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

(3) Extraposed 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 avoidence 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 committment 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]

[1] The 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 conversation consists of recorded telephone conversations between American adults, and is roughly the same size as the WSJ corpus.

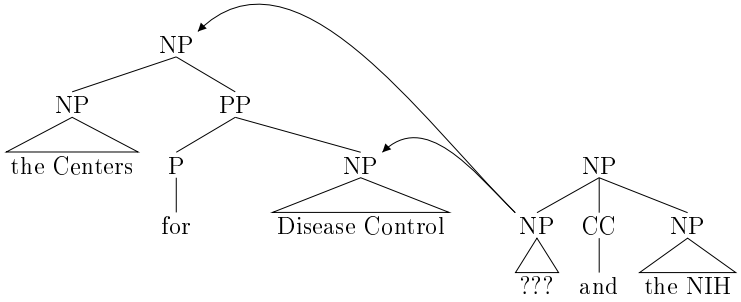


Figure 1: Mismatch between corpus attachment frequency (low attachment) and comprehension preference (high attachment); c.f. Example 6. (Gibson and Schütze, 1999)

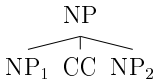
- Gross statistical generalization: unlike coordinations are rare (Table 1).

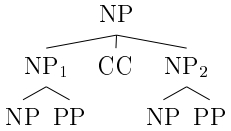
	WSJ	Brown	Switchboard
Unlike Coord. containing NP	60	35	98
NP coordination	9201	2470	3083
% Unlike Coord.[2]	0.6%	1.4%	3.1%

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 cruddy, 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. 

b. 

The distribution of expansions of NP_1 and NP_2 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.

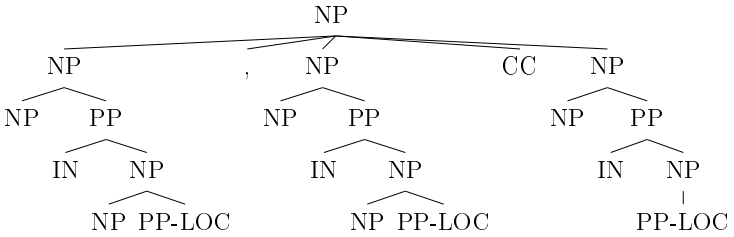
	Right dtr	
Left dtr	NP →NP PP	other
NP →NP PP	408_{121}	249_{536}
other	515_{802}	3853_{3566}

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

- There is a statistically significant correlation between right and left conjunct expansions for all corpora
- The strength of the effect for PP expansions, as measured by mutual information[4], is strongest for newswire text (WSJ) and weakest for the spoken corpus (Switchboard).

(8) a drawing of Pinocchio and a photograph of Mr. Florio's rival, Republican Rep. Jim Courtner (NP PP and NP PP)

(9) a. the phase-out of a battery facility in Greenville, N.C., the recent closing of a Hostess cake bakery in Cincinnati and a reduction in staff throughout the company

b. 

	WSJ	Brown	Switchboard
Mutual Information	**0.108	**0.075	*0.00249

Table 3: Mutual information for NP → NP PP expansions of conjunct sisters. **: $p \ll 0.01$; *: $p < 0.05$

3 Conjunct Weight and Positioning

Data Source: Treebank WSJ section

Method: compare lengths in words of NP conjuct 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.[5]
- 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 in 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.

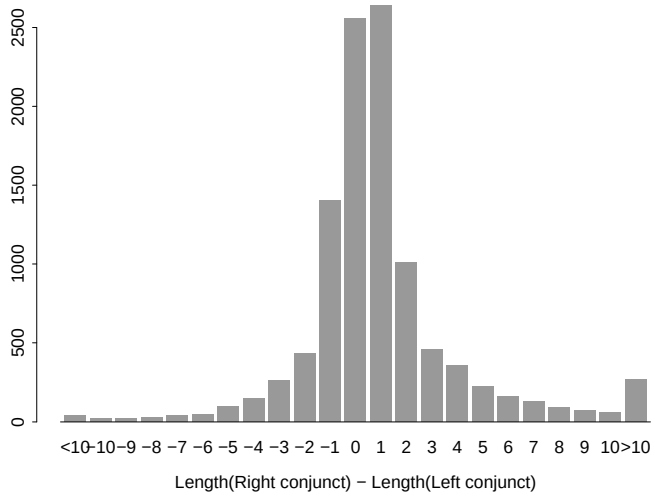


Figure 2: Preference for increasing NP conjunct length, all coordinate NPs, WSJ (Length(right sister) - Length(left sister)). Mean length difference is 0.6

3.1 Conjunct 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

[5] In this section, an NP conjunct pair is a valid datapoint if it is split by a conjoining category (CC).

