Stochastic OT as a model of constraint interaction

ELENA MASLOVA

1 Introduction[1]

Stochastic OT is a set of specific assumptions about the mechanism of interaction between universal linguistic constraints: the choice of optimal candidates is determined by the rankings of constraint weights, but the rankings can be "perturbed" by a normally distributed "evaluation noise" (Boersma and Hayes 2001); the combined weight of a set of cooperating constraints is thus assumed to equal the maximum of their own weights. This implies that the combined effect of a set of cooperating constraints is only slightly stronger than the effect of the strongest constraint in this set. An interesting question is whether this hypothesis is correct, i.e. whether (and to which extent) it accurately predicts the universal properties of grammatical variation. Whereas this paper does not presume to answer this question, it provides necessary computational tools and defines a minimal condition on StOT grammars under which the StOT assumptions on the mechanism of constraint interaction can be tested against cross-linguistic data.

Section 2 examines the general properties of StOT grammars viewed as parametric statistical models. Knowledge of these properties considerably facilitates all StOT-related computations (Section 3) and will, I hope, be useful independently of the theoretical agenda of this paper. Section 4 returns to the problem of typological predictions entailed by StOT grammars and shows that they can but need not be determined by the specific properties of the StOT mechanism of constraint interaction, depending on the configuration of constraint sets. Obviously, the plausibility of this mechanism can be corroborated or falsified by cross-linguistic evidence only insofar as StOT grammars do entail StOT-specific linguistic predictions. Finally, Section 5 establishes a minimal condition that must be satisfied by a StOT grammar in order to have this property and the corresponding criterion of linguistic plausibility of the StOT mechanism of constraint interaction.

[1] This paper has its origin in discussions I have had with Joan Bresnan over the last couple of years and would hardly ever be written without her interest and encouragement.

2 Stochastic OT as a parametric statistical model

A StOT model $\mathcal{M}$ can be represented as a pair <**C**, μ>, where **C** is a vector of **constraints** and μ is a vector of **constraint strengths**. Constraints can be represented as *n*-by-*m* Boolean arrays, where *n* is the number of environments (input conditions) described by the model, and *m*, the number of competing candidates:

(1) $C_l(i, j) = 1$ if and only if C_l is violated by the j^{th} candidate in the i^{th} environment.

Since constraints are assumed to be universal, the constraint set can be interpreted as a universal model of (some domain of) grammatical variation. **Constraint weights** W_i are independent normally distributed random variables W_i with mean values μ_i and identical standard deviations $\sigma_i = \nu$, interpreted as evaluation noise. Constraint strengths μ are parameters of the model, and their values are language-specific. A vector of values w_i of variables W_i determines a ranking of constraints from **C**: if $w_k > w_l$, then C_k is ranked higher than C_l, and a ranking determines the choice of optimal candidate according to the standard OT evaluation procedure. A StOT model can be called **complete** if any ranking of constraints uniquely determines the optimal candidate for each environment. Obviously, a model is incomplete if and only if there exist at least two candidates that violate exactly the same set of constraints in at least one environment; in what follows, I assume that a StOT model must be complete.

Under these assumptions, the probability of a candidate being evaluated as optimal obviously does not depend on the absolute values of constraint strengths, but only on the differences between them. Accordingly, an *r*-constraint StOT model has *r*–1 **degrees of freedom**. In what follows, it will be convenient to normalize any StOT model by setting the value of one constraint strength to 0 and the noise value to 1; in what follows, the normalizing constraint is denoted as C_0 (i.e. $\mu_0 = 0$).

