Frequency Effects in Differential Object Marking

based on joint work in progress with

Judith Aissen, Chris Manning, and Kenjiro Matsuda

Joan Bresnan

Stanford University
[227B, Spring 2002–3]
1. Differential object marking (DOM) in Spanish

Classic markedness theory: Disharmonic argument types are either avoided or flagged by formal marking. Aissen’s (2002) formalization:a

*Markedness subhierarchy:*

\[ *\text{Obj}_{Hum} \gg *\text{Obj}_{Anim} \gg *\text{Obj}_{Inan} \]

*Avoidance interpretation:*

\[ \text{AvoidObj}_{Hum} \gg \text{AvoidObj}_{Anim} \gg \text{AvoidObj}_{Inan} \]

*Formal marking interpretation:*

\[ \text{MarkObj}_{Hum} \gg \text{MarkObj}_{Anim} \gg \text{MarkObj}_{Inan} \]

DOM is the formal marking or flagging of what is disharmonic.b

---


bIn Aissen (2000) flagging is implemented by local conjunction with \( \star \text{Ø}_{case} \), ‘Express case’.
Opposing these hierarchically related markedness constraints is an economy constraint to avoid the structural complexity of case marking: NoCase.\(^a\)

**Reranking of opposed constraints is limited by the subhierarchy:**

- a) NoCase \(\gg\) MarkO\(_{Hum}\) \(\gg\) MarkO\(_{An}\) \(\gg\) MarkO\(_{Inan}\)
- b) MarkO\(_{Hum}\) \(\gg\) NoCase \(\gg\) MarkO\(_{An}\) \(\gg\) MarkO\(_{Inan}\)
- c) MarkO\(_{Hum}\) \(\gg\) MarkO\(_{An}\) \(\gg\) NoCase \(\gg\) MarkO\(_{Inan}\)
- d) MarkO\(_{Hum}\) \(\gg\) MarkO\(_{An}\) \(\gg\) MarkO\(_{Inan}\) \(\gg\) NoCase

These constraints are hypothesized to be present in all grammars, but will be more or less active depending on their interactions. In (a) the avoidance of case is paramount, with the result that no objects are marked; in (d) the reverse holds, and all object types are marked. By the ranking in (b) the avoidance of case is more important than marking all object types except human, so only human objects are marked, and by (c) only animate objects are marked.

\(^a\)[=Aissen’s ‘*Struc\(_c\)*’]
In Spanish, animacy and definiteness interact in determining object marking; this is modelled by taking the cross-product of the constraint hierarchies.

**Two-dimensional DOM:**

Animacy x Definiteness hierarchy = pairs $< a, d >$ such that $< a, d > \geq < a', d' >$ iff $a \geq a'$ and $d \geq d'$. 
Two-dimensional DOM:
Two-dimensional DOM:
Two-dimensional DOM:
Two-dimensional DOM:
Two-dimensional DOM:
Two-dimensional DOM:
Two-dimensional DOM:
Two-dimensional DOM:
The Expansion of DOM to Human Definite Objects in Spanish
(based on Laca to appear\textsuperscript{a})

<table>
<thead>
<tr>
<th></th>
<th>Stage I Prior to 12\textsuperscript{th} c.</th>
<th>Stage II 12\textsuperscript{th} c. to 1830</th>
<th>Stage III Contemporary</th>
</tr>
</thead>
<tbody>
<tr>
<td>Hum-Pron</td>
<td>Obligatory</td>
<td>Obligatory</td>
<td>Obligatory</td>
</tr>
<tr>
<td>Hum-PN</td>
<td>Obligatory</td>
<td>Obligatory</td>
<td>Obligatory</td>
</tr>
<tr>
<td>Hum-Def</td>
<td>*</td>
<td>Optional</td>
<td>Obligatory</td>
</tr>
<tr>
<td>Hum-Indef</td>
<td>*</td>
<td>*</td>
<td>Optional</td>
</tr>
</tbody>
</table>

Stage I: \textbf{NoCase} $\Rightarrow$ MarkO\textsubscript{DefHum} $\Rightarrow$ \ldots
Stage II: \ldots $\Rightarrow$ \textbf{NoCase}, MarkO\textsubscript{DefHum} $\Rightarrow$ \ldots
Stage III: \ldots $\Rightarrow$ MarkO\textsubscript{DefHum} $\Rightarrow$ \textbf{NoCase}

\textsuperscript{a}Brenda Laca, to appear, ‘El objeto directo’, in \textit{Sintaxis histórica del español. Tomo I: La frase verbal}. Mexico City: UNAM.
Within Stochastic OT, the growth of DOM in human definite objects—which comes from the gradual movement of the opposed NoCase and MarkO_{Hum\text{Def}} constraints on the continuous ranking scale—is not quantal but continuous.

Hence the object marking properties which are categorically absent in Stage I and categorically present in Stage III are just points on a continuum of frequentistic variation. The diachronic variation itself shows a pattern that mirrors the structure of the markedness hierarchy: more prominent objects are more frequently marked. This is the stochastic generalization in the diachronic domain.
Growth of Human Definite Object Marking in Spanish As a Function of Ranking Distance of NoCase and MarkO_{HumDef} (data from Laca to appear)
2. Stylistic Case ellipsis in Japanese and Korean\textsuperscript{a}

In style shifting speakers systematically alter their grammars to express social aspects of the utterance situation, such as the degree of casualness or social closeness of the interlocutors. In Japanese and Korean ellipsis of case markers is involved in style shifting.