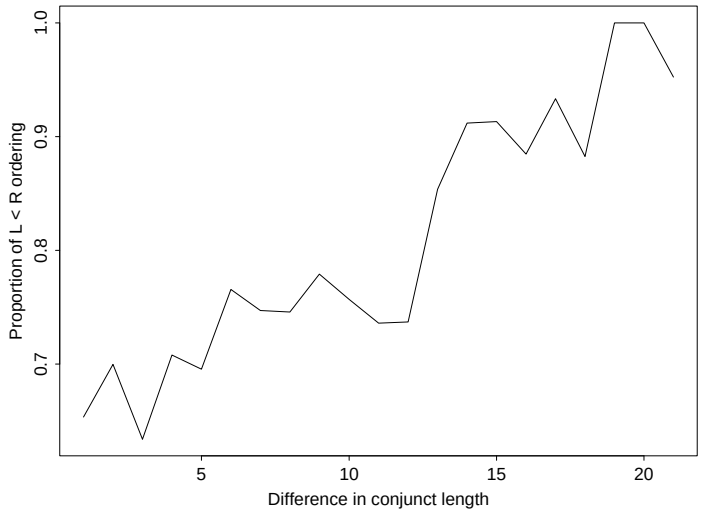


Figure 3: Proportion of L < R orderings by difference in NP conjunct length

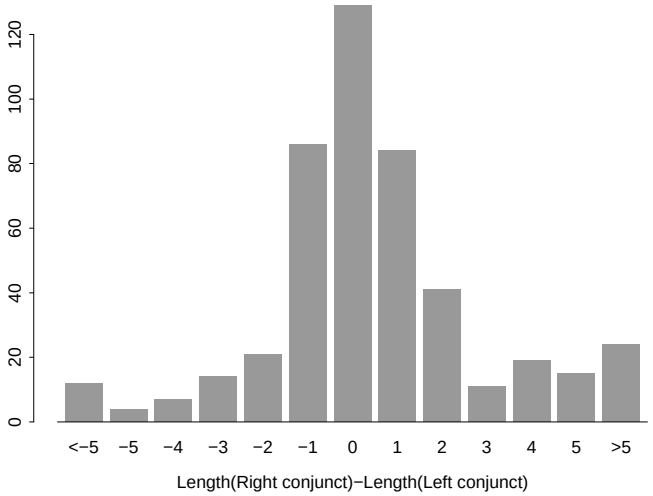


Figure 4: Distribution of sister NP conjunct size difference for sentence-initial positions

- 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

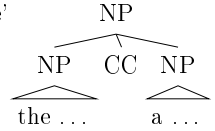
(11) Given before new in sentence-initial coordinate NPs
  a. Ray White in Utah and Walter Bodmer, a researcher in Great Britain, (WSJ)

[6] One possible alternative explanation is that subjects tend to be human or animate actors and there may be more stringent non-linguistic constraints on their ordering within a coordinate (such as priority among multiple acting parties). Assuming length of name is independent of such constraints, this would tend to reduce any ordering preference by weight.

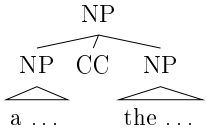
  b. The latter two and Judge Daniel M. Friedman, 73, (WSJ)
  c. The city park and a street bearing the Rothschild name

– 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'



(13) 'Post'



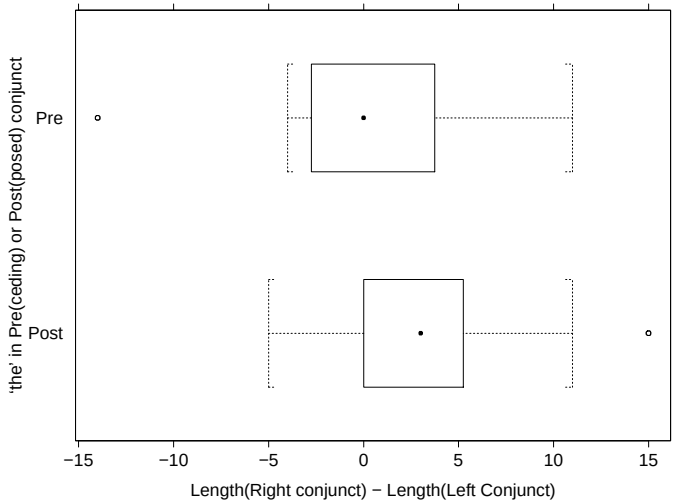


Figure 5: Difference in Length(R)-Length(L) for *the*-initial conjunct before/after *a*-initial conjunct ($p < 0.01$; with outliers removed, $p < 0.02$)