The probability p_{ij} of candidate O_j occurring in a certain environment E_j can be represented as a function of language-specific constraint strengths μ under the universal conditions determined by the constraint set:

(2) $p_{ij} = \pi_{\mathrm{C}}(\mu)$

Since the constraint weights are independent normally distributed random variables, their joint probability distribution function is the product of *r* normal distribution functions:

(3) $$g(w_0,...,w_{r-1}) = \prod_{k=0}^{r-1} \phi(w_k;\mu_k,1)$$

Here $\phi(x;\ \mu,\ \sigma^2)$ denotes, as usual, the normal probability density function with mean value μ and standard deviation σ evaluated at x. The probability of each ranking ρ can be obtained by integration of this function over the corresponding domain:

(4) $$p(\rho) = \int_{\rho} g(w_0,..., w_{r-1})dw_0...dw_{r-1}$$

Then, p_{ij} is the sum of probabilities of all rankings that select O_j as optimal in environment E_i. However, this direct approach can hardly be effective for actual computations, since it would require numerical integration over r variables (where r is the number of constraints). Presumably for this reason, the probability distributions predicted by StOT models are usually evaluated by running a computer simulation of the StOT grammar and counting the frequencies of different candidates in the generated “corpus”. The goal of this section is to show that the function $\pi_C(\mu)$ can be considerably simplified, so that the computation of p_{ij} is reduced to integration over $m – 1$ variables (where m is the number of competing candidates); this would make the StOT-related computations simpler and more effective than simulations. Informally, the simplification in question is possible because the optimal candidate is determined only by the relation between **maximal** weights of constraints violated by different candidates. I will describe the proposed approach to computation of p_{ij} for the case of two competing candidates (which is not only the simplest one, but also the most common one in actual linguistic practice) and then outline its extension to the general case of competition between m candidates. In the rest of this section, I consider the distribution of candidates in one fixed environment; thus, the index specifying the environment is dropped in the notations, and p_i denotes the probability of the i^{th} candidate.

In the case of two candidates, the set of relevant constraints includes only those constraints that are violated by exactly one of the candidates (if the constraint is violated by both candidates, it is irrelevant). Let $\mathbf{C}_i$ be the set of constraints violated by candidate O_i and $\mathbf{W}_i$, the corresponding set of constraint weights. Let $M_i = \mathbf{max}(\boldsymbol{W}_i)$. The maximum of several independent normally distributed random variables is a random variable with the following cumulative distribution function (simply put, this equation says that the maximum of several constraint weights is less than *m* if and only if each of these constraint weights is less than *m*):

$$(5)\ F_{M_i}(m) = P(M_i \le m) = \prod_{W_k \in \mathbf{W}_i} P(W_k \le m) = \prod_{W_k \in \mathbf{W}} \Phi(m; \mu_k)$$

Here and below, $\Phi(x; \mu)$ denotes a cumulative normal distribution function with mean value μ and variance 1 evaluated at x. The corresponding probability density function is:

$$(6)\ f_{M_i}(m) = F'_{M_i}(m) = \sum_{W_l \in \mathbf{W}_i} \left[\phi(m; \mu_l) \prod_{W_k \in \mathbf{W}_i, k \neq l} \Phi(m; \mu_k) \right]$$

The output of the competition is determined by the difference between M_i and M_j (where O_j is the other candidate). Setting $R = M_j - M_i$, the probability of O_i can be represented as the cumulative distribution function of R evaluated at zero (i.e. the probability that M_j is greater than M_i):

$$(7)\ p_{ji} = P(R_{ji} < 0) = F_{R_{ji}}(0) =$$

$$= \int_{-\infty}^{0} \int_{-\infty}^{\infty} f_{M_{jl}}(t + r) f_{M_{ji}}(t) dtdr = \int_{-\infty}^{\infty} f_{M_{jl}}(t) F_{M_{ji}}(t) dt$$

Substituting (5) and (6) in (7), we get the following expression for p_i as a function of constraint strengths:

$$(8)\ \ p_i = \sum_{W_t \notin \mathbf{W}_i} \int_{-\infty}^{\infty} \phi(m; \mu_t) \prod_{k \neq t} \Phi(m; \mu_k) dm$$

