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

Gerhard Jäger & Anette Rosenbach

Robert Munro, for

LING 289: Quantitative and Probabilistic Explanations in Linguistics.

03 Dec 2007

Synopsis

The paper argues that Maximum Entropy (MaxEnt) models are preferable to Stochastic Optimality (StOT) models, as MaxEnt models allow low-ranked constraints to 'gang-up' on high ranked constraints. That is, they allow cumulativity.

In addition to *ganging-up cumulativity* the authors distinguish *counting cumulativity*. Counting cumulativity is simply being sensitive to the number of violations of a single constraint. In a sense it is no different to ganging-up cumulativity, simply allowing a higher ranked constraint to be ganged-up on by multiple violations of the *same* lower-ranked constraint. Many of the arguments follow from existing comparisons of MaxEnt and StOT made by Goldwater and Johnson (2003). The authors give a worked example modeling English genitive variation, demonstrating that MaxEnt models give a better account of the observed data than StOT.

Background: Stochastic OT

Researchers have generally found that StOT is better than standard OT in predicting the relative frequency of the outcomes in observed data.[1]

StOT is similar to standard OT, but instead of there being hard-divisions between constraints, the ranking of constraints is defined by normal distributions on a continuum. Instead of one constraint outranking another in 100% of cases, as in standard OT, it will outrank another constraint *p*% of the time, where *p* is determined by the degree to which the two distributions intersect. In order words, StOT extends OT by defining a *probability of outcomes*, for each ranking, not just a single dominant outcome.

[1] But not always, cf: Paul Kiparsky (2005).

(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	*		

If we have data where we observe d_2 with more frequency than d_1 in the context of violations of c_2 and c_3, we need a probabilistic model of constraints that has the possibility of predicting this.[2] The authors demonstrate that MaxEnt models are a good way to achieve this.

Maximum Entropy models

MaxEnt models are also known as log-linear models. They differ mostly from what we've seen in class in the terminology:

- *Bias.* The amount by which a model differs from the observed data. The *least biased* of all possible models is therefore the best fit.
- *Entropy.* An information theoretic notion that quantifies the bias. The entropy H of a probability distribution p is defined as:

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

For all intents and purposes they are using logistic-regression, but note the higher the entropy, the lower the bias.

They set up the models as follows:

- Each feature represents a constraint; with each value the number of observed violations.
- The 'rank' of a constraint is the weight given to that feature after the model has been fit.

[2] The authors note, citing pc from Paul Boersma, that standard OT could be extended to allow d_2 to be the winner by simply modeling that c_2 and c_3 combined outrank c_1. The authors call this *strong cumulativity*, as opposed to the cumulativity implemented in the paper which is *weak cumulativity.* Strong cumulativity entails the weak.

English genitive variation

The worked example in the paper is on English genitive variation, looking at the various factors that contribute to it:

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

They demonstrate that while animacy is the most important factor, the others factors can interact. Referring to earlier work, they looked at the weight of the NP, but needed to tease NP-weight apart from animacy, as the two can correlate. They find that the relative strength of animacy and weight is not absolute, but depends on the NP-weight of the possessor. They define NP-weight as the number of pre-modifiers, hence it is an example of counting cumulativity, that is, every pre-modifier is modeled as a violation of a NP-weight constraint against modifiers for prenominal genitives:

	*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	

They build models of the data using both MaxEnt and StOT, comparing the two to the observed data using the Kullback-Leibler distance. The results show that MaxEnt is a (slightly) better fit. The main set of results are on the next page, where Figure 1 is the observed distribution, Figure 4 the predictions of StOT, and Figure 5 the predictions of MaxEnt.

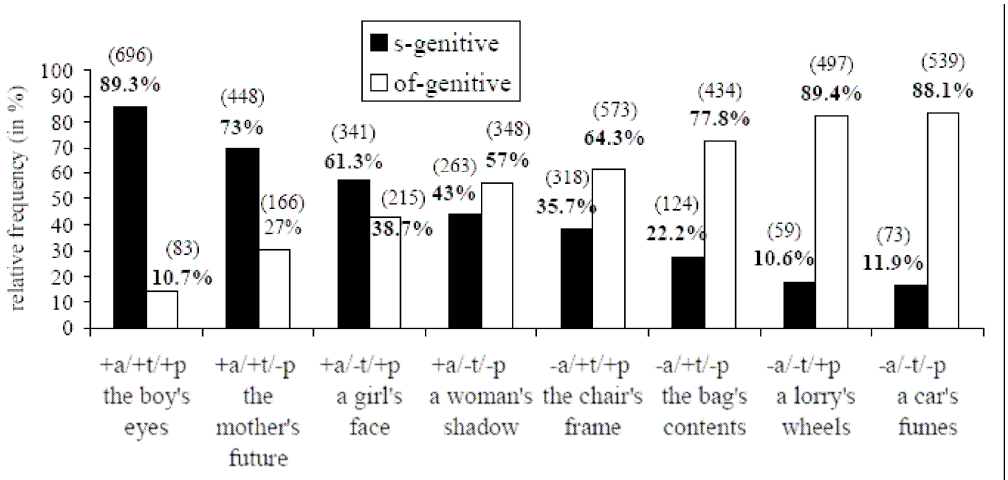

Figure 1: Animacy, topicality, and possessive relation – results of experimental study, British subjects (n=56), absolute number of token given in brackets above each column

