The winner takes it all – almost.
Cumulativity in grammatical variation

Gerhard Jäger
Gerhard.Jaeger@uni-bielefeld.de

Anette Rosenbach
ar@phil-fak.uni-duesseldorf.de

1 Introduction

Classical Optimality Theory in the sense of Prince and Smolensky (2004/1993) implements the intuition that **grammars cannot count**. The grammaticality of a candidate is fully determined by the ranking of the relevant constraints. Numerical constraint weights play no role. Furthermore, if a competition between two candidates is decided by a constraint *c*, it only matters which candidate violates *c* more often. The numerical proportions of the amount of constraint violations play no role for grammaticality. This feature of OT is usually called **strict constraint domination**.

Several researchers have considered probabilistic generalizations of classical OT in recent years (see for instance Anttila 1997, Boersma 1998, Johnson and Goldwater 2003). It is common to these approaches that the evaluation component does not just assign categorical grammaticality values to the candidates ("grammatical" or "ungrammatical"), but rather probabilities which may take real values between 0 and 1. It does not follow necessarily though that constraint violations can add up in a probabilistic version of OT. Different versions of Stochastic OT in fact differ as to whether they predict cumulativity effects in constraint evaluation.

In this paper we show that cumulativity is necessary to account for probabilistic variation found in actual language use, and we compare the accuracy of the predictions that different versions of Stochastic OT make. We distinguish two versions of cumulativity, namely *ganging-up cumulativity* and *counting cumulativity*. We will compare how Paul Boersma's version of Stochastic OT on the one hand and Maximum Entropy models on the other hand deal with cumulativity. The second part of the paper reports empirical data on English genitive variation. It turns out that both versions of cumulativity do obtain in the empirical data. In the last part of the paper we compare the predictions of the two theories with respect to this empirical domain. The Maximum Entropy model proves to be clearly superior, both with respect to the accuracy of its predictions and to its learnability properties.

2 Cumulativity and Stochastic OT

2.1 Two kinds of cumulativity

In standard OT, the evaluation follows the slogan "The winner takes it all". The optimal candidate is grammatical, and all sub-optimal candidates are ungrammatical. In a competition between two candidates, the constraints are evaluated one by one according to their strength, and once a competition is decided, lower ranked constraints have no impact on the outcome.

Stochastic generalizations of OT have to reserve some probability mass to the loser, so to speak. In this setting the issue whether constraints and constraint violations can add up and have a cumulative effect arises anew. Different stochastic generalizations of OT vary in this respect. Before we look at these models though, let us first make precise what we mean by "cumulativity" in a probabilistic setting.

There are actually two notions of cumulativity that standard OT rejects. One way in which OT is non-cumulative can, metaphorically speaking, be paraphrased as "Two weak constraints can never gang up to jointly beat a stronger constraint". If a method of constraint evaluation does not follow this prinicple, we therefore talk about **ganging-up cumulativity**. The other notion of cumulativity pertains to the amount of violations of a constraint. OT

follows the principle "A single violation of a stronger constraint is more severe than any amount of violations of a weaker constraint". If this fails to hold, we talk about **counting cumulativity**.

What would ganging-up cumulativity mean in a probabilistic setting? There are several options here. Generalizing from the categorical case, ganging-up cumulativity entails that **the presence of a dominated constraint can matter**. To see this point, consider the tableaux in (1).

(1)

	c_1	c_2	c_3
a_1		*	
a_2	*		

	c_1	c_2	c_3
b_1			*
b_2	*		

	c_1	c_2	c_3
d_1		*	*
d_2	*		

Suppose the constraints are ranked as indicated, i.e., $c_1 > c_2 > c_3$. Under categorical evalution, ganging-up cumulativity would obtain if a_1 would be the winner of the first competition, b_1 of the second competition, but d_2 would win the third competition. The difference between the second tableaux and the third is that c_2 is inactive in the former but active in the later. In both cases, c_2 is dominated by an opposing constraint, but still the presence of c_2 alters the outcome. We take this pattern to be the essence of ganging-up cumulativity. That several constraints can "gang up" is a side effect. If c_2 allone would be sufficient to ensure the victory of the second candidate – i.e., if a_2 would win the first competition – this would effectively mean that c_2 dominates c_1.

Given this, the generalization to probabilistic evaluation is straightforward. A method of stochastic constraint evaluation shows ganging-up cumulativity if it is possible that the presence of a dominated constraint increases the probability of the candidates that are optimal according to this constraint.

To make this formally precise, we have to pin down what it means for a constraint to be dominated, and we have to do this in a theory-independent way. This is not fully possible because the notions of constraint ranking and domination are something theory-internal. However, we can state a meta-theoretical constraint for all conceivable notions of constraint domination. We are using the notion of "weak domination" here, which includes the possibility of a tie between two constraints. Strong domination means that the first constraint weakly dominates the second one, but not the other way round. (A note on terminology: when we say that a constraint is active in a competition, we mean that it does not assign the same number of violations to each candidate, and a constraint is, of course, inactive iff it is not active.)

Constraint 1 (Constraint domination): *Suppose two competitions are identical except that the constraint c_1 is only active in the first competition and c_2 only in the second competition. Suppose furthermore that the candidate x is optimal according to c_1, but not according to c_2. If c_1* weakly dominates *c_2, then the probability of x in the first competition must be at least as high as in the second competition.*

Probabilistic ganging-up cumulativity, as we conceive it, means that a strongly dominated constraint has an effect. Thus we define:

Definition 1 (Ganging-up cumulativity): *A probabilistic constraint evaluation method* predicts ganging-up cumulativity *iff the following situation is possible:*

1. *Constraint c_1 strongly dominates c_2.*

2. *The competitions A and B are identical (involve the same candidates, constraints, and constraint ranking) except that c_2 is active in A but inactive in B.*
3. *The candidate x, which is part of both A and B, is optimal according to c_2 but not according to c_1.*
4. *The probability that the evaluation assigns to x relative to A is higher than the probability that it assigns to x relative to B.*

In a categorical context, i.e., if all probabilities are either 0 or 1, this notion of ganging-up cumulativity reduces to the standard notion (provided the constraint above is fulfilled). So even though the definition does not, strictly speaking, involve a "ganging-up" of several constraints, it is a genuine probabilistic generalization of the categorical notion of ganging-up cumulativity.

Paul Boersma (p.c.) suggests another probabilistic notion of ganging-up cumulativity. Consider again the tableaux in (1). As said above, in a non-probabilistic setting, ganging-up cumulativity would obtain for instance if a_1 would win the first competition, b_1 the second, but d_2 the third (provided the constraint ranking is kept constant across the competitions). According to Boersma's notion, a probabilistic evaluation displays ganging-up cumulativity if for each ε, there is a ranking such that, $P(a_2|\{a_1,a_2\}) < \varepsilon$, $P(b_2|\{b_1,b_2\}) < \varepsilon$, but $P(d_2|\{d_1,d_2\}) > 1-\varepsilon$. We might call this notion of cumulativity **strong ganging-up cumulativity**, while the one that was defined in Definition 1 is fittingly dubbed **weak ganging-up cumulativity**. These names are appropriate because strong ganging-up cumulativity entails categorical ganging-up cumulativity, which in turn entails weak ganging-up cumulativity. Weak ganging-up cumulativity, finally, does not entail strong ganging-up cumulativity.

In the remainder of this paper, we will exclusively be concerned with weak ganging-up cumulativity. Therefore we do not give a formally precise definition of the strong notion here.

Let us now turn to counting cumulativity. Consider the tableaux in (3):

(3)

	c_1	c_2
☞ a_1		*
a_2	*	

	c_1	c_2
b_1		***
☞ b_2	*	

Suppose a_1 would win the first competition while b_2 would win the second. This would be an instance of counting cumulativity. In general, counting cumulativity admits that a single violation of a constraint c_1 is less severe than a single violation of c_2, but n violations of c_2 (for some $n>1$) are more severe than a single violation of c_1.

The essential point here is that both constraints define isomorphic orderings in both competitions, while the concrete numerical values differ. In standard OT, this should never make a difference. In a system with counting cumulativity, it could make a difference. In the categorical setting, "to make a difference" means "switching from ungrammatical to grammatical". This can be generalized to the probabilistic setting. Here, to make a difference simply means to change the probabilities that the evaluator assigns. This leads to the following definition:

Definition 2 (Counting cumulativity): *A probabilistic constraint evaluation method* predicts counting cumulativity *iff the following situation is possible:*

1. *The two competitions A and B are completely identical except that the constraint c_1 assigns more violation marks to the candidate x in A than in B.*
2. *Despite this difference, c_1 induces the same ranking of candidates on A as on B.*
3. *The evaluation assigns a higher probability to x in A than in B.*

If the range of possible probabilities is restricted to 0 and 1, the definition covers categorical counting cumulativity as a special case.

2.2 Boersma's Stochastic OT

We will only give a brief sketch of Boersma's model here – the interested reader is referred to Boersma and Hayes (2001) for a more thorough introduction. StOT shares the generator component with standard OT. It also uses a set of ranked and violable constraints as the basis for grammatical evaluation. The constraints are not ranked on an ordinal scale though, but each is assigned a real number on a continuous scale, its *rank*. This way it is possible to speak of the distance between two constraints in a meaningful way. In each evaluation event, some random noise is added to each constraint rank. The ranking of a constraint after adding the noise is called the *selection point*. The constraints can be ordered according to the value of their selection points, and this ordering can be used as ranking in the standard OT sense. However, adding the noise value to the ranks of the constraint may change their ordering, so the ranking of selection points may differ from evaluation to evaluation. In this way the ranking on the continuous scale defines a probability distribution over ordinal rankings. This in turn defines a probability distribution over the set of candidates – the probability of a candidate to be optimal is the sum of the probabilities of all ordinal rankings that make it optimal. The noise that is added to each constraint rank is a normally distributed random variable with mean 0 and standard deviation 1. The probability of a constraint ranking $c_1 > c_2 > \ldots > c_n$ can thus be given by the following formula (where r_i is the rank of constraint c_i and N is the standard normal distribution):

$$P(c_1 > \cdots > c_n) = \int_{-\infty}^{+\infty} dx_1 N(x_1 - r_1) \int_{-\infty}^{x_1} dx_2 N(x_2 - r_2) \cdots \int_{-\infty}^{x_{n-1}} dx_1 N(x_n - r_n)$$

A StOT grammar adequately describes a language if it assigns probabilities to the linguistic signs (sentences, syllable structures or whatever) in the corpus that match with their empirical relative frequencies in this language. If the distances between the ranks of the constraints are very high, the probability of the ordinal ranking that matches the ordering of the ranks converges towards 1. Standard OT, where there is only one ranking, can thus be seen as a borderline case of StOT.