Simply put, we calculate the probability that some C_t violated by the **other** candidate outranks all other constraints, and then sum up these probabilities for all such constraints; the computation of p_i is thereby reduced to numerical integration over one variable.[2] Note that, for the purposes of actual computations, the integral boundaries can be safely defined as $\mu_t \pm 5\nu$ due to the properties of normal distribution. In the simplest case of a competition determined by interaction of exactly **two** constraints, R can be defined

[2] Since the normal probability density function and cumulative distribution function are standard functions, implementation of this formula should be a relatively simple task.

as the difference between two independent normally distributed variables (W_0 and W_1) and thus itself a normally distributed variable with mean value $\mu_i - \mu_j$ and variance $\sigma^2 = 2\nu^2 = 2$ (assuming, without loss of generality, that O_i violates C_i) Accordingly, equation (8) can be replaced by (9):

(9) $p_i = P(R < 0) = \Phi(0; \mu_i - \mu_j, 2)$

Here $\Phi(x; \mu, \sigma^2)$ denotes, as usual, the cumulative normal distribution function with the mean value μ and standard deviation σ evaluated at x.

In the general case of m candidates, the situation is complicated by the need to take into account possible overlaps between sets of constraints violated by O_i and some (but not all) other candidates. However, the computations can still be reduced to integration over m–1 variables. Let me briefly show how for the case of three candidates.

Let $\mathbf{C}(\mathbf{O^*}; \mathbf{O}^+)$ be the set of all constraints violated by candidates that belong to $\mathbf{O^*}$ and not violated by the elements of $\mathbf{O}^+$ (if a candidate does not belong to either set, then the inclusion of a constraint in $\mathbf{C}(\mathbf{O^*}; \mathbf{O}^+)$ does not depend on whether it penalizes this candidate. Let $M(\mathbf{O^*}; \mathbf{O}^+)$ be the maximum weight of constraints from $\mathbf{C}(\mathbf{O^*}; \mathbf{O}^+)$. Then the rankings under which candidate O_1 is optimal fall into three sets that can be described by the following conditions:

(10) a. $M(\{O_2, O_3\}; \{O_1\}) > M(\{O_1\}; \varnothing)$
b. $M(\{O_2\}; \{O_1, O_3\}) > M(\{O_3\}; \{O_1\}) > M(\{O_1\}; \varnothing)$
c. $M(\{O_3\}; \{O_1, O_2\}) > M(\{O_2\}; \{O_1\}) > M(\{O_1\}; \varnothing)$

These conditions are deliberately defined in such a way that the corresponding sets of rankings do not overlap, so that their probabilities can be summed up in order to obtain the probability of O_1. It can be easily verified that if a ranking does not belong to one of these sets, O_1 cannot be optimal. The probability of a ranking belonging to the first set can obviously be defined in terms of competition between two constraint sets and thus calculated by using equation (8). The latter two sets of ranking can only be described in terms of competition between three constraint sets and thus will require integration over two variables to compute the corresponding probabilities. Thus, the maximal number of variables of integration is $m - 1$. This approach can be easily extended to any number of candidates.

3 Language-specific applications: evaluation of parameters

Statistical datasets described by means of a specific StOT model can be represented as $n{\times}m$ matrices of integers, where rows correspond to different environments (inputs) and columns, to competing candidates; d_{ij} is the number of occurrences of candidate O_j in environment E_i. An example of such dataset is given in Table 1.[3] This fragment of grammar can be described by a variety of different StOT models, one of which is shown in Table 2. Note that this constraint set does not distinguish between two environments with pronominal theme (E_1 and E_2).

