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:

<table>
<thead>
<tr>
<th></th>
<th>c1</th>
<th>c2</th>
<th>c3</th>
</tr>
</thead>
<tbody>
<tr>
<td>a1</td>
<td>*</td>
<td></td>
<td></td>
</tr>
<tr>
<td>b1</td>
<td></td>
<td>*</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>c1</th>
<th>c2</th>
<th>c3</th>
</tr>
</thead>
<tbody>
<tr>
<td>a2</td>
<td>*</td>
<td></td>
<td></td>
</tr>
<tr>
<td>b2</td>
<td></td>
<td>*</td>
<td>*</td>
</tr>
</tbody>
</table>

\[ P(b_1) > P(b_2) ? \]
Counting cumulativity

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

```
<table>
<thead>
<tr>
<th></th>
<th>c1</th>
<th></th>
<th>c1</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>a1</td>
<td>*</td>
<td></td>
<td>a2</td>
<td>***</td>
</tr>
<tr>
<td>b1</td>
<td></td>
<td></td>
<td>b2</td>
<td></td>
</tr>
</tbody>
</table>
```

\[ P(b2) > P(b1) ? \]
• “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 $\Rightarrow$ s-genitive (a-$\Rightarrow$s)
2. animate possessor $\Rightarrow$ of-genitive (a-$\Rightarrow$of)
3. topical possessor $\Rightarrow$ s-genitive (t-$\Rightarrow$s)
4. topical possessor $\Rightarrow$ of-genitive (t-$\Rightarrow$of)
5. prototypical possessor $\Rightarrow$ s-genitive (p-$\Rightarrow$s)
6. prototypical possessor $\Rightarrow$ of-genitive (p-$\Rightarrow$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:

\begin{align*}
  a\rightarrow s & : 11.69 \\
  a\rightarrow of & : -11.69 \\
  t\rightarrow s & : 1.69 \\
  t\rightarrow of & : -9.69 \\
  p\rightarrow s & : 8.18 \\
  p\rightarrow of & : -8.18 \\
  *s & : 10.74
\end{align*}
Predicted relative probabilities

StOT, GLA

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

MaxEnt, GLA

KL-divergence between predicted and observed probabilities: 0.0002
4.2. Counting cumulativity

• four constraints:

1. animate possessor $\Rightarrow$ s-genitive (a$\rightarrow$s)
2. animate possessor $\Rightarrow$ of-genitive (a$\rightarrow$of)
3. inanimate possessor $\Rightarrow$ s-genitive (ia$\rightarrow$s)
4. inanimate possessor $\Rightarrow$ of-genitive (ia$\rightarrow$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:

\[
\begin{align*}
\text{a→s} & : 1.55 \\
\text{a→of} & : -1.55 \\
\text{ia→s} & : 1.9 \\
\text{ia→of} & : -1.9 \\
\text{*s} & : -0.55
\end{align*}
\]
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:

\[
\begin{align*}
\text{a-} & \text{>s} \quad 0.859 \\
\text{a-} & \text{>of} \quad -0.859 \\
\text{ia-} & \text{>s} \quad -0.781 \\
\text{ia-} & \text{>of} \quad 0.781 \\
* & \text{s} \quad 0.884
\end{align*}
\]
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