StOT predicts (weak) ganging-up cumulativity. Consider a situation where we have three constraints, c_1, c_2, and c_3, and two candidates, *a* and *b*, such that $c_1(a) < c_1(b)$, $c_2(a) > c_2(b)$, and $c_3(a) > c_3(b)$. (We construe constraints as functions from candidates to numbers of violation marks here. Hence an expression like "$c_1(b)$" denotes the number of violation marks that c_1 assigns to *b*.) Suppose all three constraints are equally ranked. Then each ordinal ranking between them is equally likely. There are two rankings where c_1 is the strongest constraint, and four where one of the other two wins. Hence $P(a) = 1/3$ and $P(b) = 2/3$. Now suppose c_1 is promoted by a very small amount. Then it will be the strongest constraint, but if the promotion step is small enough, the probabilities of *a* and *b* are still very close to *1/3* and *2/3* respectively. To make the argument mathematically water-proof, suppose we have an infinite descending sequence of rankings for c_1 which converges towards the ranking of c_2 and c_3. Since the function that maps vectors of ranks to probabilities in StOT is continuous, the probabilities of *a* and *b* will converge to *1/3* and *2/3* respectively. Hence there are rankings where c_1 is the highest ranked constraint but $P(a) < 50\%$. Technically, if c_1 would only compete with c_2, it would win with a probability of > 50%, but in a competition with both c_2 and c_3 it wins with less than 50% probability. So while c_3 is dominated by c_1, it still has an impact on the probabilities that are assigned.

It can also be seen from this discussion that StOT does not predict strong ganging-up cumulativity. The probability of *a* will always be *> 1/3*, so a value of 0.2 for ε would falsify strong ganging-up cumulativity.

StOT does not predict counting cumulativity either. The probability of a candidate is defined indirectly, via probabilities of categorical OT-competitions. Since categorical OT does not have counting cumulativity, StOT does not predict it either.

2.3 Maximum Entropy models

Goldwater and Johnson (2003) compare StOT with Maximum Entropy models (or, as they are sometimes called, log-linear models) that are state of the art by now in computational linguistics (see for instance Berger et al. 1996 or Abney 1997). Let us briefly explain what "maximum entropy" means.

Suppose we know that a certain experiment has two possible outcomes, *A* and *B*, but we do not know anything else about it. Which probability should we assign to *A* and *B*? The best answer seems to be: 50% probability for each. Likewise, if there are five possible outcomes, *A*, *B*, *C*, *D* and *E*, the best estimate is to assign 20% to each if we don't have further information. Every other distribution of the probability mass would represent a bias which is not supported by knowledge. And if we also know that the outcome will be *A* or *B* with a probability of 70%, then the least biased estimate is to assign 35% to both *A* and *B*, and 10% to each of the three other events. There is a clear intuition that the least biased hypothesis is the most parsimonious one.

The information theoretic notion of *entropy* quantifies the bias of a probability distribution. The entropy *H* of a probability distribution *p* is defined as

$$H(p) = \sum_x p(x) \log \frac{1}{p(x)}$$

The least biased distribution has the highest entropy, and vice versa. If we have partial knowledge about a stochastic process and we have to estimate the underlying probability distribution, the best guess is to choose among all distributions that are compatible with our knowledge the one with the highest entropy.

Let us assume that the unknown probability distribution is a language *L* in the sense of probabilistic OT, i.e., a probability distribution over a set of input-output pairs. We know the set **GEN** of possible elements of the language (the generator) and a set of constraints. We also know how many violations each constraint incurs on each candidate, the marginal probabilities of the different inputs, and – crucially – we know how often each constraint is on average violated per input in the language in question.[1] This may be the result of investigating a large sample of *L*, but the only empirical facts we are able to observe are the inputs and the number of constraint violations of each observation. So we are looking for a relative probability distribution over the potential output for each input which predicts the correct average degree of violation of each constraint, and among all distributions with this property, we will choose the one with the highest entropy. It can be shown (see for instance Della Pietra et al. 1995 for a proof) that this distribution takes the following form:

$$p_r(o|i) = \frac{1}{Z_r(i)} \exp(-\sum_j r_j c_j(i,o))$$

where $Z_r(i)$ is a normalization constant which ensures that the probabilities of all candidates sum up to 1. It holds that

[1]Note that we do not claim that this information is available to the learner. Rather, this is the kind of information that is (ideally) available to the linguist, and it can be used to test the adequacy of theoretical models.

$$Z_r(i) = \sum_{o:\mathrm{GEN}(i,o)} \exp(-\sum_j r_j c_j(i,o))$$

Taking the logarithm on both sides yields

$$\log p_r(o|i) = -\sum_j r_j c_j(i,o) - \log Z_r(i)$$

So the logarithm of the probability of a candidate is a linear function of its constraint violations. (Therefore these probability distributions are called "log-linear".) Of course there are infinitely many log-linear distributions, depending on the values of the rank parameters r_j. Della Pietra et al. (1995) also show that among all these log-linear distributions, the one which maximizes the likelihood of the language *L* is the one which assigns the correct average degree of violations to each constraint. In other words, the unique log-linear distribution which assigns maximal likelihood to *L* is at the same time the unique distribution with the empirically correct predictions of average constraint violations that maximizes entropy.

The parameters r_j, the **weights** or **ranks** of a constraint, can be interpreted as measures of the **perplexity** that the constraint induces. (Technically, the weight is actually related to the logarithm of this perplexity.) The higher the rank of a constraint, the more surprised (or "perplex") I will be to see it violated, judging from the experience from the training corpus.

It is worth noting that the predecessor of Optimality Theory, **Harmonic Grammar** (HG henceforth, see Legendre et al. 1990) has a very similar mathematical setup to MaxEnt models. In HG, each constraint has a numerical weight (analogous to the rank of constraints in MaxEnt models), and the **harmony** of a candidate is the negated weighted sum of its constraint violations. The winner of a competition is the candidate with the highest harmony. The harmony of a candidate thus differs from the logarithm of its probability under the MaxEnt interpretation only by the constant $Z_r(i)$. Since this constant is identical for all candidates in a competition, the winner under the HG interpretation is always the candidate with the highest probability under the MaxEnt interpretation.

Despite their similarity, the motivations for the two models are very different. MaxEnt models are derived from first principles of information theory, while HG models are a high level description of a certain class of connectionist networks. While this kinship has been noted before (Johnson 1998), we are not aware of further explorations of this connection.[2]

The general setup of a maximum entropy model is also quite similar to a StOT grammar. The main difference between StOT and MaxEnt is the evaluation component, i.e., the way in which constraint ranks are interpreted as a probability distribution. Like StOT, MaxEnt models can be seen as a generalization of standard OT. If the ranks (or "weights", as the parameters *r* are usually called in the MaxEnt tradition) of the constraints are very high and spread far apart, the probabilities of candidates that would be sub-optimal in classical OT converge towards 0 in the MaxEnt interpretation.

It follows from the definitions that MaxEnt evaluation predicts ganging-up cumulativity in its weak and strong form, as well as counting cumulativity. As for weak ganging-up cumulativity, consider the scenario in (1), which is repeated here as (4) for convenience.

(4)

	c_1	c_2	c_3
a_1		*	
a_2	*		

	c_1	c_2	c_3
b_1			*
b_2	*		

[2]The restriction to positive weights is no serious restriction. A constraint with negative weight is equivalent to its negation with the corresponding positive weight, and a constraint with weight 0 is as good as non-existent.

	c_1	c_2	c_3
d_1		*	*
d_2	*		

Suppose $r_1=3$ and $r_2=r_3=2$. Then the probabilities of a_1 and b_1 are both $e^{-2}/(e^{-2}+e^{-3})\approx 73\%$, and the probability of d_1 is both $e^{-4}/(e^{-3}+e^{-4}))\approx 27\%$. So if everything else remains equal, activating c_3 has an impact even though it is dominated by c_1.

Now suppose the same constraints, but the weights are *3k and 2k* instead of *3* and *2*, for some positive constant *k*. Consider the scenario in (4) again. If *k* grows to infinity, the probability of a_1 and b_1 converges to *1*, while the probability of d_1 converges to *0*. This illustrates that MaxEnt evaluation also predicts strong ganging-up cumulativity.[3]

Finally, consider the tableaux in (5).

(5)

	c_1
a_1	*
a_2	

	c_1
b_1	**
b_2	

Suppose $r_1=log\ 2$. Then the probabilities of a_1 and b_1 are *1/3* and *1/5* respectively. Hence MaxEnt evaluation predicts counting cumulativity.

As a side remark: (5) also illustrates another important difference between StOT and MaxEnt: a_1 is harmonically bounded. Therefore it would have probability 0 under StOT. MaxEnt, however, assigns a non-zero probability to it. Generally, no candidate is ever strictly speaking impossible under MaxEnt. We will return to the issue of harmonic bounding later.

3 Empirical evidence for cumulativity: English genitive variation

In this section we will show that we actually need cumulativity to describe empirical facts adequately, and which versions of cumulativity are necessary.

Our study deals with English genitive variation, which represents a case of grammatical variation in the noun phrase. In English, very often the *s*-genitive (*the king's palace*) and the *of*-genitive (*the palace of the king*) can be used to express a possessive relation.

English genitive variation					
s-genitive			*of*-genitive		
possessor	POSS *'s*	possessum (head)	possessum (head)	*of*	possessor
the king	*'s*	*palace*	*the palace*	*of*	*the king*

However, the choice between these two genitives is not random, but determined by various factors. These factors do not determine categorically which construction is to be used, but rather the likelihood with which the two genitives are used, i.e., their frequency distribution. Therefore, English genitive variation represents a case of probabilistic variation.[4]

[3] Paul Boersma (p.c.) pointed out to us that StOT and MaxEnt make different predictions with regard to strong ganging-up cumulativity, even though they behave similar with respect to the weak notion.

[4] Note that only determiner *s*-genitives (*the girl's eyes*) and *of*-genitives where the possessor is a complement (*the frame of the chair*) were compared in this study. Possessors functioning as modifiers (*women's magazines, a*

In Rosenbach (2002) three such factors were investigated in an experimental study, i.e., animacy, topicality, and possessive relation, and the results provide evidence for ganging-up cumulativity. In Rosenbach (2003) the factors animacy and weight were compared in an experimental study as well as a corpus analysis. The results of this study provide evidence for counting cumulativity. In the following, we will report the rationale and findings of these two studies and point out in how far they provide evidence for cumulativity.