	Theme	Goal	O_1= Theme *to* Goal	O_2 =Goal Theme	Total
E_1	Pro	NP	13	0	13
E_2	Pro	Pro	123	2	125
E_3	NP	NP	23	13	36
E_4	NP	Pro	16	242	258

Table 1. Dative shift vs. overt dative marking in English

Theme	Goal		C_0	C_1	C_2
E_1: Pro	NP	Pro *to* NP	*		
		NP Pro		*	*
E_2: Pro$_1$	Pro$_2$	Pro$_1$ *to* Pro$_2$	*		
		Pro$_2$ Pro$_1$		*	*
E_3: NP$_1$	NP$_2$	NP$_1$ to NP$_2$	*		
		NP$_2$ NP$_1$		*	
E_4: NP	Pro	NP to Pro	*		*
		Pro NP		*	

Table 2. A three-constraint StOT model for dative shift in English

The first question to be answered is whether this set of constraints can account for English data, that is, whether there exists a vector of constraint strength that would predict probability distributions that are sufficiently close to the observed frequencies. This problem is usually solved by running the Gradual Learning Algorithm and finding out whether it converges to an appropriate vector of constraint strengths. However, the results described in Section 2 suggest a more reliable and computationally effective method.

[3] The dataset is based on the Switchboard corpus and has been kindly provided by Joan Bresnan for testing of the algorithm described here.

As in the previous section, we normalize the model by setting $\mu_0 = 0$ and $\nu = 1$. Applying the results of Section 2, we get the following expressions for probabilities of competing candidates in different environments:

(11) $$p_{11} = p_{21} = \sum_{i=1,2}\int_{-\infty}^{\infty} \phi(m;\mu_i)\prod_{k\neq i}\Phi(m;\mu_k)dm; p_{12} = p_{22} = 1 - p_{11}$$

$$p_{31} = \Phi(m;-\mu_1,2); p_{32} = 1 - p_{31}$$

$$p_{41} = \int_{-\infty}^{\infty} \phi(m;\mu_1)\prod_{k\neq 1}\Phi(m;\mu_k)dm; p_{42} = 1 - p_{41}$$

Once the probabilities of candidates in different environments are expressed as functions of constraint strengths, our StOT model can be dealt with as any other parametric statistical model; in particular, we can use general statistical methods (for example, the method of maximum likelihood) to evaluate the values of relevant parameters (constraint strengths) for any dataset. The numbers of tokens of each structure occurring in a given environment follow the multinomial distribution, that is, the probability of a set of values <d_{i1}, …, d_{im}> for a given total number d_i of occurrences of environment E_i is distributed as follows:

(12) $$P(d_{i1},\ldots,d_{im}) = \frac{d_i!}{d_{i1}!\ldots d_{im}!}\prod_{j=1}^{m} p_{ij}(\boldsymbol{\mu})^{d_{ij}}$$

where $p_{ij}(\boldsymbol{\mu})$ is the probability of candidate O_j occurring under conditions E_i. The corresponding log-likelihood function is:

(13) $$L(\boldsymbol{\mu};\boldsymbol{D}) = \sum_{\mathrm{i,j}} d_{ij} \ln p_{ij}(\boldsymbol{\mu})$$

The optimal vector μ of constraint weights for the given dataset is the one that maximizes this function. According to my counts, μ = <0, 0.429, 2.572>; the predicted probability distributions are shown in Table 3, along with the observed frequency distributions. According to χ^2-test, the dataset in Table 1 can be assumed to follow the predicted distributions, which means that the constraint set in Table 2 can satisfactorily account for English data. Note that the predicted probabilities most closely match the observed frequencies for the case of lexical Theme and pronominal Goal (E_4). This is because this is the most frequent environment; accordingly, the frequencies observed in this environment are supposed to approximate the underlying probabilities better and are therefore given more “weight” in the log-likelihood function.

