*Stanford University*
*Margaret Jacks Hall*
*Building 460*
*Stanford, California 94305-2150*

*asudeh@csli.stanford.edu*