3.1 Ganging-up cumulativity

Animacy, topicality, and the type of the possessive relation are well-known factors determining the choice of genitive construction (see e.g. Altenberg 1982; Quirk et al. 1985; Jucker 1993; Taylor 1996; Leech et al. 1994; Anschutz 1997; Biber et al. 1999; Huddleston & Pullum 2002). Table 1 illustrates how these factors affect English genitive variation:[5]

factors	**preference for the *s*-genitive**	**preference for the *of*-genitive**
animacy	[+ animate] possessor: *the boy's eyes > the eyes of the boy*	[-animate] possessor: *the frame of the chair > the chair's frame*
topicality	[+topical] possessor: *the boy's eyes > the eyes of the boy*	[-topical] possessor: *the headlamps of a car > a car's headlamps*
possessive relation[6]	[+ prototypical] possessive relation: *the boy's eyes > the eyes of the boy*	[- prototypical] possessive relation: *the condition of the car > the car's condition*

Table 1: Animacy, topicality, and possessive relation as factors determining English genitive variation

In general, the *s*-genitive is preferred if the possessor is animate, topical, or in a prototypical possessive relation. If not, the *of*-genitive appears to be the preferred choice. The example of *the boy's eyes* illustrates an important methodological problem: The factors animacy, topicality, and possessive relation correlate to quite an extent with each other. Usually, topics are animate, and prototypical possessors are animate. So, in the example of *the boy's eyes* it is very difficult to tell whether the *s*-genitive is preferred because the possessor *the boy* is animate, or, as a definite noun phrase, high in topicality,[7] or because the kin relation represents a prototypical possessive relation. That is, when the three factors cluster, we simply cannot tell how the three factors contribute to the choice of the *s*-genitive. For this reason, these three factors need to be teased apart in the empirical analysis.

This was done in an experimental study in Rosenbach (2002). In a questionnaire, subjects were presented with little text passages adapted from crime fiction novels which provided contexts for genitive constructions, and subjects had to choose then as spontaneously as possible to use the *s*-genitive or the *of*-genitive in the given contexts. Here's an example from the questionnaire to illustrate the task:

(6) *He passed through the entrance where a sign identified the park as Island Gardens. At its far west end, a circular brick building stood, domed in glass and mounted by a white and green lantern cupola. A*

---

man of honour) were systematically excluded from this study, as they are not subject to the same systematic variation.

[5] A first analysis of the single factors animacy, topicality, and possessive relation in Rosenbach (2002) confirmed this pattern.

[6] Note, that the factor of 'possessive relation' is notoriously difficult to define (cf. also Rosenbach 2002: §4.3). The Rosenbach (2002) study follows Koptjevskaja-Tamm's (2001) binary classification of possessive relations into prototypical and non-prototypical ones for the languages of Europe, with the former comprising kin relations, body parts, and legal ownership. For further details pertaining to this classification, we again refer to Rosenbach (2002: §6.2.2).

[7] The factor of topicality was defined both in terms of definiteness and discourse givenness in this study (for further details see again Rosenbach 2002:112-113).

movement of white shimmered against the red bricks, and Lynley saw Jimmy Cooper trying ***[the door of the building/ the building's door]****.* (Elizabeth George, *Playing for the Ashes*, 585)

Crucially, only such contexts were chosen where both the *s*-genitive and the *of*-genitive are possible; note that this is not always the case.[8] So, for example, indefinite possessive NPs cannot be expressed by the (determiner) *s*-genitive since the possessor renders a possessive NP definite (cf. e.g. Huddleston 1984:253; Lyons 1989, 1999:23), even if the possessor itself is indefinite (Woisetschlaeger 1983); see e.g. *a book of a teacher ≠ a teacher's book* ('the book of a teacher').

To test for the relative strength of the factors animacy, topicality, and possessive relation all logically possible combinations of the 3 factors were tested, resulting in 8 conditions to be tested. There were at least 10 items per conditions, in all 93 items were tested. Table 2 illustrates what the conditions and items looked like:

[+animate]				**[-animate]**			
[+topical]		**[-topical]**		**[+topical]**		**[-topical]**	
[+proto]	**[-proto]**	**[+proto]**	**[-proto]**	**[+proto]**	**[-proto]**	**[+proto]**	**[-proto]**
the boy's eyes/ the eyes of the boy	*the mother's future/ the future of the mother*	*a girl's face/ the face of a girl*	*a woman's shadow/ the shadow of a woman*	*the chair's frame/ the frame of the chair*	*the bag's contents/ the contents of the bag*	*a lorry's wheels/ the wheels of a lorry*	*a car's fumes/ the fumes of a car*

Table 2: Experimental study (Rosenbach 2002): conditions and items

Note, that in Table 2 the factors are already arranged in such a way that stipulates animacy as the most important factor, followed by topicality, and then possessive relation. If this hierarchy holds true, we'd expect the *s*-genitive to become less frequent from left to right. Figure 1 shows the results for the British subjects.[9]

First of all, we can notice that the relative frequency of the *s*-genitive decreases steadily – and significantly – from left to right, except for the difference between the last two conditions, which is random.[10] Therefore, the relative importance of the three factors is indeed:[11]

(7) animacy > topicality > possessive relation

Note, however, that animacy is not *per se* the strongest factor. While for the first three animate conditions the *s*-genitive is always preferred to the *of*-genitive, irrespective of the values for topicality and possessive relation, this picture changes in the fourth animate condition (the *a woman's shadow* type). Here the possessor is not topical and the possessive relation is a non-prototypical one, i.e. both the values for topicality and possessive relation favor the *of*-genitive in this case. And indeed we can see that the *of*-genitive becomes the

[8] While it is certainly interesting to know in which contexts the *s*- and the *of*-genitive are used categorically, it would be fatal for any study of genitive variation to include such contexts in the empirical analysis, since this would seriously confound the quantitative results.

[9] The same items were tested with 48 American English subjects. Since the general pattern is essentially same as for the British speakers, the results for the American English group will be neglected here.

[10] Most likely, the difference between the last two conditions is random is because they represent the 'worst' context for the occurrence of the *s*-genitive. At the lower end of the scale, subjects might have been simply more insecure in their choices. Note also that non-topical, inanimate possessors are particularly prone to receive a compound interpretation instead of a phrasal (determiner) reading. Although the contexts had been carefully chosen as to force a specific interpretation of the possessor, it cannot be completely ruled out that subjects may have interpreted *a lorry's wheel* as *a [lorry's wheel]* instead of *[a lorry's] wheel*. For further discussion on the deviate behavior of the last two conditions see Rosenbach (2002:171-176).

[11] For the statistical analyses we refer to Rosenbach (2002). Note also, that Figure 1 shows that the three factors are separate (if naturally correlating with each other), i.e., none can be reduced to the other(s).

preferred option here (57%): *the shadow of a woman* is more frequent than *a woman's shadow*. So, although individually, topicality and possessive relation are weaker constraints on the choice of *s*-genitive they can both together 'knock out' animacy. This is clear evidence for ganging-up cumulativity.

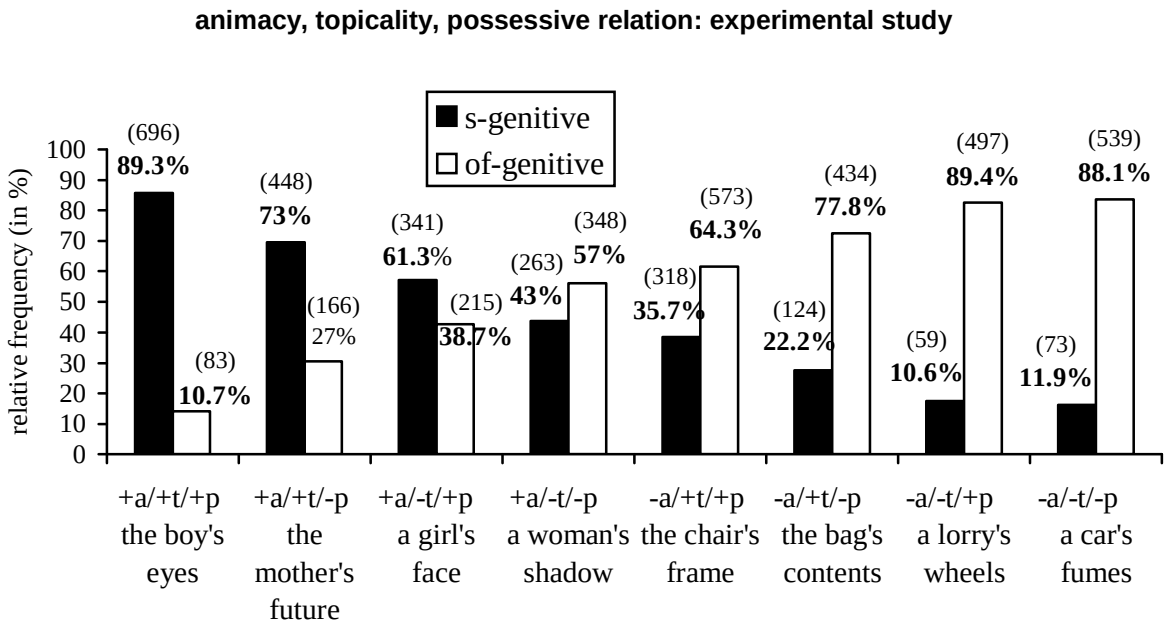

Figure 1: Animacy, topicality, and possessive relation – results of experimental study, British subjects (n=56), absolute number of token given in brackets above each column

3.2 Counting cumulativity

In Rosenbach (2003) the relative strength of the factors animacy and weight were compared. Among the factors determining English genitive variation, syntactic weight is certainly another important one. Weight can be defined in two ways: If we only look at the weight of the possessor (= absolute weight), we can notice that the *s*-genitive is preferred if the possessor is short (cf. Biber et al. 1999: 304-5). If we also take a look at the relative weight between possessor and possessum, then the prediction is that possessives should show a preference for 'short before long', following Behaghel's (1909/10) *Gesetz der wachsenden Glieder*.[12] This predicts the *s*-genitive to be preferred with a short possessor (and a long possessum), as in *John's two elder brothers*, while the *of*-genitive should be preferred with a long possessor and a short possessum, as in *the house of the London real estate agent John Miller*. Note, however, that there is also a correlation between animacy and weight: Animates tend to be shorter than inanimates (see e.g. Wedgwood 1995, cited in Kirby 1999: 118-9), so, again, it is difficult to tell whether in examples such as *John's mother* the *s*-genitive is chosen because the possessor *John* is animate or because it is short.[13] Again, the two factors need to be teased apart. To this end, another experimental study was carried out in Rosenbach (2003), which was basically identical in design to the Rosenbach (2002) study, if, naturally, differing in the conditions to be tested. Most crucially, animacy and weight were teased apart, comparing two conditions where animacy and weight do not go together, i.e. a long animate possessor (& short possessum), as in *the dark man's hand*, and an inanimate short possessor

[12] As far as we are aware of, this question has only been addressed by Altenberg (1982) in his study of genitive variation in 17th-century English. For an analysis of the impact of relative weight on modern English genitive variation, see Rosenbach (2003).

