The Statistical Properties of Coordinate Noun Phrases

Roger Levy
Stanford University

March 11, 2004

1 Introduction

Two major points about the nature of grammar, both illustrated in new ways using the coordination of noun phrases:

1. The nature of distributional patterns in syntactic structure
    - Traditional notion of coordination as an *internally* constrained relation: combine two things that are alike in some respect
    - Recent theoretical work on the syntax of coordination suggests that there are *no* internal categorical constraints operative between conjuncts
    - I show that a strong *non-categorical* constraint operates between NP conjuncts. Conjuncts tend toward *parallelism* – they are unusually similar to one another. This pattern holds at a variety of granularity levels.
    - With modern syntactically annotated corpora, we can investigate these patterns across genres and languages
2. The role of functional considerations in determining linear order of constituents
    - Linear ordering is sensitive to constituent *weight* in many cases
    - Proposed explanations vary from memory conservation to center-embedding avoidance to discourse-based information status
    - In English, domain of investigation has been post-verbal, and all proposed explanations make the same predictions
    - Coordinate NPs can occur *preverbally*, so we can tease apart the proposals

2 Parallelism: a non-categorical principle of *Conjoin Likes*

2.1 Categorical backdrop

- Coordination under Context-Free Grammars (CFGs) has traditionally been taken as a combination of categorically identical elements:

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

- But *Conjoin Likes* has been demonstrated to be false. Constituents of unlike syntactic category can be coordinated (Peterson, 1986; Sag et al., 1985); more recently, it's been shown that NPs of unlike case can be coordinated too (Przepiórkowski, 1999; Levy, 2001):

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

  (3) Včera vec' den' on prožd̆al [$_{\text{NP}}$ svoju podrugu Irinu] i
  yesterday all day he expected self's.ACC girlfriend.ACC Irina.ACC and
  [$_{\text{NP}}$ zvonka [$_{\text{PP}}$ ot svoego brata Grigorija]]. (Russian)
  call.GEN from self's brother Gregory
  "Yesterday he waited all day for his girlfriend Irina and for a call from his brother Gregory."

- Recent theoretical work (Ingria, 1990; Bayer and Johnson, 1995; Bayer, 1996; Dalrymple and Kaplan, 2000; Levy, 2001; Levy and Pollard, 2001; Sag, 2002) indicates that the constraints on so-called "X Conj X" coordination are all *extrinsic*: every conjunct must individually satisfy all the relevant external constraints, but there are no constraints that actually operate solely between conjuncts.

Assertion: 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. Furthermore, the statistical trend of conjunct similarity *cannot* entirely be explained away by external constraints. It is a preference for pure structural similarity between conjuncts.

2.2 Non-categorical parallelism in a statistical framework

- Data sources: LDC Penn Treebank of English (ETB), Wall Street Journal (WSJ), Brown, and Switchboard (SWBD) sections (Marcus et al., 1994); LDC Penn Treebank of Chinese (CTB) (Xue et al., 2002).[1]

[1]The WSJ section of ETB 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 Switchboard consists of

	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

- Intuitive overall level of formality: WSJ > Brown > Switchboard; Chinese Treebank perhaps similar to WSJ (both newswire)

2.2.1 Coarse-grained generalization: unlike coordinations are rare

- Unlike coordinations are attested in all three genres of English corpus:

  gently, and with minimum pain at each stage [AP & PP] (Brown)
  52 years old and a 27-year Reuters veteran [AP & NP] (WSJ)
  not cruddy, but not a dress either [AP & NP] (Switchboard)

- However, unlike coordinations are rare (Table 1). *They are rarer in less formal corpora.*

At the level of gross syntactic category, *Conjoin Likes* is stronger in more formal language.

2.2.2 *Conjoin Likes* beyond gross syntactic category: constituent substructure

- Intuitively, some coordinations just sound more parallel (and often nicer) than others:

  Japan's Haruki Murakami and China's Gao Xingjian
  Haruki Murakami of Japan and China's Gao Xingjian
  Japan's Haruki Murakami and Gao Xingjian of China
  Haruki Murakami of Japan and Gao Xingjian of China

- This type of parallelism could also be a type of *Conjoin Likes*. If it is reflected in corpus data, then the internal modificational structures of conjuncts should tend to be similar.

