Cumulativity in Variation: testing different versions of Stochastic OT empirically

Workshop on Optimality Theoretic Syntax 7
October 27-28, 2003, Katholieke Universiteit Nijmegen

Gerhard Jäger
University of Potsdam
jaeger@ling.uni-potsdam.de
www.ling.uni-potsdam.de/~jaeger/
and
Anette Rosenbach
University of Düsseldorf
ar@phil-fak.uni-duesseldorf.de

1. Overview

- cumulativity and stochastic OT
  - ganging-up cumulativity
  - counting cumulativity
- floating constraints and maximum-entropy models
- empirical evidence for cumulativity: the syntax of English genitives
- comparison: how deal the two theories with the data?
- conclusion

2. Cumulativity and stochastic OT

- basic assumption of OT: *The winner takes it all*
  - Once a competition is decided, lower-ranked constraints play no role, and
  - it plays no role how high the winner wins.
- several stochastic generalizations of OT on the market
- how do they deal with cumulativity?

Ganging-up cumulativity

- question: can dominated constraints have an impact on probability of a candidate?
- for instance:

	c1	c2	c3
a1	*		
b1		*	

	c1	c2	c3
a2	*		
b2		*	*

$$P(b1) > P(b2) \ ?$$

Counting cumulativity

- question: can numerical amount of constraint violations have an impact on probability of a candidate?
- for instance:

	c1
a1	*
b1	

	c1
a2	***
b2	

$$P(b2) > P(b1) \ ?$$

- “partial ranking” approach (Anttila) and “floating constraints” approach (Boersma) agree
  - ganging-up cumulativity exists
  - counting cumulativity does not exist
- alternative: Maximum Entropy (MaxEnt) models

3. Maximum Entropy models

- state of the art in machine learning and computational linguistics (Della Pietra et al. 1996, Abney 1997)
- very similar to OT, and even more similar to Harmony Grammar (see Goldwater and Johnson 2003)
- derived from first principles: best hypothesis must
  - confirm with the data (the average degree of violations for each constraint), and
  - given this, be as unbiased as possible, i.e.
  - have the maximal entropy

- each constraint has numerical weight
- probability of a candidate is proportional to its exponentiated harmony:

$$H(a) \ = \ \sum_i w_i \cdot -c_i(a)$$

$$P(a) \ \sim \ \exp(H(a)$$

$$P(a) \ = \ \frac{\exp(H(a))}{\sum_{a'} \exp H(a')}$$

- predicts both kinds of cumulativity

4. Comparison

4.1. Ganging-up cumulativity

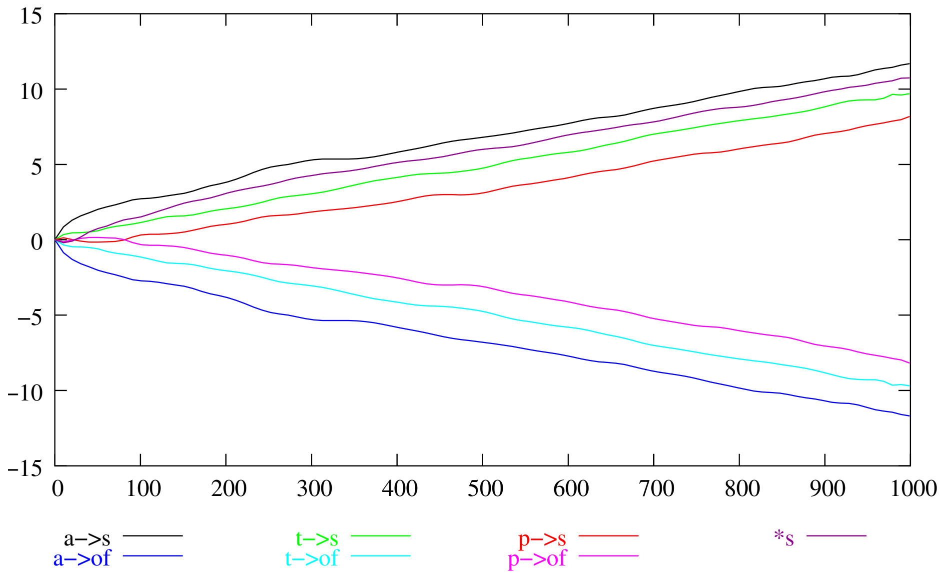
- seven constraints:

  1. animate possessor ⇒ s-genitive (a->s)
  2. animate possessor ⇒ of-genitive (a->of)
  3. topical possessor ⇒ s-genitive (t->s)
  4. topical possessor ⇒ of-genitive (t->of)
  5. prototypical possessor ⇒ s-genitive (p->s)
  6. prototypical possessor ⇒ of-genitive (p->of)
  7. avoid s-genitives (*s)

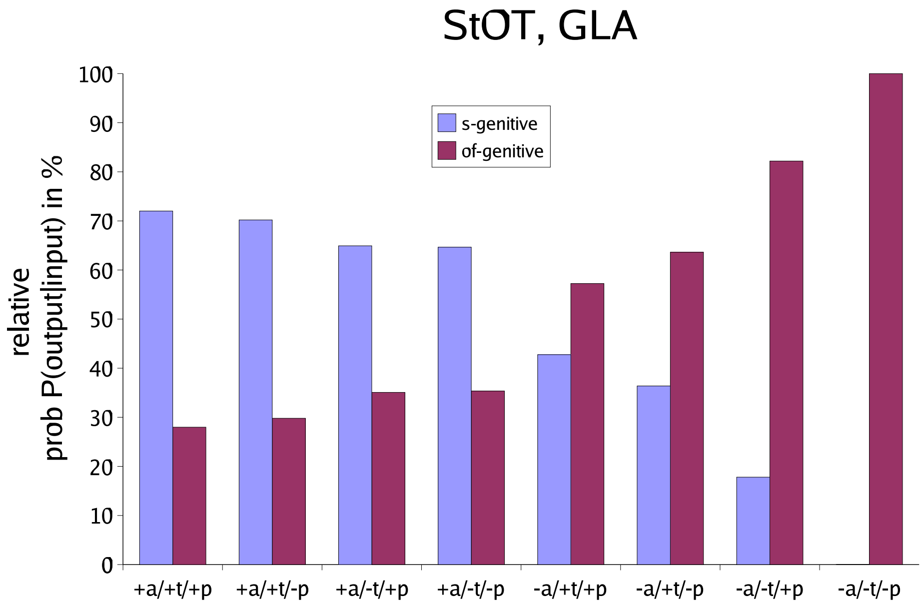
Stochastic OT

- evaluator component: Stochastic OT in the sense of Boersma 1998
- Learning algorithm: Gradual Learning Algorithm
- acquired grammar:

a->s	11.69
a->of	-11.69
t->s	1.69
t->of	-9.69
p->s	8.18
p->of	-8.18
*s	10.74



Predicted relative probabilities

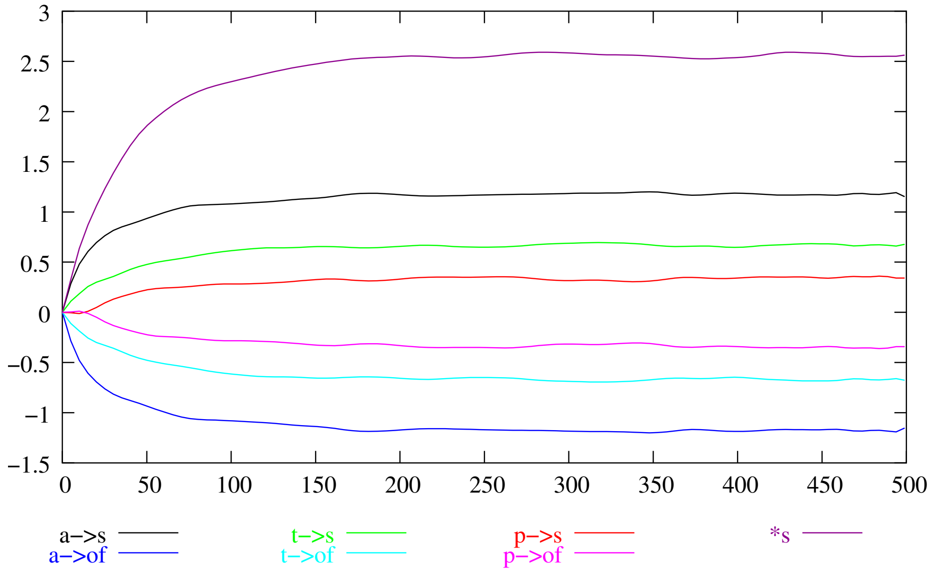


KL-divergence between predicted and observed probabilities: 0.0576

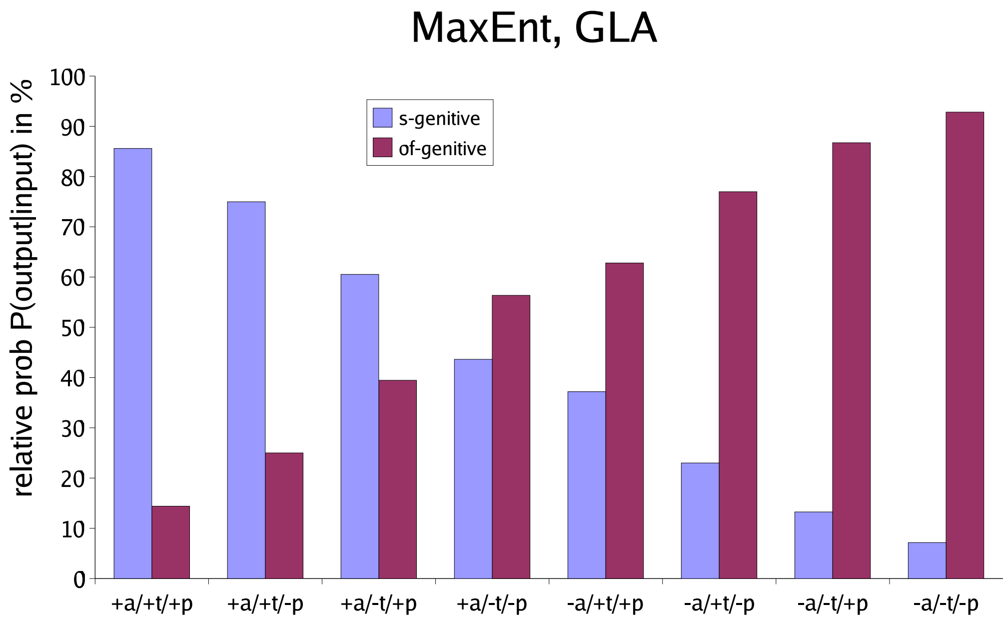
MaxEnt

- evaluator component: log-linear probabilities (proportional to exponentiated harmony)
- Learning algorithm: Gradual Learning Algorithm (= Stochastic Gradient Ascent)
- acquired grammar:

a->s	1.153
a->of	-1.153
t->s	0.677
t->of	-0.677
p->s	0.342
p->of	-0.342
*s	2.562



Predicted relative probabilities



KL-divergence between predicted and observed probabilities: 0.0002

4.2. Counting cumulativity

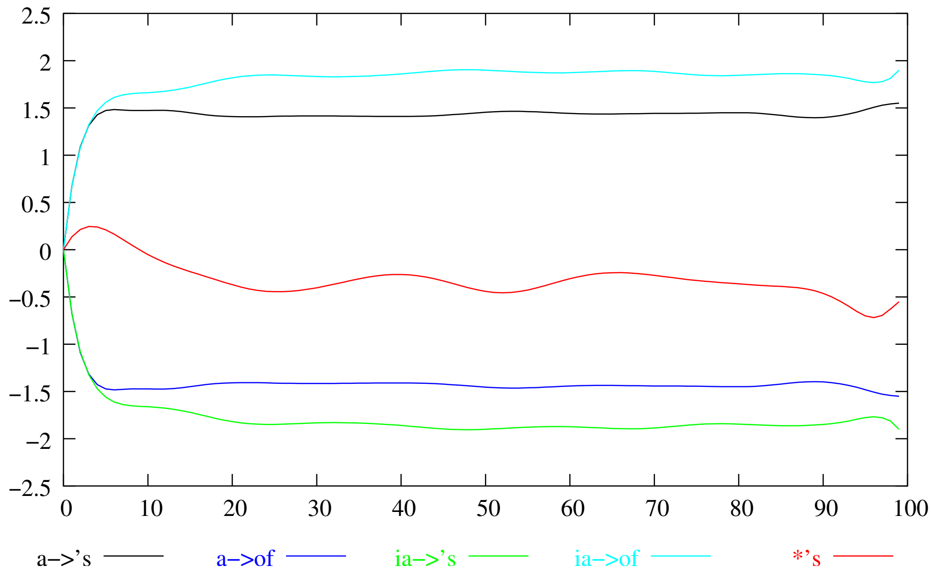
- four constraints:

  1. animate possessor $\Rightarrow$ s-genitive (a-$>$s)
  2. animate possessor $\Rightarrow$ of-genitive (a-$>$of)
  3. inanimate possessor $\Rightarrow$ s-genitive (ia-$>$s)
  4. inanimate possessor $\Rightarrow$ of-genitive (ia-$>$of)
  5. avoid s-genitives (*s): counts number of words in prenominal genitive