[13] Given the correlation between animacy and weight, Hawkins (1994: 337) even speculates that animacy is an epiphenomenon of weight. For a refutation, see Rosenbach (2003).

(& long possessum), as in *the hotel's elegant lobby*; there were also two baseline conditions which were neutralized as to weight, see Table 3.[14]

animate		**inanimate**	
neutral	**long possessor/short head**	**short possessor /long head**	**neutral**
the boy's eyes/ the eyes of the boy	*the dark man's hand/ the hand of the dark man*	*the hotel's elegant lobby/ the elegant lobby of the hotel*	*the chair's frame/ the frame of the chair*

Table 3: Experimental study: animacy vs. weight – conditions and items (at least 10 items per condition)

If animacy is a stronger factor than weight, then the *s*-genitive should be more frequent with *the dark man's hand* than with *the hotel's elegant lobby*. If, however, weight is stronger than animacy, it should be the other way round, i.e., the *s*-genitive should be more frequent with *the hotel's elegant lobby* than with *the dark man's hand*.

A questionnaire study with 39 American subjects revealed the following results, see Figure 2:

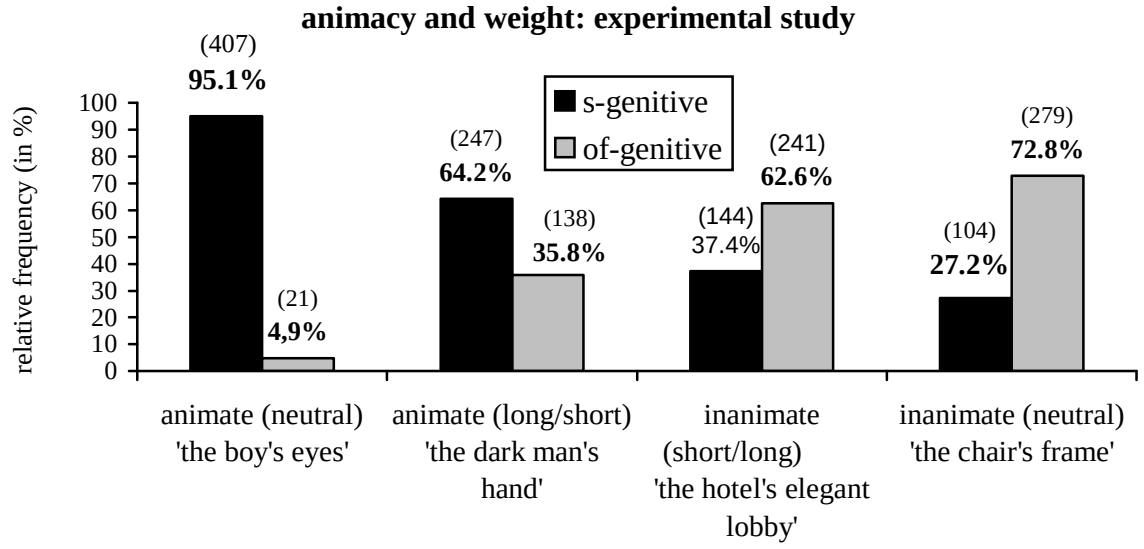

Figure 2: Experimental study: relative frequency of the *s*-genitive and the *of*-genitive (number of subjects: N = 39); absolute number of tokens given in brackets above each column

Figure 2 clearly indicates that animacy is a stronger factor than weight, since the *s*-genitive (64.2%) is more frequent in the animate long/short condition (e.g. *the dark man's hand*) than in the inanimate long/short condition (*the hotel's elegant lobby*), 37.4%. Moreover, the *s*-genitive is also more frequent than the *of*-genitive in the animate long/short condition, i.e., it is more likely to use *the dark man's hand* (64.2%) than *the hand of the dark man* (35.8%). Note, however, that in this experimental study a long possessor was invariably defined by being premodified by 2 elements, a determiner and an adjective, as in *the dark man's hand*. But what about longer possessors? Is animacy *per se* the stronger factor no matter how long the possessor is? To test for this question some additional data from the British component of

[14] Note, that only premodifying elements were considered here. As argued by Altenberg (1982), only premodification is a manifestation of weight in the sense of length (i.e. number of words) while postmodification, consisting of syntactically far more complex constructions (e.g. clauses) are rather a manifestation of syntactic complexity. In this respect, weight is defined here in terms of length.

the *International Corpus of English* (ICE-GB) was analyzed. Figures 3a and 3b shows the relative frequency of the *s*-genitive and the *of*-genitive according to the number of premodifiers to the possessor, Figure 3a for human possessors, and Figure 3b for inanimate possessors.[15]

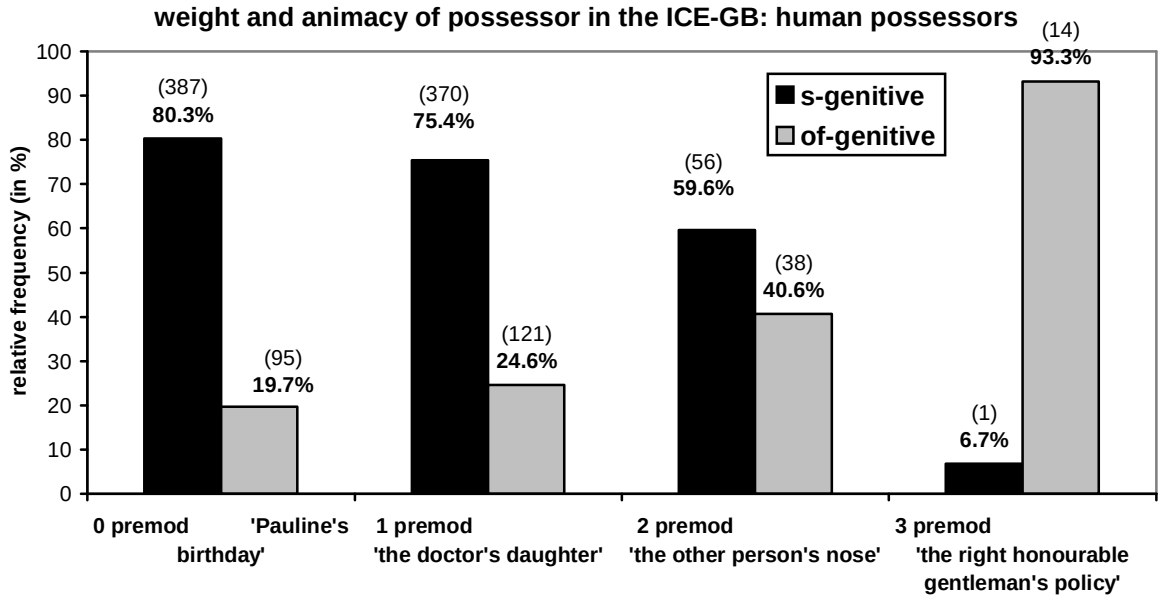

Figure 3a: Weight and animacy of the possessor in the ICE-GB (absolute number of tokens indicated in brackets above each column): human possessors

[15] As in the experimental study, only genitive constructions where both the *s*-genitive and the *of*-genitive could be used were considered here. Also, only premodification was considered and any postmodification was left out. To control for relative weight, premodified heads were systematically excluded. Note also, that only a subcorpus of all possessive NPs in the ICE-GB was considered here, i.e., definite possessive NPs where the possessor was either a proper noun or definite common noun. For further details on the data set and the analysis, see Rosenbach (2003).

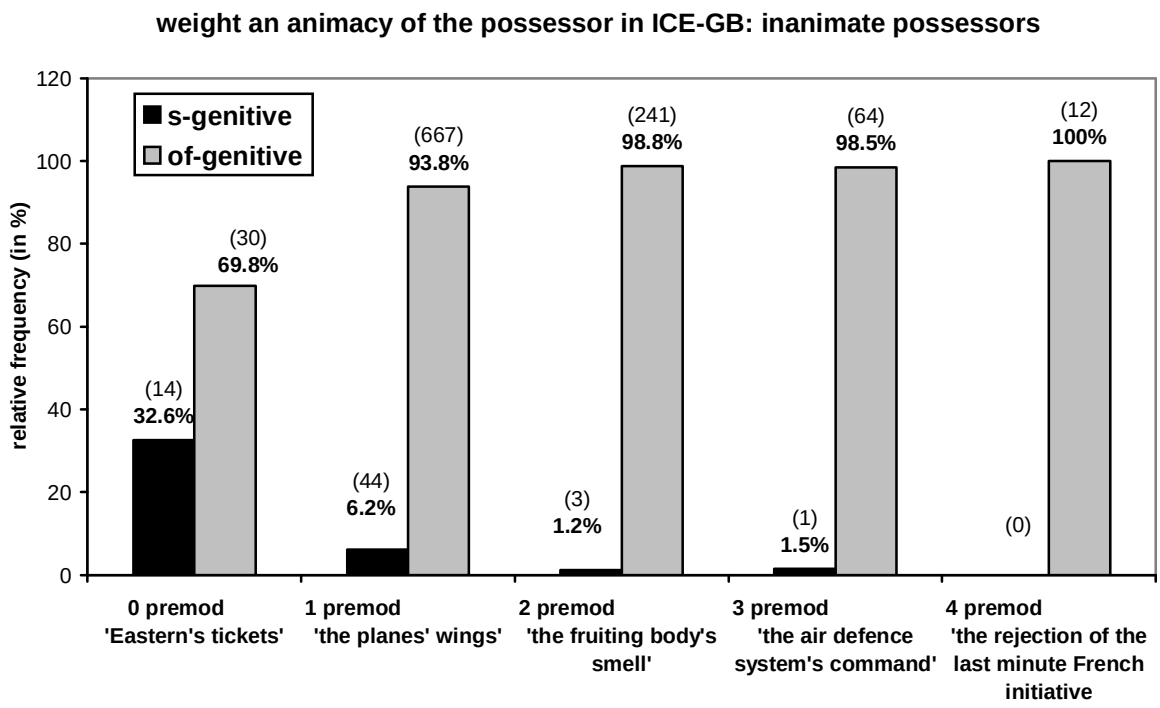

Figure 3b: Weight and animacy of the possessor in the ICE-GB (absolute number of tokens given in brackets above each column): inanimate possessors

First, we can notice that the *s*-genitive becomes less frequent the longer the possessor is, for both human and inanimate possessors. For inanimate possessors, the *s*-genitive is always less frequent than the *of*-genitive, no matter how short the possessor is. For human possessors, however, it depends on the number of premodifiers whether the *s*-genitive or the *of*-genitive is preferred. For possessors premodified by up to two elements, the *s*-genitive is preferred. That is, *the other person's nose* is still more likely than *the nose of the other person*. This corresponds to the finding of the experimental study reported above, where *the dark man's hand* was preferred to *the hand of the dark man.* For any longer possessor, however, the *of*-genitive becomes very clearly the preferred choice. So, a four-word possessor as in *the right honourable gentleman's policy* is much more likely to be expressed by a corresponding *of*-genitive (*the policy of the right honourable gentleman*). Note, however, that such long premodified possessors were as such a very infrequent context in the corpus (both in the *s*-genitive as well in the *of*-genitive), and the contexts of more than 3 premodifiers were so rarely represented in the corpus that they were not quantified. However, if such possessives occur, then the *of*-genitive is by far the more frequent construction, so the same pattern holds. So, we can notice the following factor hierarchy:

- animacy > weight: for possessors premodified by up to 2 elements
- weight > animacy: for possessors premodifed by 3 or more elements

To conclude, the relative strength of animacy and weight is not absolutely fixed but depends on the particular weight of the possessor (which is defined gradually here in terms of number of words).[16] This is evidence for counting cumulativity

[16] Different definitions of weight are on the market. See however Wasow (1997, 2002) for arguing that the various definitions of weight (as e.g. number of words/phrases/nodes) reveal basically the same results. By now, such an orthographical definition of weight has become the established operational definition of syntactic weight in the literature.