- *Conjoin Likes* Test cases:

recorded telephone conversations between American adults, and is roughly the same size as the WSJ corpus. I use version 3 of the Chinese Treebank, which contains roughly 250,000 segmented words of printed Chinese text.

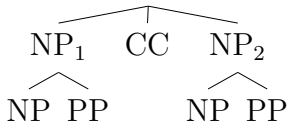
[2]Though it's tempting to draw conclusions about the relationship of corpus type to unlike coordination frequency, the results given should be considered preliminary. The Treebank is highly inconsistent in its annotation of unlike coordinations, and many may be best analyzed as like coordinations.

– English: right modification by prepositional phrases (4b) and relative clauses and sentential complements (4c)[3]

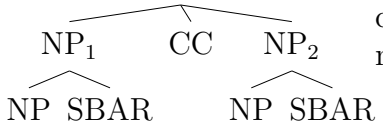
– Chinese: left modification by genitive and clausal *de*- phrases (4d)[4]

(4) a. NP (Base)

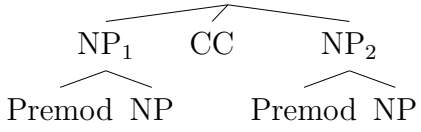
b. NP (English – parallel PP modification)



c. NP (English – parallel relative clause/sentential complement modification)



d. NP (Chinese – parallel premodification)



- The *parallelism effect*: rewrites of NP_1 and NP_2 from 4a are *correlated*. That is, the tree fragments in 4b, 4c, and 4d are seen more often than would be expected if the internal structures of NP_1 and NP_2 were statistically independent (Table 2).[5]

[3]I group relative clauses and sential complements under the syntactic category SBAR with which they are annotated in the treebank.

[4]The Chinese Treebank distinguishes between two types of prenominal *de*- modification: (i) genitive modification by NPs:

Běihǎi shì *de* juéqǐ
Beihai city rise
'The rise of the city of Beihai'

and (ii) premodification by a clause, including relative clauses and sentential complements:

Pǔdōng xīnqū guīdìng *de* fǎguīxìng wénjiàn
Pudong new-district formulate legal document
'The legal documents formulated by Pudong New District'

Due to the lack of morphology distinguishing parts of speech in Chinese, the dividing line between the two types is often unclear; nevertheless, I treat the two classes separately whenever possible, as it is a more stringent test.

[5]All p-values are given based on two-tailed tests. Contingency-table p-values are calculated with Fisher's exact test.

Chinese

	Right	
Left	hasCP	noCP
hasCP	42_{7}	36_{71}
noCP	48_{83}	844_{810}

	Right	
Left	hasDNP	noDNP
hasDNP	66_{15}	46_{97}
noDNP	48_{113}	800_{749}

English – WSJ

	Right	
Left dtr	hasPP	noPP
hasPP	498_{138}	228_{502}
noPP	544_{904}	3567_{3293}

	Right	
Left dtr	hasSBAR	noSBAR
hasSBAR	35_{2}	26_{59}
noSBAR	128_{161}	4561_{4528}

English – Brown

	Right	
Left dtr	hasPP	noPP
hasPP	95_{31}	52_{116}
noPP	174_{238}	946_{882}

	Right	
Left dtr	hasSBAR	noSBAR
hasSBAR	15_{1}	7_{21}
noSBAR	52_{66}	1166_{1152}

English – Switchboard

	Right	
Left dtr	hasPP	noPP
hasPP	78_{36}	76_{118}
noPP	325_{367}	1230_{1188}

	Right	
Left dtr	hasSBAR	noSBAR
hasSBAR	15_{2}	21_{34}
noSBAR	71_{84}	1596_{1583}

Table 2: Contingency table of left and right NP conjunct daughter expansions from 4a (subscripts are expected values under independence of sister expansions). $p \ll .001$ in all cases.