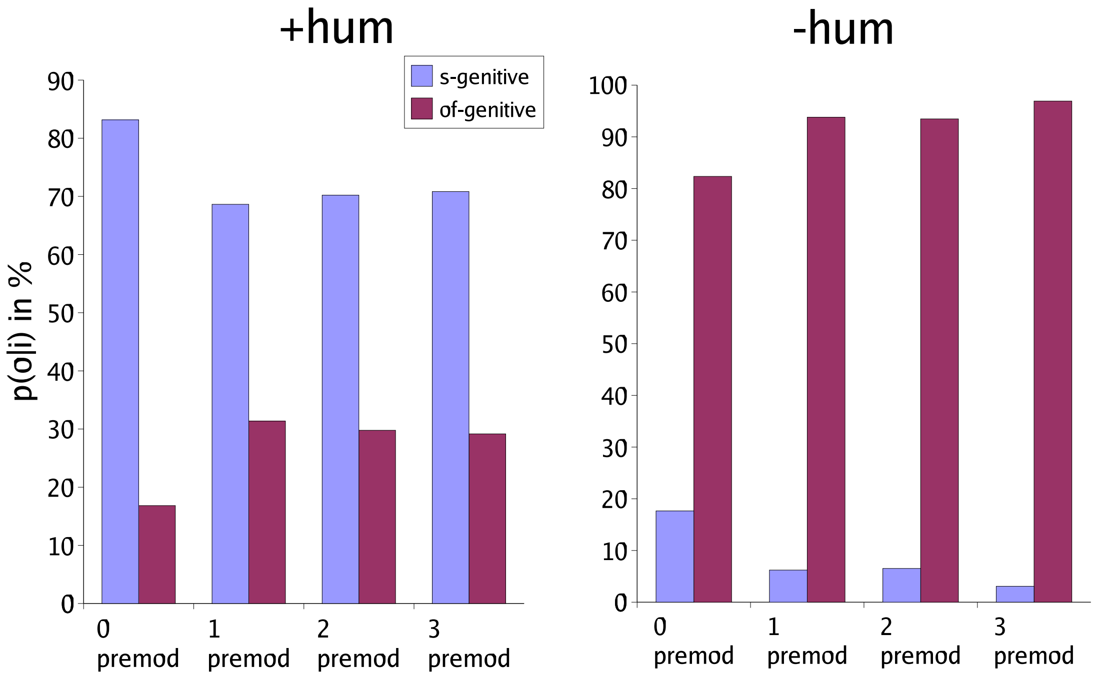
Stochastic OT

- evaluator component: Stochastic OT in the sense of Boersma 1998
- Learning algorithm: Gradual Learning Algorithm
- acquired grammar:

a->s	1.55
a->of	-1.55
ia->s	1.9
ia->of	-1.9
*s	-0.55



Predicted relative probabilities

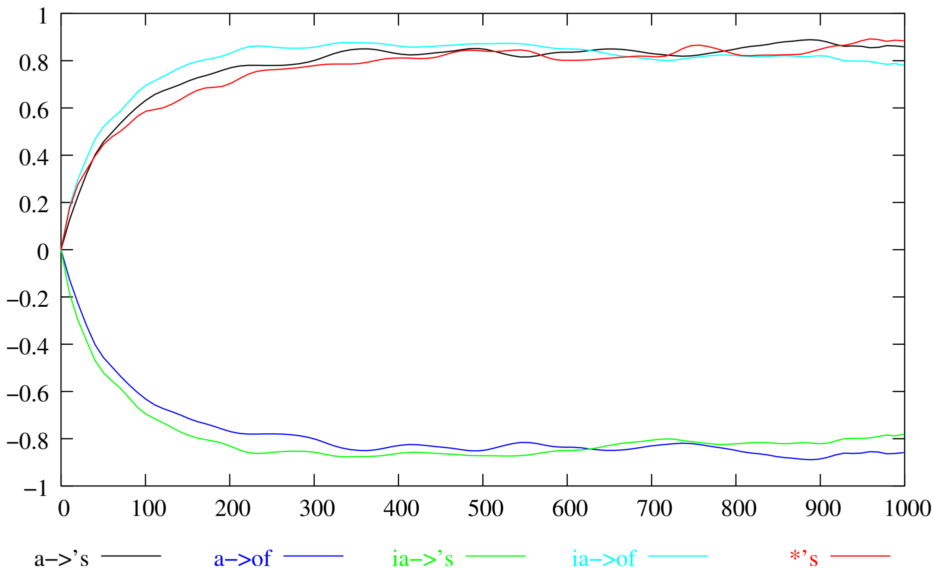


KL-divergence between predicted and observed probabilities: 0.0214

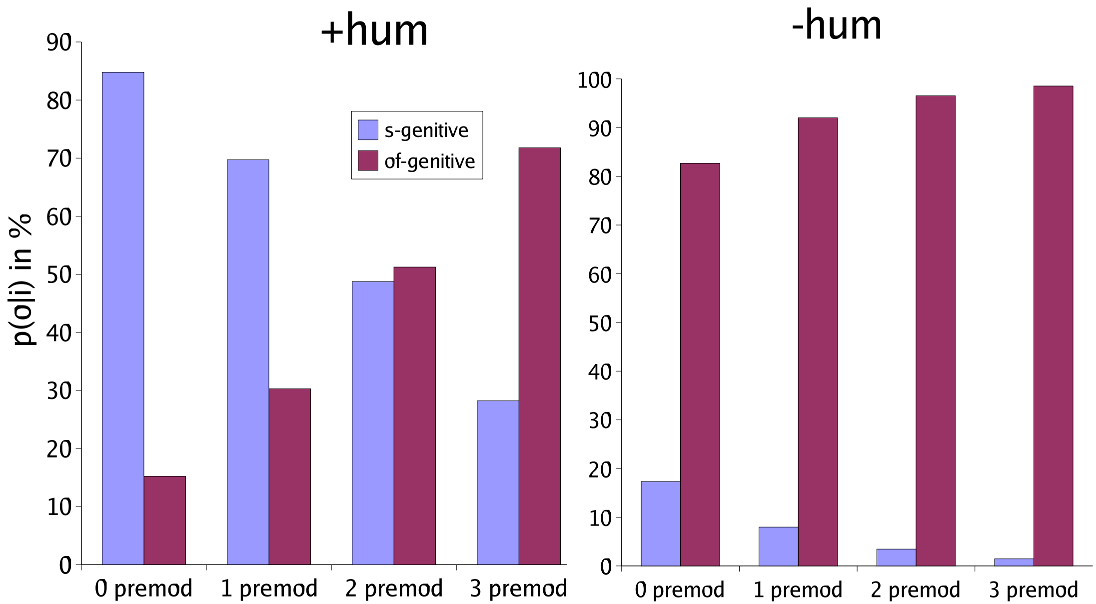
MaxEnt

- evaluator component: log-linear probabilities
- Learning algorithm: Gradual Learning Algorithm
- acquired grammar:

a->s	0.859
a->of	-0.859
ia->s	-0.781
ia->of	0.781
*s	0.884



Predicted relative probabilities



KL-divergence between predicted and observed probabilities: 0,0103

- Maxent model
  - accounts for both kinds of cumulativity
  - provides better fit of the data
- additional advantages of Maxent philosophy
  - sound philosophical motivation
  - several provably convergent learning algorithms are applicable (GLA, improved iterative scaling, conjugate gradient ascent, ...)
  - learning algorithms are robust — they always converge to the maximum entropy probability distribution

References

Abney, S. (1997). Stochastic attribute-value grammars. *Computational Linguistics*, **23**, 597–618.

Berger, A., Della Pietra, S., and Della Pietra, V. (1996). A maximum entropy approach to natural language processing. *Computational Linguistics*, **22**(1), 39–71.

Boersma, P. (1998). *Functional Phonology*. Ph.D. thesis, University of Amsterdam.

Della Pietra, S., Della Pietra, V., and Lafferty, J. (1995). Inducing features of random fields. CMU Technical Report CMU-CS-1995-144.

Goldwater, S. and Johnson, M. (2003). Learning OT constraint rankings using a maximum entropy model. In J. Spenader, A. Eriksson, and Ö. Dahl, editors, *Proceedings of the Stockholm Workshop on Variation within Optimality Theory*, pages 111–120. Stockholm University.

Jäger, G. (2003). Maximum entropy models and Stochastic Optimality Theory. manuscript, University of Potsdam. available from the Rutgers Optimality Archive.

Rosenbach, A. (2002). *Genitive Variation in English. Conceptual factors in synchronic and diachronic studies*. Mouton de Gruyter, Berlin/New York.

Rosenbach, A. (2003). Comparing animacy vs. weight as determinants of grammatical variation in English. submitted manuscript.