Note, finally, that the preferences for the two genitives in the tested contexts for both ganging-up and counting cumulativity are not meant to be absolute but only hold for the contexts tested. We do, for example, not claim that any 3-word possessor is preferably realized by the *s*-genitive. It is well possible that for different contexts the preferences shift. So, for example, the context of a 3-word possessor might well be preferably realized by the *of*-genitive, if the possessor is indefinite (*the hand of a dark man/a dark man's hand*), or if the possessive relation is not a prototypical one (*the future of the dark man/the dark man's future*). What is crucial for the present argumentation is that in the contexts tested (which we take to be empirically valid contexts, if not covering all possible contexts) such cumulativity effects do occur – and hence need to be accounted for.

4 Comparison

The results from the last section indicate that an adequate modeling of grammatical variation requires both kinds of cumulativity. In this respect MaxEnt models seem to be better suited for this task than StOT. In this section we investigate how well these two approaches are able to model the empirical data from the last section exactly. There are learning algorithms both for StOT and for MaxEnt on the market that induce constraint rankings from corpora. The acquired constraint rankings in turn define a probability distribution, and this distribution can be compared with the empirical distribution from the experiments and the corpus study.

4.1 Ganging-up cumulativity

In the first pair of experiments, we used the results from the experimental study from Rosenbach (2002) (see Figure 1) as a training corpus. The generator thus contains eight inputs (all configurations of the three binary features animacy, topicality and possessive relation), and two outputs for each input, namely the prenominal (> *s*-genitive) and the postnominal construal (> *of*-genitive). We adopted the OT-system for the analysis of these constructions that was proposed by Aissen and Bresnan (2002). Using the technique of Harmonic Alignment, Aissen and Bresnan derive 12 constraints that are relevant here, one for each combination of an input feature with an output feature. The constraints take the form "Avoid +anim prenominal possessors" etc. We abbreviate them as "*+a/s", "*+a/of" etc.

4.1.1 Predictions of StOT

Boersma (1998) developed the Gradual Learning Algorithm (GLA), an algorithm that induces stochastic constraint rankings from a frequency distribution over the set of input-output pairs (provided the constraints are known). We simulated a training corpus by drawing 100,000 samples according to the empirical frequency distribution.[17] The GLA acquired the following constraint ranking:

*+a/s	-2.17
*+a/of	2.17
*-a/s	2.76
*-a/of	-2.76
*+t/s	-1.26
*+t/of	1.26
*-t/s	1.85
*-t/of	-1.85
*+p/s	-0.65
*+p/of	0.65
*-p/s	1.24
*-p/of	-1.24

[17] The plasticity value was 0.01, and we kept it constant. The initial value of all constraints was 0.

This constraint ranking defines a probability distribution over the possible outputs for each input. It is not possible though to determine these probabilities analytically. Therefore we used a random generator to estimate their values. The results are shown in Figure 4.

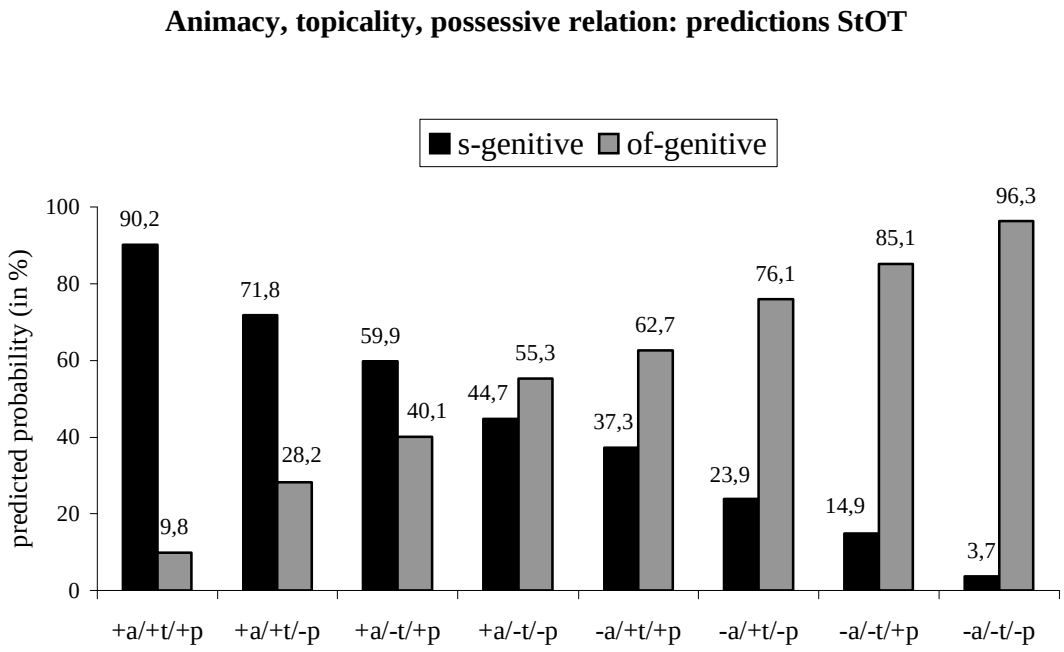

Figure 4: Animacy, topicality, possessive relation: predictions StOT

The ganging-up cumulativity pattern is very clear here. Notably, for the input combination +a/-t/-p the *of*-genitive is the preferred option, even though +a generally favors the *s*-genitive, and the animacy related constraints are each stronger than every other constraints. A comparison with Figure 1 reveals that the predictions fit the empirical data rather well. A standard tool to measure the difference between two probability distribution is the Relative Entropy (also called Kullback-Leibler distance, see for instance Cover and Thomas 1991). Here the entropy of the prediction relative to the empirical distribution is about 0.0121 bit, which is a comparatively low value.

4.1.2 Predictions of MaxEnt

There are several standard machine learning algorithms around that can be applied to induce constraint weights in a MaxEnt model. It is worth noting that Boersma's GLA can be applied almost unchanged to MaxEnt models. In the context of these models (but not in the context of StOT), the GLA belongs to the family of "Stochastic Gradient Ascent" algorithms that are frequently used in machine learning, especially for the training of neural networks (see Mitchell 1997 and the references cited therein).[18] Fischer (2005) proves that Stochastic Gradient Ascent is a proper learning algorithm for MaxEnt models in the sense that the algorithm approximates the probability distribution that is defined by the "teacher"'s grammar with arbitrary precision provided the training corpus was generated by a MaxEnt grammar.[19]

Based on the same training data and the same constraint set that were used in the previous subsection, this algorithm acquired the following constraint weights:[20]

[18]To be precise, the GLA is a stochastic gradient ascent algorithm as long as all constraints are binary. When counting constraints are used, the two algorithms differ slightly. See Jäger (2003) for a detailed discussion.
[19]A similar proof for the correctness of the GLA for StOT does not exist so far.
[20] Here the initial value of all constraint weights was set to 10.

*+a/s	9.476
*+a/of	10.524
*-a/s	10.644
*-a/of	9.356
*+t/s	9.746
*+t/of	10.254
*-t/s	10.374
*-t/of	9.626
*+p/s	9.895
*+p/of	10.105
*-p/s	10.225
*-p/of	9.775

As in the previous experiment, animacy turns out to be the strongest factor, followed by topicality and possessive relation, and for all three features, the value "+" favors the *s*-genitive and vice versa. The absolute values of the StOT model and the MaxEnt model cannot really be compared because the evaluation procedure is different.

From a vector of constraint weights in the MaxEnt setting, it is possible to determine the predicted probabilities of the different outputs relative to the inputs simply by applying the definitions. The results are given in Figure 5.

Animacy, topicality, possessive relation: predictions Maxent

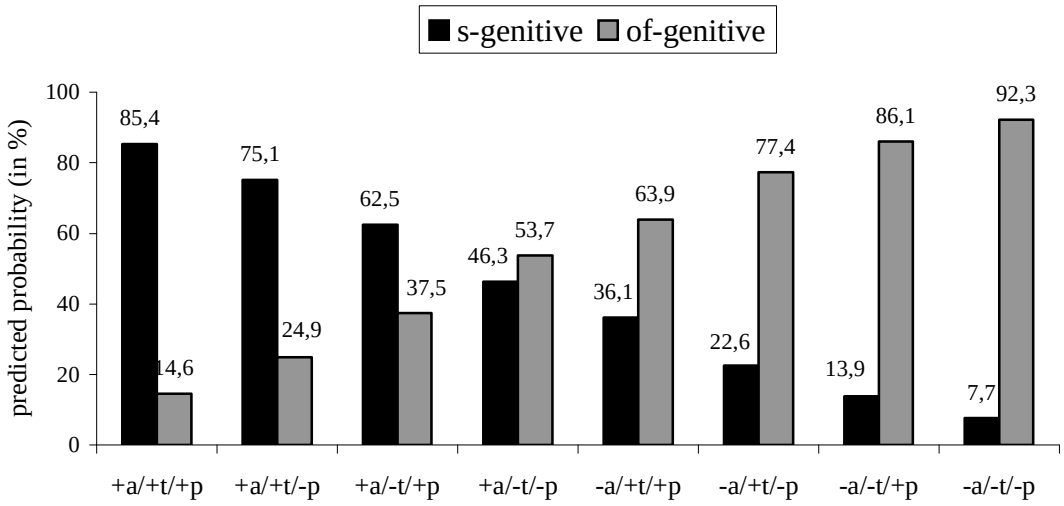

Figure 5: Animacy, topicality, possessive relation: predictions MaxEnt

Again we find a well-articulated ganging-up cumulativity pattern. The predictions fit the data even better here than in the previous experiment; the Kullback-Leibler divergence between the MaxEnt predictions and the empirical data is only 0.00472 bit.