- In both English and Chinese, presence of nominal modifier of a given type is highly correlated between conjuncts
- In English, right conjuncts are far more likely ($p \ll 0.001$) than left conjuncts to have PP or SBAR postmodifiers (likely due to weight effects discussed in Section 3)
- In Chinese, difference between premodifier frequency for conjuncts is insignificant
- We can measure the *strength* of the parallelism effect by the *odds ratio*, defined as:

$$O = \frac{P(both)P(neither)}{P(left)P(right)}$$

where:

  - P(*both*) = probability that both conjuncts have modifier
  - P(*neither*) = probability that neither conjunct has modifier
  - P(*left*), P(*right*) = probability that only left (or right) conjunct has modifier

- Example
  - Estimated odds ratio for WSJ PP's

$$\begin{aligned} \text{Total examples :} \quad & 498 + 228 + 544 + 3567 = 4837 \\ O &= \frac{\frac{498}{4837} \times \frac{3567}{4837}}{\frac{228}{4837} \times \frac{544}{4837}} \\ &= 14.32 \end{aligned}$$

  - Estimated odds ratio for Switchboard PP's

$$\begin{aligned} \text{Total examples :} \quad & 78 + 76 + 325 + 1230 = 1709 \\ O &= \frac{\frac{78}{1709} \times \frac{1230}{1709}}{\frac{76}{1709} \times \frac{325}{1709}} \\ &= 3.88 \end{aligned}$$

- Odds ratio is higher for SBAR than for PP; also, it is much higher for WSJ and somewhat higher for Brown than for Switchboard (Figure 1).[6]

At the level of PP postmodification, the parallelism effect is stronger (as measured by the odds ratio) in corpora of written language than in a corpus of conversations.

[6]Many Switchboard coordinate structures whose right conjuncts have no PP are of the form "X and/or {something/anything/all/all that/such" and perhaps should be not included as "genuine" conjuncts. Even if these are included, however, the estimated odds ratio for SWD remains considerably lower than for Brown or WSJ.

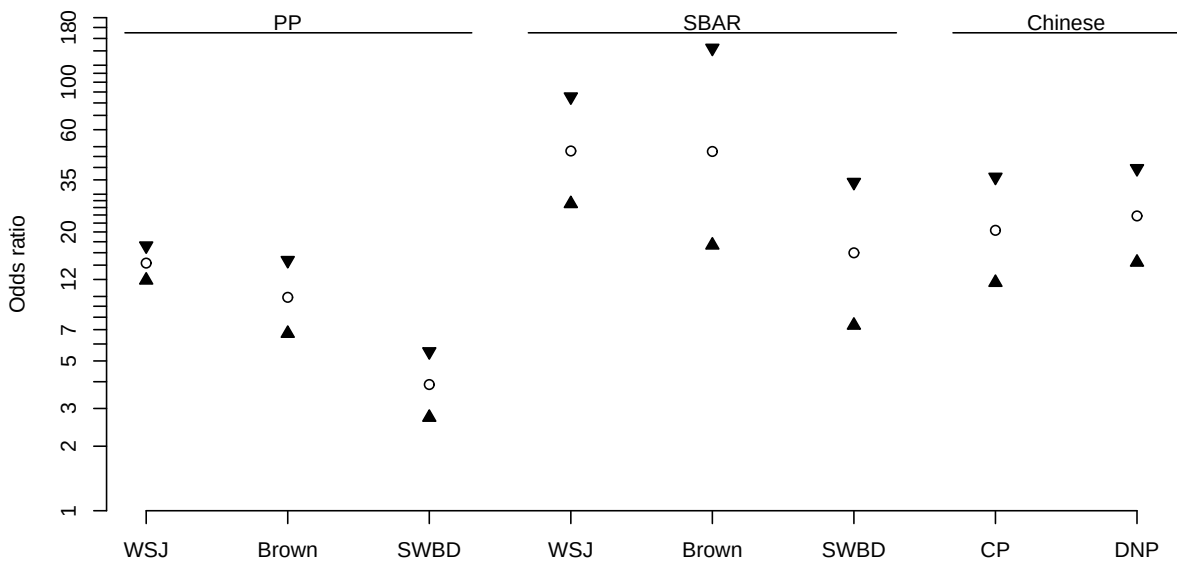


Figure 1: Estimated odds ratio for modifiers in different corpora. Triangles indicate 95% confidence intervals.