	Frequency-based Estimates		Model predictions μ = <0, 0.429, 2.572>	
	O_1	O_2	O_1	O_2
E_1	1.0	0.0	0.972	0.028
E_2	0.984	0.016	0.972	0.028
E_3	0.639	0.361	0.619	0.381
E_4	0.062	0.938	0.060	0.940

Table 3. Probabilities of overt dative marking (O_1) vs. dative shift (O_2)

To sum up, the results of Section 2 allow us to use general statistical methods to decide whether a given StOT model can account for the attested cross-linguistic variation and to find the corresponding vector of constraint strengths for each language under consideration without running the GLA (or any other learning algorithm, for that matter). Accordingly, the problem of **existence** of an appropriate vector of constraint weights for the observed dataset can be solved independently of the problems of convergence for learning algorithms. On the other hand, the proposed method of finding constraint strengths can also be thought of as a learning algorithm of another sort (based on the ability of a learner to memorize the frequencies of competing structures in the data).

4 Typological applications: StOT-specific predictions of StOT models

In what follows, I will distinguish between two multidimensional typological "spaces" defined by a StOT model. The dimensions of **primary space** correspond to constraints; the boundaries of this space are determined by a set of conditions like in (14), which define the maximum relevant difference between two constraint strengths and delimit the absolute values of constraint strengths by setting $\mu_0 = 0$:

(14) $\mu_0 = 0; \quad \forall i = 0,...,r, \exists j \neq i : | \mu_i - \mu_j | \leq 10\nu$

	$C_0 > C_1 > C_2$	$C_1 > C_0 > C_2$	$C_2 > C_0 > C_1$	$C_2 > C_1 > C_0$
	$C_0 > C_2 > C_1$	$C_1 > C_2 > C_0$		
E_1, E_2	O_2	O_1	O_1	O_1
E_3	O_2	O_1	O_2	O_1
E_4	O_2	O_1	O_2	O_2

Table 4. Primary vs. secondary typological space

The dimensions of **secondary space** correspond to the predicted probabilities p_{ij} of competing candidates in different environments. Its "mathematical" boundaries are obviously defined by conditions (15), but a StOT model is generally supposed to define a narrower space, i.e. to impose additional linguistic conditions on the possible values of p_{ij} and their relations.

(15) $\forall i, j\ 0 \leq p_{ij} \leq 1; \quad \forall i \sum_{j=1}^{m} p_{ij} = 1$

Informally, the primary typological space corresponds to the factorial typology of the standard version of OT (where different "points" correspond to different rankings), while the points of secondary space correspond to different types of distributions of candidates across inputs/environments. This difference is demonstrated in Table 4 for the constraint set of Table 2; the columns correspond to the points of secondary typological space.

It is the secondary space that constitutes testable typological predictions of an OT grammar. In our example, the standard OT prediction amounts to two implicational universals: if O_2 (=dative shift) is used in E_1 or in E_2, it is also used in all other environments; and if it is used in E_3, it also used in E_4. For the corresponding StOT model, this prediction transforms into the following statement about probabilities of dative shift in different environments, which imposes additional conditions on the secondary typo-logical space:

(16) $p_{12} = p_{22} \leq p_{32} \leq p_{42}$

The predictions of the corresponding standard OT model are obviously entailed by (16): if $p_{12} = 1$ or $p_{22} = 1$, then the probability of dative shift equals 1 in all other environments; and if $p_{32} = 1$, then $p_{42} = 1$. However, the StOT version also accounts for non-categorical (stochastic) preferences and, accordingly, imposes additional conditions on the possible relations between probabilities of competing candidates.

It must be recognized, however, that a prediction like in (16) does **not** depend on the specific StOT assumptions on the interaction of linguistic constraints. In particular, these (in)equalities would be the same independently of whether the combined weight of a set of cooperating constraints is defined as their maximum (as in StOT) or, for example, as their sum. For instance, p_{22} cannot be greater than p_{32} simply because the model contains no constraints that would penalize dative shift in absence of pronouns but not in the context of pronominal theme. The specific mechanism of constraint interaction has nothing to do at all with this prediction. Accordingly, the ability of StOT to predict this sort of constraints on cross-linguistic variation cannot, in principle, corroborate its specific assumptions about constraint interaction. However, the StOT model in our example also entails some more subtle predictions, such as:

(17) If overt dative marking is preferred in constructions with two nominal objects, then the preference for dative shift for pronominal goals is weaker than the preference for overt dative marking in the context of pronominal theme, and vice versa.