4.2 Counting cumulativity

We conducted two analogous experiments using the counting cumulativity data from Section 3.2 (Figures 3a and 3b). Here the correlation between animacy, weight and the choice of the syntactic construction is to be modeled. There are eight possible inputs. The possessor can be either human or inanimate, and it can have 0, 1, 2 or 3 premodifiers. There are again two outputs for each input, the *s*-genitive and the *of*-genitive. The training samples were drawn at

random according to the empirical frequencies of these 16 possible patterns in the ICE-GB corpus. The correlation of animacy and the choice of genitive construction was again modeled by means of the four alignment constraints *+a/s, *+a/of, *-a/s, and *-a/of.[21] To take the potential correlation between weight and the choice of genitive construction into account, we assumed another constraint, *s, which penalizes heavy prenominal genitives. The degree of violation of this constraints depends on the weight of the possessor. Put simply, each premodifier violates this constraint once. This is illustrated in the following tableaux.

	*s
Pauline's birthday	
the birthday of Pauline	
the doctor's daughter	*
the daughter of the doctor	
the other person's nose	**
the nose of the other person	
the right honourable gentleman's policy	***
the policy of the right honourable gentleman	

4.2.1 Predictions of StOT

As in the previous experiment, the GLA was fed with 100,000 samples from the 16 possible input-output pairs according to the empirically determined probability distribution. The acquired constraint ranking was:

*+a/s	-1.580
*+a/of	1.580
*-a/s	1.804
*-a/of	-1.804
*s	-0.167

This StOT grammar corresponds to the probability distribution given in Figure 6a (for human possessors) and 6b (for inanimate possessors).

As expected from the theoretical considerations in Section 2, StOT does not show counting cumulativity. The model does distinguish between "*s is violated" and "*s is not violated", but the degree of violations is not reflected in the predictions. Accordingly, the predicted probabilities of the *s*-genitive having possessors with at least one premodifier only depend on animacy, not on weight.

[21] A reviewer questioned that a constraint based on factors such as weight/length should be part of the syntactic grammar, as it belongs to processing. This of course depends on one's conception of grammar. Note, that in the recent functional OT approach by Bresnan & Aissen (2002) constraints which should ultimately be functionally grounded are explicitly included. In this conception, therefore, there is no reason to exclude a constraint from the grammar just because it is based on a processing or performance factor.

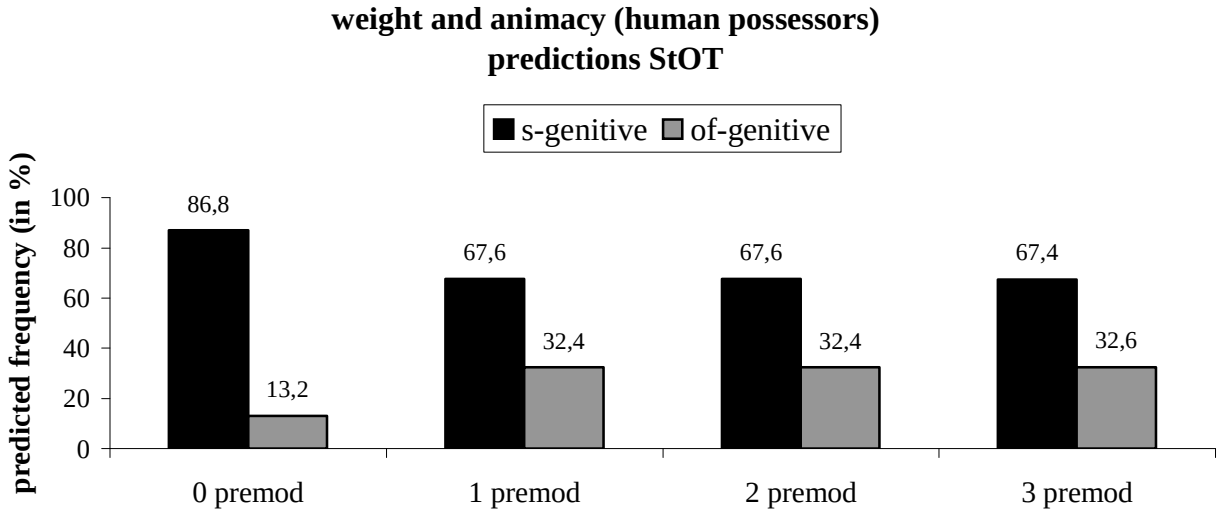

Figure 6a: Weight and animacy (human possessors): predictions StOT

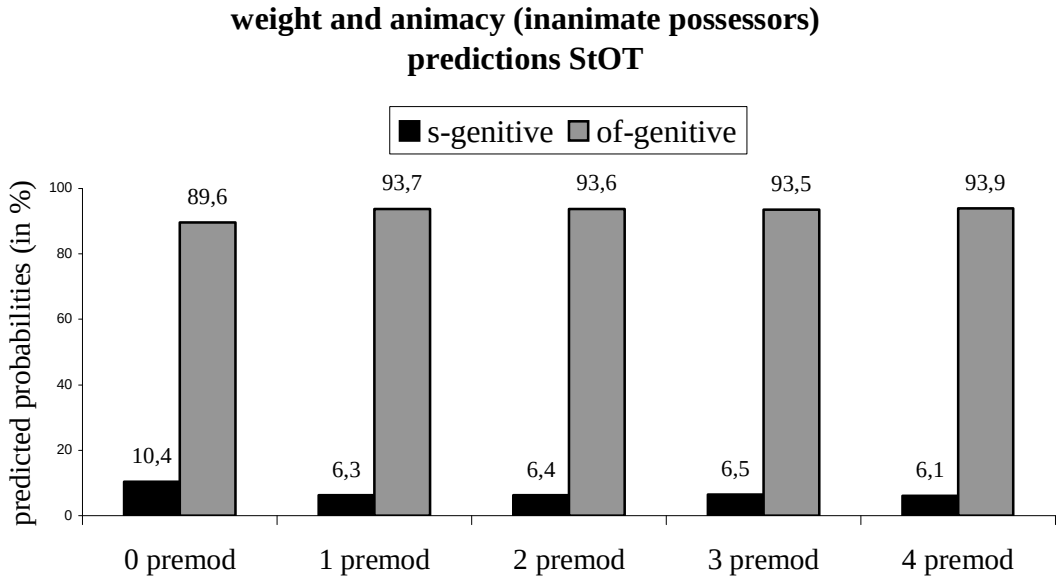

Figure 6b: Weight and animacy (inanimate possessors): predictions StOT

There is a slight variation between the values for 1, 2 and 3 premodifiers, but this is due to the fact that we obtained these probabilities by using a random generator – this variation is thus not predicted by the StOT model but it is a kind of noise. Since the data do show a dependency between weight and the choice of genitive, the fit of the model is not very good. The Kullback-Leibler distance between the model and the data is 0.0314 bit.

4.2.2 Predictions of MaxEnt

Using again the Conjugate Gradient Ascent algorithm, we obtained the following constraint weights for a MaxEnt model:

*+a/s	9.162
*+a/of	10.838
*-a/s	10.984
*-a/of	9.016
*s	0.754

This translates into the probabilities given in Figure 7a and 7b.

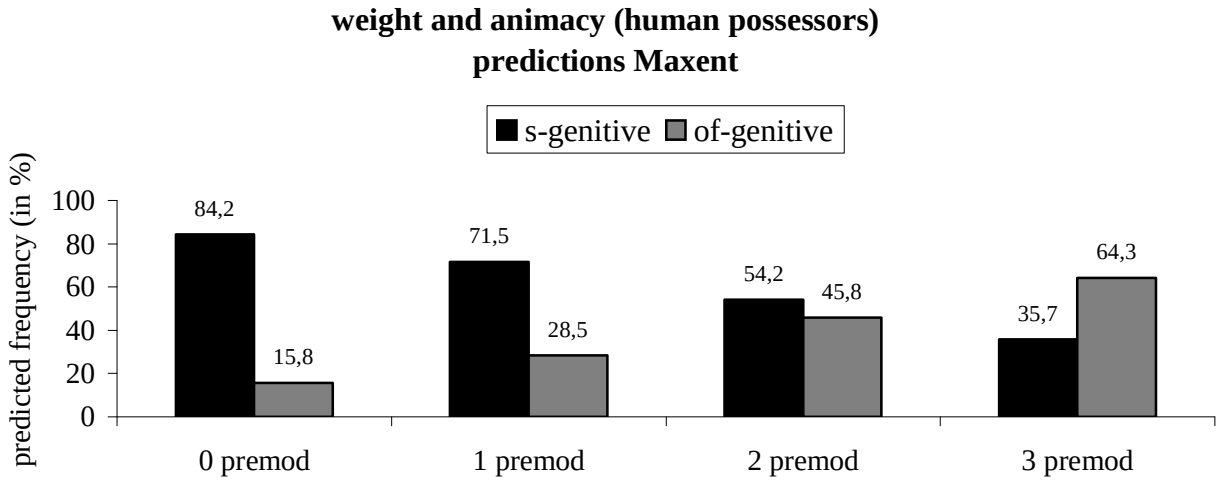

Figure 7a: Weight and animacy (human possessors): predictions MaxEnt

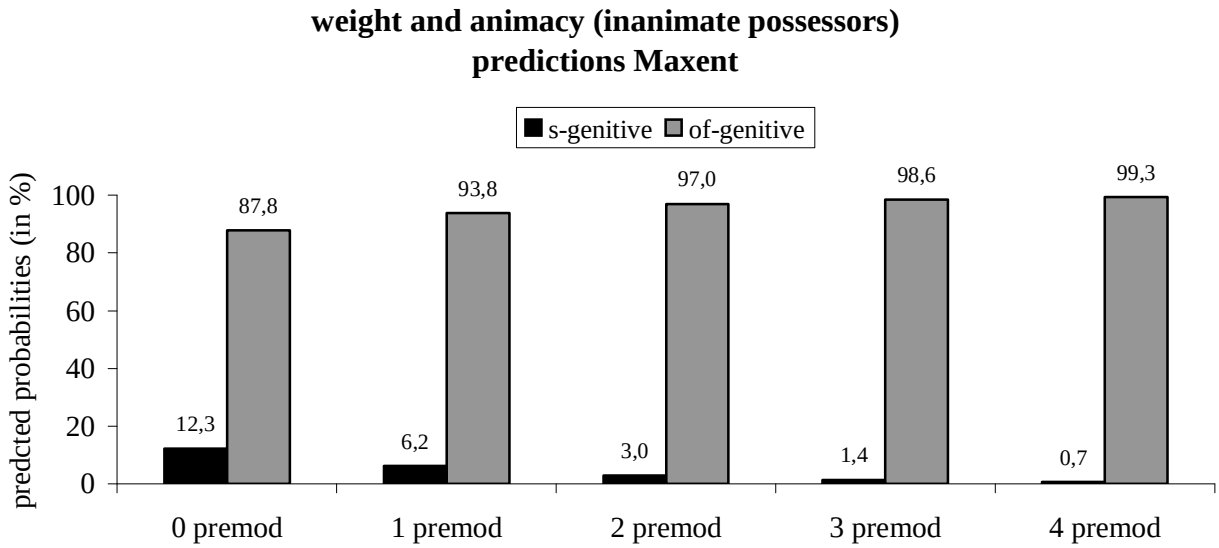

Figure 7b: Weight and animacy (inanimate possessors): predictions MaxEnt

Here we see clear counting cumulativity effects, both for human and for inanimate possessors. Especially for human possessors, the predicted cumulativity effect is actually stronger than the empirically observed one (compare Figure 3a). Nonetheless, the fit of the data is better here than for the StOT model. The Kullback-Leibler divergence between the model and the data is as low as 0.0108 bit.