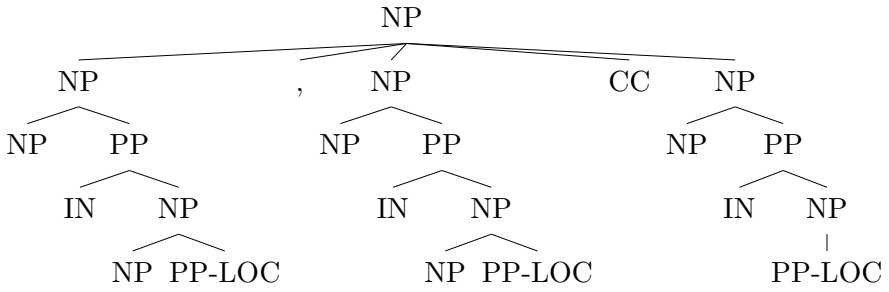
- A typical example of parallelism:

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

- An extreme example of parallelism:

  (6) 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.



[PP *of* [NP NP CC NP]]

	Right dtr	
Left dtr	hasPP	noPP
hasPP	2	14
noPP	14	98

[PP *despite* [NP NP CC NP]]

	Right dtr	
Left dtr	hasPP	noPP
hasPP	25	25
noPP	25	25

Combined	Right dtr	
Left dtr	hasPP	noPP
hasPP	27	39
noPP	39	123

Table 3: A hypothetical example of "false" parallelism, with (hypothetical) $p < 0.02$ for "Combined"

2.3 Three tests to rule out potential confounds

- The correlation of sister conjunct substructure cannot, however, immediately be ascribed to a purely structural preference for conjunct similarity; differences in conjunct expression patterns among "subpopulations" could also create such a phenomenon.
- If different external governors of coordinate NPs differ in frequencies of PP modifier in governed NP conjuncts but individually show no tendency toward parallelism, the combined sample could nevertheless show strong correlation between left and right conjunct expression. For example:
  - About half of NPs governed by *despite* have PP modifiers
  - But only about $\frac{1}{8}$ of NPs governed by *of* have PP modifiers
  - Imagine a parallelism-free corpus composed only of coordinate NPs governed by *despite* and *of*.
  - The corpus as a whole would still show correlation between presence of PPs in right and left NP conjuncts (Table 3)
- This is known as *Simpson's paradox*, and we should see if we can rule it out.[7]

2.3.1 Test 1: control for external lexical governor

- We can begin to account for possible confounds by controlling for external governor. The most common external governors for binary NP coordinations all show significant correlation between PP modification in conjunct NP complements.[8] The counts in

[7] Simpson's paradox is a widespread and well-known phenomenon in statistics. False correlations do not only arise from subgroups with independent variables; correlation patterns can actually *reverse* when attention is shifted from super- to sub-population.

[8] *External governor* is defined as the word heading the node immediately above the coordinate NP mother. Node headship is the deterministic result of a simple set of rules sensitive to daughter category and position, due to Collins (1999) and widely used in work on probabilistic parsing. Nearly all of the common governors are straightforward instances of prepositional phrases, so it is unlikely that the results here are sensitive to precise formulation of headship.

the Chinese Treebank are much smaller, but in several cases there were significant correlations consistent with structural parallelism (Table 4).

English – WSJ			Chinese		
of	479	†	*yóu*/'from, due to'	24	n.s.
between	263	*	*zài*/'at'	21	n.s.
by	220	†	*shì*/'be'	19	**
in	216	†	*duì*/'toward'	16	**
with	198	†	*wèi*/'for'	15	**
for	186	†	*yǒu*/'have'	15	**
to	182	†			
are	142	†			
as	139	†			

Table 4: Most frequent external lexical governors of coordinate NPs, token counts, and significance levels in WSJ, Switchboard, and Chinese corpora. *: $p < 0.05$; **: $p < 0.01$; †: $p < 0.001$

2.3.2 Test 2: control for external governor across corpus type

- Simpson's paradox could potentially also explain the difference in degree conjunct correlation across corpora/genre, since different corpora have different word distributions etc.
- But, comparing WSJ and Switchboard while controlling for external governor, we find dramatic differences in conjunct correlation. In fact, no set of Switchboard conjuncts for a given external governor shows significant correlation (Table 5).[9]
- This finding, along with the difference in overall strength of parallelism show in Figure 1, is strong evidence for parallelism as a structural preference, differentially active in different modalities.

In a spoken corpus, the parallelism effect is completely explained away by controlling for external lexical governor. It is not explained away in written corpora.

[9]The nine most common external governors of coordinate NPs differ for WSJ and Switchboard; Switchboard's most common external governor, *have*, does show a marginal trend toward PP modification parallelism ($p < 0.1$).