This is because the preference for overt dative marking in absence of pronouns means that C_0 is weaker than C_1; accordingly, the combined effect of C_3 and C_0 against C_1 (in the context of pronominal goal) is bound be weaker than the combined effect of C_3 and C_1 against C_0. What is essential is that a preference for overt dative marking or for dative shift in absence of pronouns is **quantitatively** related to the difference between corresponding distributions for pronominal theme and pronominal goal, and this quantitative relation does prominently depend on the StOT mechanism of constraint interaction. A dataset can be in contradiction with the predictions of the model even if it is only slightly different from the English dataset in Table 1. Consider, for example, a faux dataset in Table 5, where the number of tokens of overt dative marking in absence of pronouns is increased and all other numbers are the same as in Table 1, i.e. the preference for overt dative marking in absence of pronouns is supposed to be stronger, whereas the distributions for pronominal objects are exactly as in English.

Following the procedure outlined in Section 2, we can determine the best vector of constraint strengths for this dataset: μ = <0, 0.570, 2.430>; the corresponding probabilities of dative shift are $p_{12} = p_{22} = 0.033$, $p_{32} = 0.343$, $p_{42} = 0.9125$. According to χ^2-test, the dataset in Table 5 cannot be assumed to represent this distribution; in other words, the model predicts that this combination of frequencies is impossible (it is outside the boundaries of secondary typological space defined by the model). Roughly speaking, the model predicts that if the combination of frequency distributions in the first three environments is as in Table 5, then the preference for dative shift in E_4 cannot be that strong (in fact, it is predicted to occur in about 70% of instances).

	Theme	Goal	O_1	O_2	Total
E_1	Pro	NP	13	0	13
E_2	Pro	Pro	123	2	125
E_3	NP	NP	73	13	86

E_4	NP	Pro	16	242	258

Table 5. Dative shift vs. overt dative marking: a faux dataset

These observations are, of course, not intended as a claim that these pre-dictions are "correct", i.e. that a structurally comparable grammatical variation in other languages will conform to these predictions; it may well turn out that this is not the case, which would mean that this particular constraint set just does not work as a partial universal grammar. The point is that these predictions are testable (that is, a comparable grammatical variation in other languages either can or cannot be described by the same constraint set with a different vector of constraint strengths) and the quantitative relations involved in these predictions directly depend on the StOT hypothesis on how linguistic constraints interact with each other. In other words, the predicted boundaries of secondary typological space for the same constraint set would differ depending on how exactly the combined weight of cooperating constraints is defined. Accordingly, we can hope to find cross-linguistic evidence for or against specific hypotheses about the mechanism of constraint interaction.

For Stochastic OT, the equations presented in Section 2 can be used to describe the predicted boundaries of secondary typological space, since they establish a correspondence between the points of primary and the secondary space: it suffices to perform the computations for a sufficient number of points in the primary space to determine the boundaries of the secondary one, and, consequently, the specific quantitative relations between the probabilities of candidates in different environments that have to be tested against cross-linguistic data to verify or falsify the model.

5 Non-redundancy meta-constraint on StOT grammars

As demonstrated in Section 4, the StOT mechanism of constraint interaction can entail non-trivial (StOT-specific) typological predictions in combination with **some** constraint sets. There also exist constraint sets that do not have this property; moreover, it can be shown that for any fragment of linguistic data (that is, for any set of candidates and any set of environments) there exists a StOT model that does not entail any typological predictions whatsoever; that is, for any mathematically possible point in the secondary typo-logical space (as defined in (15)) there exists a corresponding point in the primary space (the appropriate vector of constraint strengths). Such StOT models can be called **trivial**.