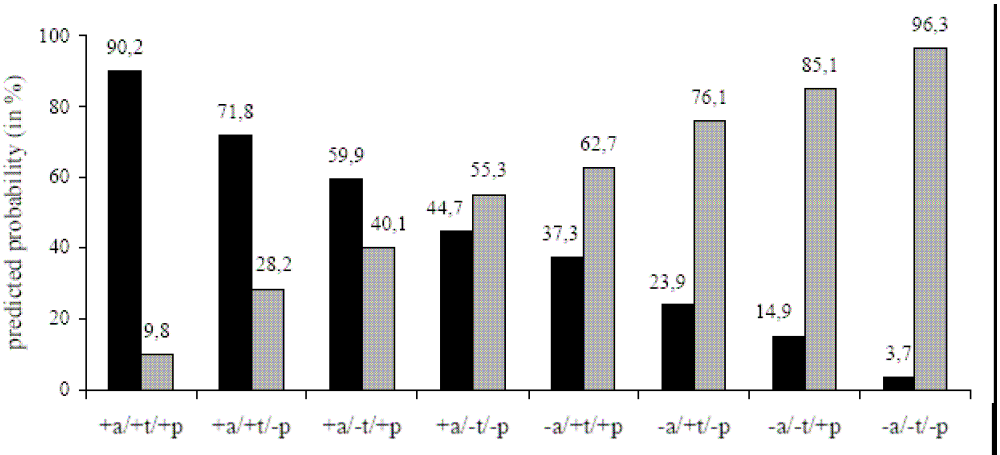

Figure 4: Animacy, topicality, possessive relation: predictions StOT

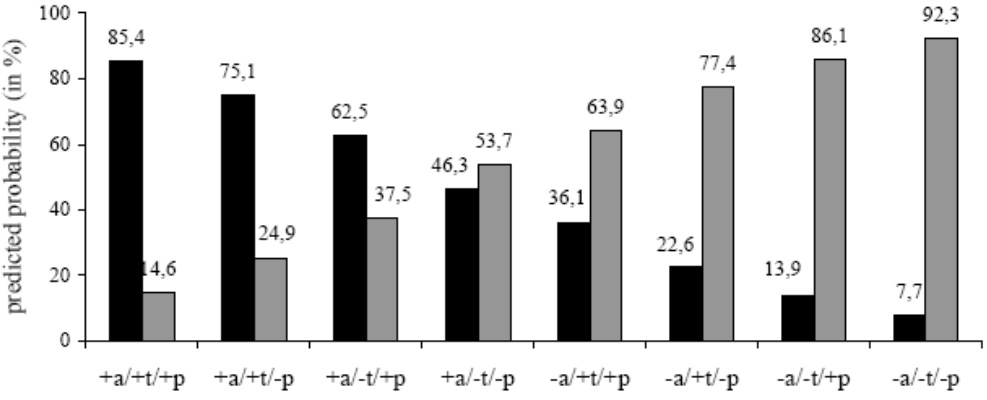

Figure 5: Animacy, topicality, possessive relation: predictions MaxEnt

Conclusions

The paper concludes with a reply to some criticisms of MaxEnt modeling that presumably originated in the StOT community:

1. *No evidence for cumulativity has been brought forward so far - this is an isolated phenomenon*.
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…*
3. *Counting cumulativity can always be avoided by binarizing constraints.*
4. *StOT is cognitively more realistic than MaxEnt, whatever the mathematical merits of the latter model may be*.

To which the authors answer:

1. True.
2. Only for categorical data.
3. True, but the MaxEnt model is simpler and you don't need to make real/integer values categorical.
4. They're equally realistic.

The paper deserves to bring more 'converts' to logistic-regression than it will probably generate (if that is its primary goal). By focusing on a fairly narrow set of syntactic features the authors aren't selling the potentially exciting scope of allowing cumulativity. There are many contextual influences that could be modeled. For example, the existing feature of topicality could be extended to any number of similar pragmatic conditions that would contribute to the outcome non-deterministically.

Some notes for discussion

Feature interaction

The authors demonstrate that a MaxEnt model is a further relaxation of standard OT. But they've relaxed the advantages of a hierarchy of constraints right out the window. For all its shortcomings, standard OT *does* model some feature interaction.[3] To adapt the terminology of the authors, standard OT allows strong cumulative *dominance*, but MaxEnt does not. Consider:

1. There is an outcome o_1 that is always observed when a constraint a_1 is inviolate.
2. For a separate task, there is an outcome o_2 that is always observed two or more constraints, $b_1 \ldots b_n$, are inviolate, *but not when most subsets of* $b_1 \ldots b_n$ *are inviolate*.

Standard OT can model both *1* and *2* perfectly by ranking these constraints highest.

Both MaxEnt and StOT will model *1* perfectly by ranking o_1 with near-100% probability in the context of a_1. But MaxEnt will model *2* by distributing weights equally among $b_1 \ldots b_n$, allowing any lower ranked constraints to factor into the predicted outcomes, therefore under-generating outcome o_2.[4]

Many linguists are already modeling cumulativity!

The authors note that functionalist, sociolinguists and connectionist linguistics have been looking at variation and feature interaction for some time, so why not compare MaxEnt to the existing models?[5]

Researchers in computational linguistics working with machine learning are also utilizing models that allow cumulativity, and if the algorithms relax the attribute independence assumption they are potentially more powerful than MaxEnt, albeit intractable to optimize.

---

[3] To the extent that feature interaction can be represented as a hierarchy. I'm uncertain whether StOT's hierarchical properties can be considered to model feature interaction …probably.

[4] It's possible that this is why MaxEnt under-generates the dominant outcome for +a/+t/+p in Figure 4. This isn't an indictment of MaxEnt, just a caveat that it needs to be further extended to allow the more complex interaction of features.

[5] As a means of distancing themselves from this literature, the authors note "[functionalist and/or sociolinguist work] presumably did not worry about the theoretical implications of their work for OT". This seems to be especially true of the functionalist and sociolinguistic work of the 60's, 70's, 80's and early 90's.