English	WSJ		SWBD		Chinese		
of	479	†	113	n.s.	*yóu*/'from, due to'	24	
between	263	*	24	n.s.	*zài*/'at'	21	
by	220	†	12	n.s.	*shì*/'be'	19	**
in	216	†	84	n.s.	*duì*/'toward'	16	**
with	198	†	90	n.s.	*wèi*/'for'	15	**
for	186	†	71	n.s.	*yǒu*/'have'	15	**
to	182	†	54	n.s.			
are	142	†	39	n.s.			
as	139	†	25	n.s.			

Table 5: Most frequent external lexical governors of coordinate NPs, token counts, and significance levels in WSJ, Switchboard, and Chinese corpora. *: $p < 0.05$; **: $p < 0.01$; †: $p < 0.001$

2.3.3 Test 3: pre/post-modification alternation

- We can also examine parallelism as a preference for structural similarity by looking at a semantically "weak" syntactic alternation, that between genitive (functionally, *possessive*) pre- and post-modification.

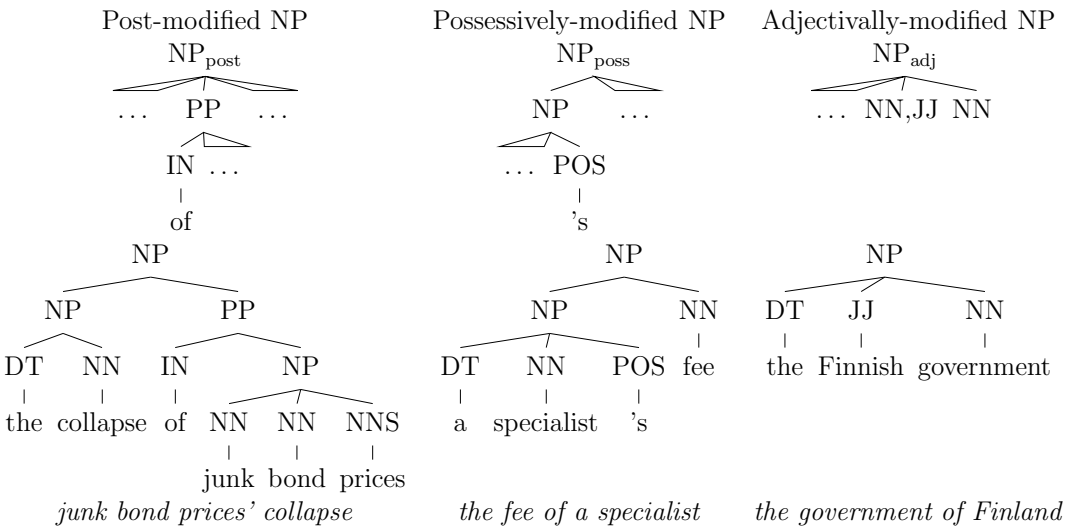
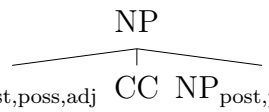


Figure 2: Tree fragments for pre/post-modification alternation

- Three different forms of genitive NP modification, together with presumed competing forms (Figure 2).[10]

[10]The rather unclear boundary in English between noun compounding and adjectival modification has led to inconsistent nominal premodifier annotation in the Penn Treebank, so premodifiers annotated as singular nouns (NN) and as adjectives (JJ) are both candidates for functional possessive modification.

- I searched the WSJ corpus for tree matches of the form



  and manually filtered out all but true examples of genitive modification.

- Examples of parallel modification:

  (7) a. Post-modification:
  the former two-time president of NBC News and creator of the Huntley-Brinkley Report

  b. Possessive modification:
  the company's stock or the specialists' performance

  c. Adjectival modification:
  solid-waste recovery and hazardous-waste cleanup

- Non-parallel examples:

  (8) a. Postnominal and possessive:
  Goldman, Sachs & Co. of the U.S. and Japan's Daiwa Securities Co
  the desert's heat and the cool of the ocean

  b. Postnominal and adjectival:
  the failure of the UAL deal and the stock-market plunge
  a temporary wage and price freeze and a devaluation of the cruzado

  c. Possessive and adjectival:
  analysts' forecasts and the year-earlier level
  a floor brokerage fee or a specialist's fee

- Modification types are highly correlated across conjuncts (Table 6). Also, possessive and adjectival modifiers seem to pattern as distinct types.[11] *The parallelism effect appears to be strong for genitive modification types.*

[11] It is possible that modifier size could be a confound for the premodification/postmodification correlation, if modifier size is correlated across conjuncts, because larger modifiers would tend to be prefer postnominal expression. The finding that possessive versus adjectival modification exhibits a similar correlation would, however, tend to undermine this objection.