4.3 Harmonic bounding

Multiple violations of a single constraint can occur in two ways. The constraint might quantify the severity of the violation – as our constraint *s in the previous subsection – or the same constraints may be violated at two positions within the same domain of evaluation. One might wonder whether the issue of counting cumulativity also arises in connection with the latter kind of multiple violations.[22] However, this question cannot be settled with regard to the particular notion of counting cumulativity that we used here with respect to the empirical

[22] Thanks to Paul Boersma (p.c.) for drawing our attention to this issue.

domain and the constraints that we considered. Recall that our definition of counting cumulativity requires two competitions that are identical except for the amount of violations of one constraint. This entails that the cardinality of the two candidate sets are identical. Multiple violations of the same constraint at two different positions would require two possessive constructions within one sentence for the constraints discussed here. This doubles the size of the candidate set. Therefore the definition is never applicable in this connection.

There is an important difference between StOT and MaxEnt though that is related to violations of the same constraint at different locations. We will briefly discuss it in this subsection.

Within OT syntax, the domain of evaluation is usually taken to be the entire sentence. Hence, if there is more than one possessive construction within one sentence, violation marks from different NPs might accumulate within the same tableaux. For concreteness, consider the following competition:

(8) a. This car's engine is louder than this car's engine.
b. This car's engine is louder than the engine of this car.
c. The engine of this car is louder than this car's engine.
d. The engine of this car is louder than the engine of this car.

Using the constraint inventory from the last subsection (but omitting those constraints that are inactive in this competition), the corresponding tableau is:

(9)

	*-a/s	*-a/of	*s
(a)	**		**
(b)	*	*	*
(c)	*	*	*
(d)		**	

In this competition, the candidates (b) and (c) are harmonically bounded. StOT thus wrongly predicts them to be ungrammatical. It should be noted that this is not an artefact of the particular constraint set that is used here. If the variation between the two constructions is governed by conflicting constraints, every strict ranking will make one of the two constructions the winner. It is thus never possible to have one construction in the subject position and the other construction in object position.

As pointed out above, MaxEnt does reserve some probability mass for harmonically bounded candidates. More precisely, the relative probability of having an *s*-genitive in subject position is predicted to be stochastically independent from the shape of the object and vice versa:

$$P(a|\{a,c\}) = P(b|\{b,d\}) = P(a|\{a,b\}) = P(c|\{c,d\})$$

We have not tested this prediction so far. It seems fair to say though that the predictions of the MaxEnt analysis are *prima facie* closer to the truth than the predictions of StOT, which amounts to some kind of non-local agreement between all possessive constructions within one sentence.

Of course, this conclusion can be avoided if the domain of optimization is a single NP rather than the entire sentence. Such an approach might work, but note that one possessive construction can be syntactically embedded into another possessive construction, as in

(10) the noise of this car's engine

Since such a construction is grammatical, the domain of optimization under StOT cannot just be the entire NP. Rather, one would perhaps need either cyclic bottom-up optimization or left-to-right optimization that applies to chunks of constituents. In either case, two evaluations, involving two different constraint rankings, would be needed to evaluate the matrix NP.

5 Conclusion and challenges

In this article we investigated the role of cumulativity in grammatical variation. We distinguished two kinds of cumulativity: 1. ganging-up cumulativity ("Every constraint matters!") and 2. counting cumulativity ("Every constraint violation matters!"). We considered several stochastic generalizations of these notions. Several empirical studies (experimental studies and a corpus study) pertaining to the grammatical variation of genitive constructions in English revealed that the weak notion of stochastic ganging-up cumulativity, as well as stochastic counting cumulativity does occur, while there was no evidence for strong ganging-up cumulativity.[23] Furthermore we compared two probabilistic generalizations of standard OT, Boersma's Stochastic OT, and Maximum Entropy models. Both models can handle weak ganging-up cumulativity. MaxEnt models can also handle counting cumulativity, while StOT cannot. Finally, we compared the empirical predictions of these two models with respect to the empirical data we considered by using standard learning algorithms. Both approaches can model weak ganging-up cumulativity effects very well, while MaxEnt is clearly superior when the data to be modeled display counting cumulativity effects.

As argued in Goldwater and Johnson (2003) and Jäger (2003), MaxEnt models are also preferable over other versions of probabilistic OT for theoretical reasons. As sketched in Section 2, the maximum entropy philosophy can be derived from basic information theoretic considerations. MaxEnt learning finds the probability distribution with the highest entropy given the empirical observations (where only constraint violation profiles can be observed). Intuitively speaking, the entropy of a stochastic process is a measure of its disorder or unpredictability. Maximizing entropy given the empirical observations thus amounts to finding a model that contains no more information (= order) than what can be derived from the data.

Last but not least, there are several provably correct learning algorithms around for MaxEnt models, while the learning problem for all other version of probabilistic OT is still open to date.

To summarize, we presented empirical arguments that both notions of cumulativity are needed to model grammatical variation adequately. Of the available probabilistic generalizations of OT, MaxEnt models implement this insight in the most natural way.

Our results and conclusions have been challenged by various proponents of OT. The main intention of this article is to make some interesting data available to the community, and a thorough discussion of the theoretical status of cumulativity goes well beyond the scope of this paper. Nonetheless we would like to state our opinion about some of the points that came up repeatedly in discussions so far.

1. *No evidence for cumulativity has been brought forward so far - this is an isolated phenomenon*. - True, we are – to the best of our knowledge – the first to put forward such evidence in the context of OT syntax, but similar cut-off points as to the length of constituents have also been reported for word order phenomena in the English verb phrase, as for example by Hawkins (2002) for postverbal complement and adjunct ordering, and by Gries (2003) and Lohse et al. (2004) for particle placement in verb-particle constructions. Quinn (2004) has argued that there is both ganging-up and

[23] See also Keller (2000), who found evidence for cumulativity with respect to graded grammaticality judgments.

cumulativity for pronoun conjunction in English. We believe that the major reason for the lack of evidence for cumulativity is lack for looking for it (at the right places). Until recently, variation of the more-or-less sort was outside the OT framework, and the first OT work on variation was to be found in the field of phonology (see e.g. Anttila 1997). Only very recently, OT approaches have begun to capture such variation in the field of morphosyntax (see Bresnan & Aissen 2002 for a programmatic sketch). So, why should anyone start looking for a phenomenon before it applies to the theoretical framework? There is a body of empirical work on grammatical variation, but so far such studies have been largely restricted to functionalist and/or sociolinguist work, and these fields presumably did not worry about the theoretical implications of their work for OT (functionalists and sociolinguists will most likely not have been aware of it at all).

2. *MaxEnt models are basically a version of Harmonic Grammar. The factorial typology that is predicted by HG is much more liberal than the predictions of OT, and the available evidence suggest that OT is closer to the truth. So moving from StOT to MaxEnt might be supported by the particular data you consider, but it leads to an overall explanatory loss that is not justified by a single study.* This criticism involves a *non sequitur*. The argument against cumulativity based on factorial typology (see for instance Legendre et al. 2005) applies to categorical data (and to our knowledge only to phonology – the issue is actually open for syntax, but this is not our concern here), while our investigation deals with quantitative data. These are different issues. The mentioned article (Legendre et al. 2005) contains a lucid discussion of the pros and cons of HG versus OT. In this connection the authors write:

> "One possibility is this. Knowledge relevant to language processing may combine (i) a system of constraints one might consider more strictly 'grammatical', interacting exclusively or primarily via strict domination, with (ii) a set of more pragmatically-based constraints, reflecting more directly, perhaps, statistical characteristics of experience, and interacting in a less restricted manner, via arbitrarily weighted constraints. The process of *grammaticalization* may be one in which constraints effectively move from the latter category to the former. The constraints interacting in the HG analysis may constitute a mixture of both types of constraints, while the constraints focused upon in most OT studies may be more completely contained in the 'grammatical' class."

   We agree that it is very well possible that strict domination holds for categorical data, while quantitative data display both kinds of (weak) cumulativity. So the argument of explanatory strength does not necessarily carry over from categorical to stochastic models.

   We do think, furthermore, that quantitative data are important, and if there is any worth in the postulation of falsifiable theories, then empirical evidence should be taken seriously. It has recently been repeatedly stressed by linguists working within a formal theoretical framework that syntactic theory should be solidly based on empirical evidence, as evidenced in various conferences/workshops on empirical linguistics/syntax in the past two or three years. We regard our work in line with such claims.

3. *Counting cumulativity can always be avoided by binarizing constraints.* We disagree for two reasons.
   - **Theory parsimony:** If two theories are identical in their empirical predictions, but the one avoids counting cumulativity while the other has fewer parameters, Occam's razor favors the second one. This, applies *ceteris paribus* also to ganging-up cumulativity, which can be simulated by a non-cumulative model by using constraint conjunction. However, this technique proliferates the number of constraints as well. So while StOT or similar models could be "tuned" to handle cumulativity, MaxEnt can do the same in an arguably more parsimonious way.
   - **Restrictiveness:** One might counter this argument by saying that admitting counting constraints as such constitutes a heavy complication of the theory, so that binarizing constraints might be the smaller price to pay. This point of view has been defended by

McCarthy (2003) as well as by Paul Boersma (p.c.). However, a single *n*-ary constraint leads to more restrictive empirical predictions than *n* binary constraints. To stick to our example, our MaxEnt analysis predicts that the probability of the *of*-genitive **strictly** increases with the weight of the possessor.[24] Using binary constraints instead only predicts weak monotonicity. Such a model would admit the existence of languages where the probability of the post-nominal genitive is constant at 10% up to 5 premodifiers, jumps to 60% starting with 6 premodiers, remains constant until 17 premodifers and approaches 100% for higher values. In fact, McCarthy uses similar considerations to show that constraint evaluation must never involve counting of any sort. Our data indicate though that this is not a viable option if one wants to capture quantitative effects.

