The Statistical Distribution of English Coordinate Noun Phrases: Parallelism and Weight Effects

N WAV 31

Roger Levy
Stanford University

October 11, 2002

1 Introduction

1.1 Coordination: Like Conjuncts

Coordination under Context-Free Grammars (CFGs):

(1) Principle of Conjoin Likes (Chomsky, 1965)
X → X Conj X

But Conjoin Likes has been demonstrated to be false (Peterson, 1986; Sag et al., 1985):

(2) Pat is a Republican and proud of it (coordination of NP and AdjP)

Claim: Although Conjoin Likes is false as a categorical claim, it is true as a statistical claim. Taking it as a statistical claim increases, rather than decreases, its explanatory power.

1.2 Constituent Ordering Preferences

Sensitivity to constituent “weight” or “heaviness” of preferred syntactic position & extrapolation:

(3) Extraposited object PP (Wasow, 1997)
   a. The prosecution showed pictures of gruesome details of the victim’s wounds to the jury.
   b. The prosecution showed pictures to the jury of gruesome details of the victim’s wounds.
   c. The prosecution showed pictures of it to the jury.
   d. *The prosecution showed pictures to the jury of it.
(4) Heavy NP shift (Hawkins, 1994)
   a. I gave the valuable book that was extremely difficult to find to Mary.
   b. I gave to Mary the valuable book that was extremely difficult to find.
   c. I gave the book to Mary.
   d. *I gave to Mary the book.

(5) Particle Movement (many researchers)
   a. She picked the books up.
   b. She picked up the books.
   c. She picked up all the folders she had forgotten the night before.
   d. *She picked all the folders she had forgotten the night before up.

Proposed explanations for apparent “weight effects”:

• Sensitivity to information status: given information precedes new information (Givón, 1983; Siewierska, 1993; Arnold et al., 2000), and given information is generally expressed more succinctly. Predicts that ordering preferences will be language- and position-independent.

• Ease of comprehension:
  – minimize the amount of structure necessary to identify the mother constituent (Hawkins, 1994). Directionality of preference is relativized to positions of functional & lexical heads of the specific language.
  – General avoidance of large center embeddings; for long constituents, preference is final > initial > medial. (Kuno, 1973; Dryer, 1992; Siewierska, 1993)

• Ease of production: saving longer constituents for later postpones commitment and facilitates production (Wasow, 1997). Predicts language-independent ordering preferences, relativized to other production-time demands.

• Ambiguity Management: constituents are ordered so as to minimize ambiguity. But ambiguity intuitions may not match corpus frequencies (Gibson and Schütze, 1999) [Example 6 and Figure 1 below].

(6) the NIH and the Centers for Disease Control (WSJ)

Claim: Weight effects occur in NP coordination, sensitive to discourse status and constituent size. Effects also vary with syntactic position.

2 Parallelism

Data source: LDC Penn Treebank, Wall Street Journal, Brown, and Switchboard sections (Marcus et al., 1994).1

1The Wall Street Journal section of this corpus consists of roughly 1 million words of 1989 Wall Street Journal text; the Brown section is about half a million words of a balanced corpus of American English, and the Switchboard conversations consist of recorded telephone conversations between American adults, and is roughly the same size as the WSJ corpus.
Figure 1: Mismatch between corpus attachment frequency (low attachment) and comprehension preference (high attachment); c.f. Example 6. (Gibson and Schütze, 1999)

- Gross statistical generalization: unlike coordinations are rare (Table 1).

<table>
<thead>
<tr>
<th>Unlike Coord. containing NP</th>
<th>WSJ</th>
<th>Brown</th>
<th>Switchboard</th>
</tr>
</thead>
<tbody>
<tr>
<td>NP coordination</td>
<td>9201</td>
<td>2430</td>
<td>3083</td>
</tr>
<tr>
<td>% Unlike Coord.(^2)</td>
<td>0.6%</td>
<td>1.4%</td>
<td>3.1%</td>
</tr>
</tbody>
</table>

Table 1: Empirical frequencies of unlike coordinations containing NP

Attestations of unlike coordination: *gently, and with minimum pain at each stage* (Brown); *52 years old and a 27-year Reuters veteran* (WSJ); *not crufty, but not a dress either* (Switchboard).

Extending the statistical Conjoin Likes generalization beyond gross syntactic category: the internal structures of conjunct daughters should also be similar.

(7)  a. 

```
NP  
NP₁ CC NP₂
```

b. 

```
NP  
NP₁ CC NP₂ 
NP PP  
NP PP
```

The distribution of expansions of NP₁ and NP₂ from 7a are correlated. Therefore, local tree 7b is seen more often in WSJ than otherwise expected (Table 2).\(^3\)

- Right daughters show two to three times the frequency of PP attachment as do left daughters (likely due to weight effects discussed in Section 3)

\(^2\)Though it’s tempting to draw conclusions about the relationship of corpus type to unlike coordination frequency, the results given should be considered preliminary. The Treebank is highly inconsistent in its annotation of unlike coordinations, and many be best analyzed as like coordinations.

\(^3\)All p values are given based on two-tailed tests.