	Right		
Left	Post	Poss	Adj
Post	77	10	5
Poss	12	23	2
Adj	8	1	13

	Right	
Left	Post	Pre
Post	77	15
Pre	20	39

Table 6: Contingency tables for postnominal, possessive, and adjectival genitive modification type in coordinate NP structure, treating possessive and adjective premodifiers separately and as a group. For both groupings, $p \ll 0.001$.

2.4 Summary

- *Conjoin Likes* is false as a categorical constraint, but as a statistical constraint is stronger than categorical grammarians ever thought, operating not only at the level of gross syntactic category but also at the finer-grained level of internal modification structure
- *Conjoin Likes* is operative cross-linguistically
- Controlling for external governor cannot explain away the parallelism effect in written text
- Three separate results point to difference in strength of parallelism effect in written versus spoken modality:
  - Greater frequency of unlike-category coordination in speech (Table 1)
  - Stronger parallelism effect for PP modification, as measured by odds ratio (Figure 1)
  - Disappearance of parallelism effect *only* in spoken text when external governor controlled for

3 Conjunct Weight and Positioning

3.1 Background: Theories of Constituent Ordering Preference

- Empirical "weight effects": in English, at points of alternation between two constituent orders, orders that shift larger constituents to the right tend to be more frequent.

(9) 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.

(10) 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.

(11) Extraposed object PP (Wasow, 1997, 2002)

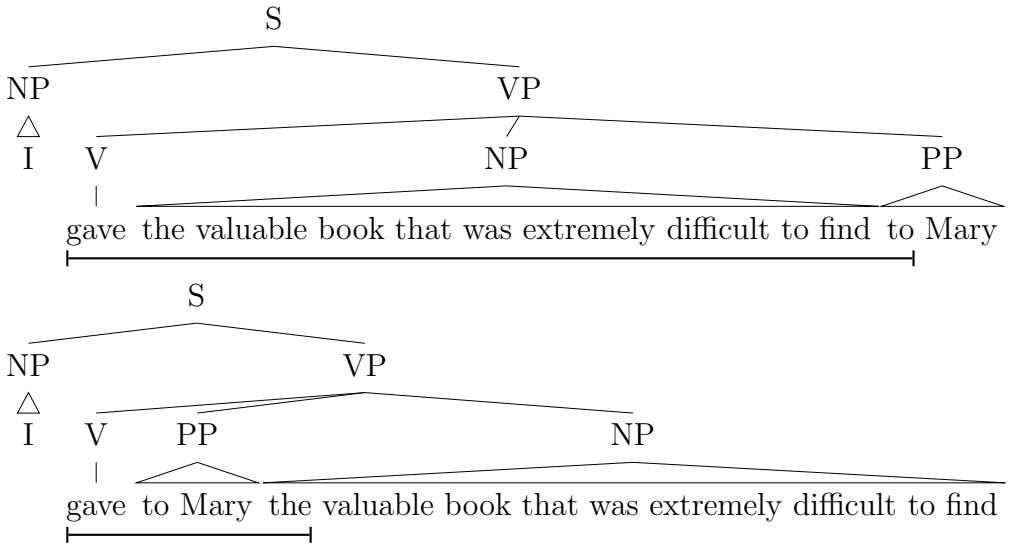
a. The prosecution showed *pictures of gruesome details of the victim's wounds* to the jury.

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.

- Proposed explanations for empirically-observed "weight effects":
  - Discourse-based 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:
    - Hawkins's memory-based theory of Constituent Recognition Domains: 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.

Claims:

1. Weight effects are empirically observed in the ordering of NP conjuncts in coordination.
2. The overall trend is consistent with English as a whole: heavier follows lighter.
3. However, there are subtrends that give evidence to theories of center-embedding avoidance and information status.

3.2 Corpus investigation

Data Source: English Treebank, WSJ section

Operative definition of *weight*: number of orthographic words dominated by constituent[12]

- Overall, there is a clear tendency for L < R: longer conjuncts follow shorter conjuncts (Figure 3).[13]
- 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" (L < R) rate of >86% for both alternations. For NP conjuncts, weight monotonicity is 68.1%.
- Figure 4 shows weight monotonicity by difference 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.