4. *StOT is cognitively more realistic than MaxEnt, whatever the mathematical merits of the latter model may be.* We disagree. There are two versions of this argument that we are aware of, and we think that they are both invalid.
   - *Boersma and Levelt (2000) present a study where the acquisition of Dutch syllable structure was simulated with StOT and the GLA. It turned out that the order of acquisition of different syllable types in the simulation matches the order in which Dutch infants acquire these structures.* This is in fact a very relevant result. However, in Jäger (2003) the experiment was replicated using MaxEnt and Stochastic Gradient Ascent (which is, as mentioned above, virtually identical to the GLA except that it applies to MaxEnt rather than to StOT). The findings of Boersma and Levelt were almost identically replicated, so this does not help to distinguish between the two models
   - *StOT, as a version of OT, is related to connectionist models and therefore indirectly to the structure of our brain, while MaxEnt is a pure data fitting device.* It is actually HG that has a connectionist foundation. (Categorical) OT-models can be seen a special case of HG models where constraint weights grow exponentially. The founders of OT are always very careful to point out that strict domination, i.e., the exponential growth of constraint weights, has no connectionist explanation so far.[25] So if the connectionist foundation of HG/OT is taken as evidence for neurophysiological plausibility, then the case for HG is actually stronger than the case for OT, because the latter restricts the class of underlying networks in a (neurophysiologically!) unmotivated way. MaxEnt, as a probabilistic version of HG, is thus at least as cognitively plausible as StOT.

Acknowledgements
The material from this article has been presented before at the University of Nijmegen and at Stanford University. We are grateful for the feedback we got at these occasions, as well as for the comments from two anonymous reviewers for *Linguistics*, and for discussions with Reinhard Blutner, Paul Boersma, Joan Bresnan, Paul Smolensky, Ralf Vogel and Henk Zeevat.

References

Abney, Steven (1997). Stochastic attribute-value grammars. *Computational Linguistics* 23(4), 597-618.
Aissen, Judith and Bresnan, Joan (2002). *Optimality Theory and Typology*, Course material of the Summer School on Formal and Functional Linguistics, Düsseldorf. (http://www.phil-fak.uni-duesseldorf.de/summerschool2002/CDV/CDAissen.htm)
Altenberg, Bengt (1982). *The Genitive v. the Of-Construction. A Study of Syntactic Variation in 17th Century English*. Malmö: CWK Gleerup.

[24] The prediction of the MaxEnt analysis is actually much stronger: the probability of the post-nominal construal must grow according to the **logistic function** $a \times e^w/(b+e^w)$ (for some constant parameters *a* and *b*) depending on the weight *w* of the possessor.

[25] For instance in Prince and Smolensky (1997:1608): „That strict domination governs grammatical constraint interaction is not currently explained by principles of neural computation; nor do these principles explain the universality of constraints that is central to optimality theory and related approaches. These are stimulating challenges for fully integrating optimality theory with a neural foundation."

Anschutz, Arlea (1997). How to choose a possessive noun phrase construction in four easy steps. *Studies in Language* 21 (1): 1-35.
Anttila, Arto (1997). Deriving variation from grammar.In: Frans Hinskens, Roeland van Hout, and Leo Wetzels (eds.). *Variation, Change, and Phonological Theory*, 35-68. Amsterdam/Philadelphia: John Benjamins.
Behaghel, Otto. 1909/10. Beziehungen zwischen Umfang und Reihenfolge von Satzgliedern. *Indogermanische Forschungen* 25.110-142.
Berger, Adam; Della Pietra, Stephen; and Della Pietra, Stephen (1996). A maximum entropy approach to natural language processing. *Computational Linguistics* 22(1), 39-71.
Biber, Douglas; Johansson, Stig; Leech, Geoffrey; Conrad, Susan; and Finegan, Edward (1999). *Longman Grammar of Spoken and Written English*. London/New York: Longman.
Boersma, Paul (1998). Functional Phonology. Ph.D. thesis, University of Amsterdam.
Boersma, Paul & Hayes, Bruce (2001). Empirical tests of the Gradual Learning Algorithm. *Linguistic Inquiry* 32(1), 45-86.
Boersma, Paul & Clara Levelt (2000). Gradual constraint-ranking learning algorithm predicts acquisition order. in *Proceedings of Child Language Research Forum 30*, Stanford: CSLI, pp. 229-237.
Cover, Thomas M.; and Thomas, Joy A. (1991). *Elements of Information Theory*. New York: Wiley.
Della Pietra, Stephen; Della Pietra, Vincent; and Lafferty, John (1995). Inducing features of random fields. CMU Technical Report CMU-CS-1995-144, Carnegie Mellon University.
Fischer, Markus (2005). *A Robbins-Monro type learning algorithm for an entropy maximizing version of stochastic Optimality Theory.* Master's Thesis. Humboldt University Berlin. In preparation.
Goldwater, Sharon; and Johnson, Mark (2003). Learning OT constraint rankings using a maximum entropy model. In *Proceedings of the Stockholm Workshop on Variation within Optimality Theory*, Jennifer Spenader, Anders Eriksson, and Östen Dahl (eds.), 111-120. Stockholm University.
Gries, Stefan. 2003. *Multifactorial analysis in corpus linguistics*. New York/London: Continuum.
Hawkins, John A. 2002. Gradedness as relative efficiency in the processing of syntax and semantics. Paper presented at the Gradedness Conference on ?-??-???-*?, October 21st – 23rd, 2002, University of Potsdam.
Hawkins, John A. (1994). *A Performance Theory of Order and Constituency*. Cambridge: CUP.
Huddleston, Rodney (1984). *Introduction to the Grammar of English*. Cambridge: Cambridge University Press.
Huddleston, Rodney & Geoffrey Pullum (2002). *The Cambridge Grammar of the English Language*. Cambridge: CUP.
Jäger, Gerhard (2003). Maximum entropy models and Stochastic Optimality Theory. Manuscript, University of Potsdam. ROA 625-1003.
Johnson, Mark (1998) Optimality-theoretic Lexical Functional Grammar. Manuscript, Brown University
Jucker, Andreas (1993). The genitive versus the of-construction in newspaper language. In *The Noun Phrase in English. Its Structure and Variability*, Andreas Jucker (ed.), 121-136. Heidelberg: Carl Winter.
Keller, Frank (2000). Gradience in Grammar: Experimental and Computational Aspects of Degrees of Grammaticality. Ph.D. Thesis, University of Edinburgh.
Kirby, Simon (1999). *Function, Selection, and Innateness. The Emergence of Language Universals*. Oxford: OUP.
Koptjevskaja-Tamm, Maria (2001). Adnominal possession. In *Language Typology and Language Universals* (Handbooks of Linguistics and Communication Science 20.1,2). Vol. II, Martin Haspelmath, Ekkehard König, Wulf Oesterreicher and Wolfgang Raichle (eds.), 960-970.
Leech, Geoffrey; Francis, Brian; and Xu, Xunfeng (1994). The use of computer corpora in the textual demonstrability of gradience in linguistic categories. In *Continuity in Linguistic Semantics*, Catherine Fuchs and Bernard Victorri (eds.), 57-76. Amsterdam/Philadelphia: John Benjamins.
Legendre, Geraldine; Miyata, Yoshiro; and Smolensky, Paul (1990). Harmonic Grammar – A formal multilevel connectionist theory of linguistic well-formedness: Theoretical foundations. In *Proceedings of the Twelfth Annual Conference of the Cognitive Science Society*, 388-395, Cambridge, MA: Erlbaum.

Legendre, Géraldine, Sorace, Antonella & Smolensky, Paul (2005). The Optimality Theory – Harmonic Grammar connection. Chapter 20 of Smolensky & Legendre (2005).
Lohse, Barbara, John A. Hawkins, and Thomas Wasow. 2004. Domain Minimization in English verb-particle constructions. *Language* 80.2. 238-261..
Lyons, Christopher (1999). *Definiteness*. Cambridge: Cambridge University Press.
Lyons, Christopher (1989). "Phrase structure, possessives and definiteness". *York Papers in Linguistics* 14: 221-228.
McCarthy, John (2003), OT Constraints are Categorical. *Phonology* 20(1), pp 75-138.
Mitchell, Tom (1997). *Machine Learning*. New York: McGraw-Hill.
Prince, Alan & Smolensky, Paul (2004/1993). *Optimality Theory: Constraint interaction in generative grammar*. Oxford: Blackwell, 2004. previously distributed as Technical Report TR-2, Rutgers Center for Cognitive Science, Rutgers University, New Brunswick, NJ, 1993.
Prince, Alan & Smolensky, Paul (1997). Optimality: From neural networks to universal grammar. *Science* 275: 1604-1610.
Quinn, Heidi (2004). "Testing and modelling the distribution of pronoun case forms in English". Paper presented at *Approaches to Empirical Syntax/WOTS-8*, ZAS Berlin, August 27-29, 2004.
Quirk, Randolph; Greenbaum, Sidney; Leech, Geoffrey; and Svartvik, Jan (1985). *A Comprehensive Grammar of the English Language*. London/New York: Longman.
Rosenbach, Anette (2003). Comparing animacy vs. weight as determinants of grammatical variation in English. Manuscript, University of Düsseldorf.
Rosenbach, Anette (2002). *Genitive Variation in English. Conceptual Factors in Synchronic and Diachronic Studies.* Berlin/New York: Mouton de Gruyter.
Smolensky, Paul & Geraldine Legendre (2005) *The Harmonic Mind: From Neural Computation To Optimality-Theoretic Grammar. Vol. 2, Linguistic and philosophical implications.* MIT Press. in press.
Taylor, John (1996). *Possessives in English*. Oxford: Clarendon.
Wasow, Thomas (2002). *Postverbal Behavior*. Stanford: CSLI Publications.
Wasow, Thomas (1997). Remarks on grammatical weight. *Language Variation and Change* 9, 81-105.
Wedgwood, Daniel (1995). Grammaticalisation by Re-analysis in an Adaptive Model of Language Change: A Case Study of the English Genitive Constructions. Master's Thesis, University of Edinburgh.
Woisetschlaeger, Ernst (1983). "On the question of definiteness in 'an old man's book". *Linguistic Inquiry* 14(1): 137-154.