A trivial StOT model can be built as follows: define one constraint violated by one candidate in all environments; for each other candidate and for each environment, define one constraint violated only by this candidate and only in this environment. It can be easily shown that such a model allows for any combination of probability distributions; for example, if there are two competing candidates, the values of constraint strengths can be directly calculated using equation (9): for each environment E_i, the strength of constraint violated by O_2 in this environment is the value of inverse normal distribution with mean value $\mu_0 = 0$ and variance 2 for the given probability of O_1 in this environment (assuming that O_1 violates C_0 in all environments).

Triviality of such models is ensured by two properties: first, the only overlap between constraint sets relevant in any pair of environments is the normalizing constraint, which means that the distributions in different environments are determined by distinct sets of parameters and are therefore absolutely independent of each other; secondly, the sets of constraints violated by different candidates in any environment do not overlap at all, which ensures that their strengths can be adjusted to "predict" any probability distribution in this environment. Note that the required number of constraints in such a model is $r = n*(m - 1) + 1$, i.e. the number of degrees of freedom must equal (or exceed) the number of independent data points in any dataset. This property of a constraint set can be referred to as **redundancy**. Of course, whether or not a model is redundant depends on the number of constraints that actually affect the output of evaluation at least in some environments: if a constraint if violated by all candidates in all environments, it must be excluded from the consideration. If a model is non-redundant, some non-trivial overlaps between constraint sets relevant in different environments are bound to exist, which would entail some additional conditions on the secondary typological space.

Trivial redundant models seem to be widely used in actual linguistic applications of StOT, whereby typological predictions are expressed by external meta-constraints in the form of universal **partial rankings** of constraints. Obviously, such models cannot entail any StOT-specific typological predictions; as a matter of fact, the very need for meta-constraints can be viewed as a signal of failure of OT as such to express typological generalizations. Moreover, in this class of applications Stochastic OT just provides a fancy – and rather complicated – way of "rewriting" attested frequencies in another form, since the number of independent parameters of the model equals the number of data points to be accounted for, and, under the conditions stated in (14), there exists a one-to-one correspondence between the vectors of parameter values and all possible sets of probability distributions.

Of course, a redundant model need not be trivial. As an example, consider the partial StOT grammar used by Clark (2004: 83-84) to describe the relative positioning of finite verb and medial adverbs in English; the crucial point is that the position of a verb depends on whether it is auxiliary or lexical: auxiliaries consistently precede adverbs, whereas the position of lexical verb gradually changed in the course of history: in the 15th century, the verb preceded the adverb in ca. 64% percent of cases, and now it always follows the verb. Clark's model contains two constraints (C_0 and C_1 in Table 6) that penalize V-Adv and Adv-V orders respectively (independently of whether the verb is lexical or auxiliary) and an additional constraint (C_2) that penalizes V-Adv order for lexical verbs only. The syntactic change in question can be modeled as gradual strengthening of C_2. In Table 6, the constraint strengths for 1425-75 and 1500-25 periods are those cited by Clark, but normalized according to conventions adopted here. The constraint strengths given for Modern English simply reflect the fact that the ranking of constraints does not vary: C_1 must always be ranked below C_0 (since adverbs cannot precede auxiliaries) and C_2 must always be ranked above C_1 (since lexical verbs cannot precede adverbs), i.e. there is no variation. In fact, the strength of C_1 could be set to –10 for the earlier periods as well without any difference for the predictions of the model (the only relevant fact is that this constraint never outranks either of the other constraints).