- Ambiguity management principles

– Punctuation (c.f. Schafer and Speer (1999) for parallels in intonation) and not-fully-parallel conjoiners, sometimes optional and sometimes obligatory, may eliminate potential ambiguities in L > R coordinations

(14) a. editor and co-owner of the Daily Tribune in Ames, Iowa, and President of NBC News in New York

b. Joni Evans, recruited two years ago to be publisher of adult trade books for Random House, and Sonny Mehta, president of the prestigious Alfred A. Knopf unit

c. a satisfactory due diligence investigation by Penn Central, a definitive agreement and regulatory approvals

d. *the most egregious violator of weight principles in WSJ:* the political manifestations of the Rowland-Molina theory (named after the researchers who found in 1974 that chlorofluorocarbons contributed to the depletion of ozone in the earth's atmosphere) and the Montreal Protocol

– Of 2261 WSJ L > R conjunct pairs (excluding those with possessive pronouns), 1206 are divided by a comma or a conjunction other than 'and'

– Postposing heavy constituents may avoid ambiguities of CC NP attaching on the right side of NP PP/S′ (Gibson and Schütze, 1999)

4 Conclusion

Corpus-based model of the composition and structuring of coordinate NPs:

- Conjunct NPs are generated by sampling from a joint distribution that reflects a statistical Coordinate Structure Constraint (Tables 2 and 3)
- The order of generated conjuncts is chosen based on discourse and processing principles (Figures 4 and 5)
- This may explain Gibson and Schütze (1999)'s mystery: why do attachment preferences of conjunctions into right-branching NPs not reflect corpus frequency? Perhaps because the corpus frequency of high attachments is deflated by positioning preferences. Alternatively, it may be that NP CC NP PP ambiguities are generally easier to resolve than NP P NP CC NP ambiguities via lexical preferences, so corpus frequencies *do* reflect ambiguity-minimization.
- Difference in strength of parallelism effect between written/formal and spoken/informal corpora (Table 3) suggests that parallelism is not production-oriented; it may be comprehension-oriented, but most likely is partly stylistic.

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

Arnold, J. E., Wasow, T., Losongco, A., and Ginstrom, R. (2000). Heaviness vs. newness: The effects of structural complexity and discourse status on constituent ordering. *Language*, 76(1):28–55.

Chomsky, N. (1965). *Aspects of the Theory of Syntax*. Cambridge.

Dryer, M. S. (1992). The Greenbergian word order correlations. *Language*, 68:81–138.

Gibson, E. and Schütze, C. T. (1999). Disambiguation preferences in noun phrase conjunction do not mirror corpus frequency. *Journal of Memory and Language*, 40:263–279.

Givón, T. (1983). *Topic Continuity in Discourse*. Amsterdam: John Benjamins.

Hawkins, J. A. (1994). *A Performance Theory of Order and Constituency*. Cambridge.

Kuno, S. (1973). *The Structure of the Japanese Language*. Cambridge, MA: MIT Press.

Marcus, M. P., Santorini, B., and Marcinkiewicz, M. A. (1994). Building a large annotated corpus of English: The Penn Treebank. *Computational Linguistics*, 19(2):313–330.

Peterson, P. G. (1986). Establishing verb agreement with disjunctively conjoined subjects: Strategies vs principles. *Australian Journal of Linguistics*, 6(2):231–249.

Sag, I. A., Gazdar, G., Wasow, T., and Weisler, S. (1985). Coordination and how to distinguish categories. *Natural Language and Linguistic Theory*, 3:117–171.

Schafer, A. J. and Speer, S. R. (1999). Intonational disambiguation in sentence production and comprehension. Presented at the 12th CUNY Conference on Human Sentence Processing.

Siewierska, A. (1993). Syntactic weight vs information structure and word order variation in Polish. *Journal of Linguistics*, 29:233–265.

Wasow, T. (1997). Remarks on grammatical weight. *Language Variation and Change*, 9:81–105.