### Table 2: Contingency table of left and right NP conjunct daughter expansions from 7a (subscripts are expected values under independence of sister expansions). \(p \ll .001\)

<table>
<thead>
<tr>
<th></th>
<th>Left dtr</th>
<th>Right dtr</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>NP → NP PP</td>
<td>408 (_{21})</td>
<td>249 (_{30})</td>
<td></td>
</tr>
<tr>
<td>other</td>
<td>515 (_{21})</td>
<td>3853 (_{56})</td>
<td></td>
</tr>
</tbody>
</table>

### Table 3: Mutual Information for NP → NP PP expansions of conjunct sisters. **: \(p \ll .01\); *: \(p < .05\)

<table>
<thead>
<tr>
<th></th>
<th>WSJ</th>
<th>Brown</th>
<th>Switchboard</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mutual Information</td>
<td><strong>0.108</strong></td>
<td><strong>0.075</strong></td>
<td>0.00249</td>
</tr>
</tbody>
</table>

### 3 Conjunct Weight and Positioning

Data Source: Treebank WSJ section

Method: compare lengths in words of NP conjunct sisters

\(^4\)Mutual information can be interpreted as the average measure of informativity between two random variables. It is zero for independent variables and increases as informativity increases—that is, as knowing the outcome of one variable helps guess the outcome of the other.
Overall, there is a clear tendency for longer conjuncts to follow shorter conjuncts, as can be seen in Figure 2.\footnote{In this section, an NP conjunct pair is a valid datapoint if it is split by a conjoining category (CC).}

However, the effect is not as strong as found for Heavy Noun Phrase Shift and Dative Alternation by Wasow (1997), who reported a “weight monotonicity” rate of >86% for both alternations. For NP conjuncts, weight monotonicity is 68.1%.

Figure 3 shows weight monotonicity by different conjunct length. Increasing weight has a gradually stronger effect on conjunct positioning, up to virtual disappearance of L > R ordering for difference 18 and higher.

![Figure 3: Proportion of L < R orderings by difference in NP conjunct length](image)

3.1 Conjoint position and theories of parsing complexity

- Theories of weight-dependent constituent ordering have generally focused on the VP, plus subject placement in free word order languages (Siewierska, 1993; Hawkins, 1994; Wasow, 1997)

- In English, this has meant that all theories predict the same constituent order for the data

![Figure 4: Distribution of sister NP conjunct size difference for sentence-initial positions](image)
Sentence-initial coordinate NP are preverbal and leftmost; they provide a testbed in English for competing theories.

- In the Hawkins (1994) theory of Constituent Recognition Domains, the positioning of non-head conjuncts is important only for the identification of the immediate mother category—in this case, the coordinate mother. Nouns and Determiners construct NPs, and the bulk of heavy NPs is post-nominal, so small before large (L < R) is optimal.

- In pragmatic theories, older (and thus shorter) material precedes newer; this should hold irrelevant of coordinate mother position.

- In theories of pure center-embedding avoidance, the sentence-initial position is superior, so large should precede small for NPs that begin sentences (L > R).

- Figure 4 shows the difference in sister NP conjunct size for sentence-initial positions. Although there is still a small preference for increasing conjunct weight (p = 0.007), the difference has shrunk considerably from Figure 2. Interestingly, for conjunct length difference ≤ 3, weight and ordering are no longer significantly correlated (p = 0.35).

- Although some L > R examples may be facilitated by ambiguity management (see 3.2 below), neither production considerations nor discourse factors seem to offer an explanation for the decrease in preference for L < R. Examples:

  \[(10)\] L > R sentence-initial conjunct NPs

  a. Last week’s uncertainty in the stock market and a weaker dollar
  b. The state-owned industrial holding company Instituto Nacional de Industria and the Bank of Spain

- Plausible explanation: leftmost position is preferred structurally over the second, preverbal position for larger constituents, consistent with simple comprehension-oriented theories of center-embedding avoidance.\(^6\)

- Result does not support Hawkins’s CRD theory, or pragmatic theories of constituent order

3.2 Other effects

- Discourse constraints

  \[(11)\] Given before new in sentence-initial coordinate NPs

  a. Ray White in Utah and Walter Bodmer, a researcher in Great Britain, (WSJ)

b. The latter two and Judge Daniel M. Friedman, 73, (WSJ)

c. The city park and a street bearing the Rothschild name

- Figure 5 compares NP conjunct sister pairs initiated by the and a. Assumption: definiteness associates with given information more often than indefiniteness. Mean conjunct length difference is significantly different for the before a than vice versa, suggesting that both discourse and weight play a role.

\[(12)\] ‘Pre’ 

\[
\begin{array}{c}
\text{NP} \\
\text{CC} \\
\text{NP}
\end{array}
\]

\[(13)\] ‘Post’ 

\[
\begin{array}{c}
\text{NP} \\
\text{CC} \\
\text{NP}
\end{array}
\]

\[\begin{array}{c}
\text{the . . .} \\
\text{a . . .}
\end{array}\]

Figure 5: Difference in Length(R)-Length(L) for the-initial conjunct before/after a-initial conjunct (p < 0.01; with outliers removed, p < 0.02)

- Ambiguity management principles
5 Further Work

- Directly test the explanatory force of semantic context in parallelism:
  - Control for external governor of coordinate mother
  - Examine genitive/premodifier pair alternations inside coordinate NPs with semantically neutral contexts (chairman of the company vs. company chairman)
  - Investigate whether coordinate mother weight considerations influence the expression of individual conjuncts

References