This model is redundant: it has two degrees of freedom for two independent data points. However, it is non-trivial, since it imposes the following condition on the secondary typological space:

(18) $p_{12} \geq p_{22}$

		C_0 (*OB-HD(IP))	C_1 (*I)	C_2 (*Lex-in-I)
	1425-75	$\mu_0 = 0$	$\mu_1 = -18.3$	$\mu_2 = -0.5$
	1500-25	$\mu_0 = 0$	$\mu_1 = -24.95$	$\mu_2 = 1.05$
	Modern	$\mu_0 = 0$	$\mu_1 = -10.0$	$\mu_2 = 10.0$
E_1 Aux	O_1 = Adv V		*	
	O_2 = V Adv	*		
E_2 Lex	O_1 = Adv V		*	
	O_2 = V Adv	*		*

Table 6. A redundant non-trivial StOT model for verb positioning in English (Clark 2004)

As discussed in Section 4, such predictions are not StOT-specific: (18) would follow from this constraint set independently of how exactly the constraints interact, simply because there are no constraints that would penalize

O_2 for auxiliaries but not for lexical verbs. Accordingly, the ability of StOT to "predict" any pair of distributions satisfying condition (18) cannot be considered as evidence in favor of the StOT mechanism of constraint interaction.

As it seems, redundant StOT models would generally entail no StOT-specific typological predictions, insofar as they minimize non-trivial overlaps between constraint sets relevant in different environments. Conversely, if the number of constraints is minimized, StOT-specific typological predictions are bound to be more significant. This implies that Stochastic OT can be considered as a plausible model of constraint interaction only insofar as linguistically plausible StOT partial grammars are non-redundant. In other words, if a certain type of grammatical variation can be described by a non-redundant StOT grammar (i.e. the typological predictions of such a grammar are not falsified by available cross-linguistic statistical data), this can be taken as evidence for the StOT mechanism of constraint interaction. Conversely, if the quantitative constraints on the secondary typological space imposed by **any** linguistically plausible and non-redundant set of constraints under the StOT assumptions on constraint interaction turn out to be too strong, this means that the StOT mechanism of constraint interaction does not work for the corresponding type of grammatical variation.

It seems worth noting that redundant models can also be ruled out on in-dependent grounds. If the constraint set of a StOT model is thought of as a part of UG, then this part of UG is redundant in the sense that it does not re-duce the amount of information to be learnt in the course of language acquisition by a single bit: formally, such "learning" is equivalent to memorizing frequencies of the candidates in the primary data, albeit in a non-direct way. If language learners can in some sense memorize $n*(m-1)$ constraint strengths, they can just as well memorize n independent discrete probability distributions for m candidates and than randomly draw candidates from these distributions; in fact, both the learning process and the evaluation process would be considerably simplified. These considerations suggest that Stochastic OT, as a linguistic model of constraint interaction, must include a general meta-constraint prohibiting redundant grammars.

A possible objection to this meta-constraint is that the linguistic constraints participating in partial StOT grammars usually come from an underlying linguistic theory and are motivated by other considerations as well. The existence of independent motivation means, however, that these constraints are also supposed to account for some other linguistic data, apart from the datasets under consideration. If so, then these other datasets must also be taken into account in establishing the values of constraint strengths, since the constraint strengths are supposed to be constant within a single language. This means that in order to find the appropriate values of language-specific parameters of the model we must consider a more comprehensive StOT grammar, which would account for all relevant pieces of statistical data. In this case, the non-redundancy meta-constraint applies to this more comprehensive model. If this model proves to be too complex to be dealt with in actual practice, the constraint weights can be established on the basis of a partial dataset and then tested against the other relevant data. In any event, the differences between strengths of shared constraints must be the same across all relevant datasets, which obviously reduces the actual number of degrees of freedom of each partial StOT grammar, so that the non-redundancy constraint is satisfied.

To sum up, the StOT assumptions about the mechanism of constraint interaction can but need not entail testable typological predictions depending on the structure of constraint sets involved in StOT grammars. In particular, these assumptions can be corroborated by cross-linguistic data only insofar as these data are described by a non-redundant StOT grammar.

References

Boersma, P., Hayes B. 2001. Empirical tests of the Gradual Learning Algorithm. *Linguistic Inquiry* 32: 45-86.

Clark, B.Z. *A Stochastic Optimality Theory Approach to Syntactic Change.* Dissertation. Department of Linguistics. Stanford University.