Rate of Zero Marking of Objects in Tokyo Japanese

(based on Matsuda 1995\textsuperscript{b})

<table>
<thead>
<tr>
<th>Newspaper Asahi Shimbun</th>
<th>Careful Speech Styles</th>
<th>Casual Speech Styles</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.59% \newline N=1178</td>
<td>41% \newline N=3083</td>
<td>65% \newline N=4446</td>
</tr>
</tbody>
</table>

\textsuperscript{a}A stochastic OT analysis of case ellipsis is first proposed by Lee (2002a) in a corpus study of Korean. See Lee (2002b) for further analysis and implications.

Style Shifting as Continuous Constraint Ranking

newspaper:

MarkO  NoCase
105.7  94.3

careful:

MarkO  NoCase
100.3  99.7

casual:

NoCase  MarkO
100.5  99.5
Boersma and Hayes (2001: 83–84): hypothesis that style is a linear factor in constraint ranking:

\[ selectionPoint_i = rankingValue_i + styleSensitivity_i \cdot Style + noise \]

(maximally casual) \[0 \leq Style \leq 1\] (maximally formal)

\(styleSensitivity_i\) is a constraint-specific value which allows groups of constraints to be coordinated in style shifting.

In the Japanese case ellipsis example, MarkO is positively style sensitive and NoCase is negatively style sensitive.

---

The style shifting phenomenon leads to two important conclusions:

—(i) not all syntactic variation can be attributed to incomplete information about relevant grammatical factors;
—(ii) there is a common grammatical structure that underlies the varying frequencies in specific corpora.
Question: Does it make sense to derive frequencies of usage from grammar? (Keller and Asudeh 2002: 240)

After all, unlike the grammaticality of a linguistic form, which is an idealization over usage, the actual frequency of usage of a form is a function of both grammatical structure (as determined by the theory embodied in the constraint set) and extra-grammatical factors such as memory limitations, processing load, and the context. These extra-grammatical factors are not represented by constraints in the stochastic grammars. Therefore the grammars that derive the given output distributions must be bogus, because their constraint rankings completely determine the distributions, when in fact non-constraint factors play an important role in determining frequency.

Answer: Knowledge of the grammatical structure of a particular language is represented by the (mean) ranking values of the constraints. Extra-grammatical factors affecting language use are represented by the variables that perturb the rankings. So each ‘competence’ grammar (= set of ranking values) is embedded in a ‘usage’ grammar (the style and noise variables). This embedding enables a much richer array of evidence to be used in studies of grammar than with classical approaches.
In fact, “MarkO” encapsulates a hierarchy of markedness constraints created by harmonic alignment; e.g. “MarkO \(\gg\) NoCase” abbreviates:

\[
\text{MarkO}_{Hum} \gg \text{MarkO}_{Anim} \gg \text{MarkO}_{Inan} \gg \text{NoCase}
\]

The latent hierarchy is revealed in the frequency patterns of case ellipsis in casual spoken Japanese: cases are most frequently ellipsed on objects \textit{low} in prominence.

Rate of Zero Marking of Objects in Conversations in Japanese Novels and Comics

(from Minashima 2001: 185)

<table>
<thead>
<tr>
<th></th>
<th>Zero</th>
<th>Full</th>
<th>Total</th>
<th>Percentage</th>
</tr>
</thead>
<tbody>
<tr>
<td>Personal pronoun</td>
<td>1</td>
<td>115</td>
<td>116</td>
<td>0.86</td>
</tr>
<tr>
<td>Proper noun</td>
<td>5</td>
<td>100</td>
<td>105</td>
<td>4.76</td>
</tr>
<tr>
<td>Human noun</td>
<td>21</td>
<td>136</td>
<td>157</td>
<td>13.38</td>
</tr>
<tr>
<td>Animal noun</td>
<td>15</td>
<td>78</td>
<td>93</td>
<td>16.13</td>
</tr>
<tr>
<td>Inanimate noun</td>
<td>401</td>
<td>1181</td>
<td>1582</td>
<td>25.35</td>
</tr>
</tbody>
</table>

Recall:
Harmonic alignment yields a *markedness reversal* for subjects. Compare:

\[
\text{MarkO}_{Hum} \Rightarrow \text{MarkO}_{Anim} \Rightarrow \text{MarkO}_{Inan} \Rightarrow \text{NoCase}
\]

\[
\text{MarkS}_{Inan} \Rightarrow \text{MarkS}_{Anim} \Rightarrow \text{MarkS}_{Hum} \Rightarrow \text{NoCase}
\]

Hence, cases are most frequently ellipsed on subjects high in prominence.

Rate of Zero Marking of Subjects in Japanese Telephone Conversations (CallHome Corpus)

(based on Fry 2001 128)\(^a\)

<table>
<thead>
<tr>
<th></th>
<th>Zero</th>
<th>Full</th>
<th>Total</th>
<th>Percentage</th>
</tr>
</thead>
<tbody>
<tr>
<td>Animate noun</td>
<td>873</td>
<td>1642</td>
<td>2515</td>
<td>35</td>
</tr>
<tr>
<td>Inanimate noun</td>
<td>829</td>
<td>1926</td>
<td>2755</td>
<td>30</td>
</tr>
</tbody>
</table>


---

The stochastic generalization:

The case marking which is categorical in one style of Japanese shows up as statistical preferences in others.

Further, the same implicational relations which structure the typological space across languages also structure frequency within individual language styles.