[12]Wasow (1997) found that number of words, number of nodes, and number of phrasal nodes were all highly correlated and analyses by the three measures led to essentially indistinguishable results.

[13]Histogram significance values are given according to the Mann-Whitney test.

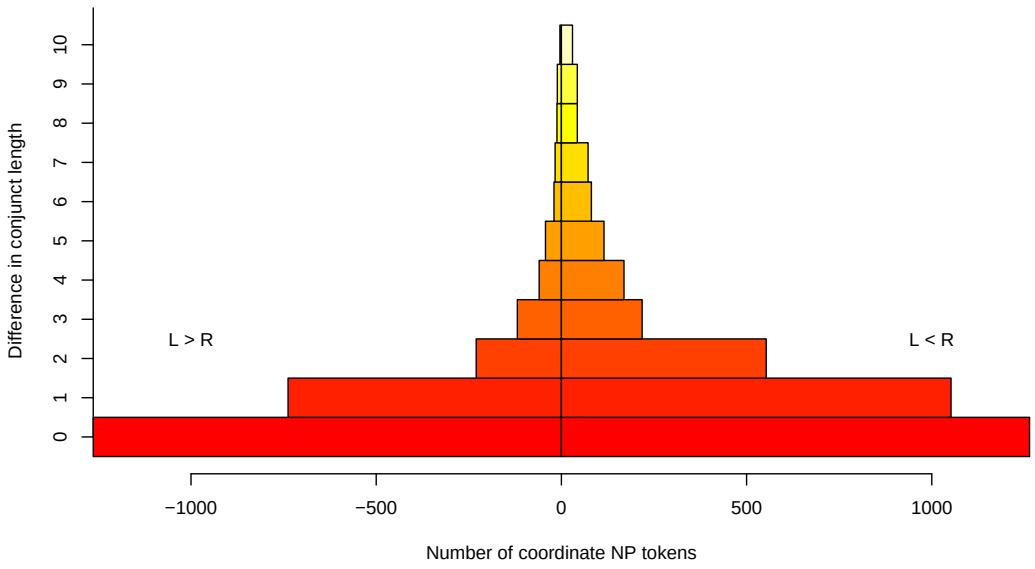


Figure 3: Preference for increasing NP conjunct length, all coordinate NPs, WSJ (Length(right sister) - Length(left sister)). Mean length difference is 0.6; $p \ll 0.001$

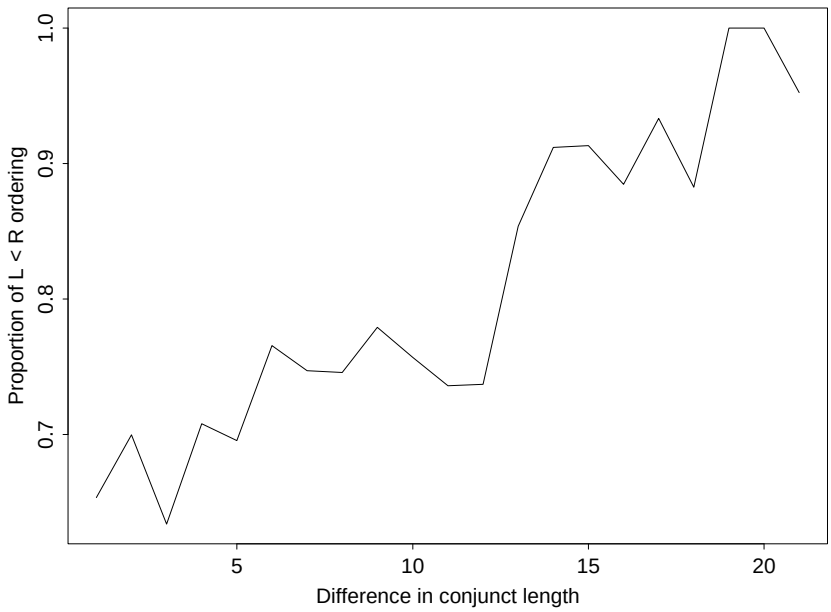


Figure 4: Proportion of L < R orderings by difference in NP conjunct length

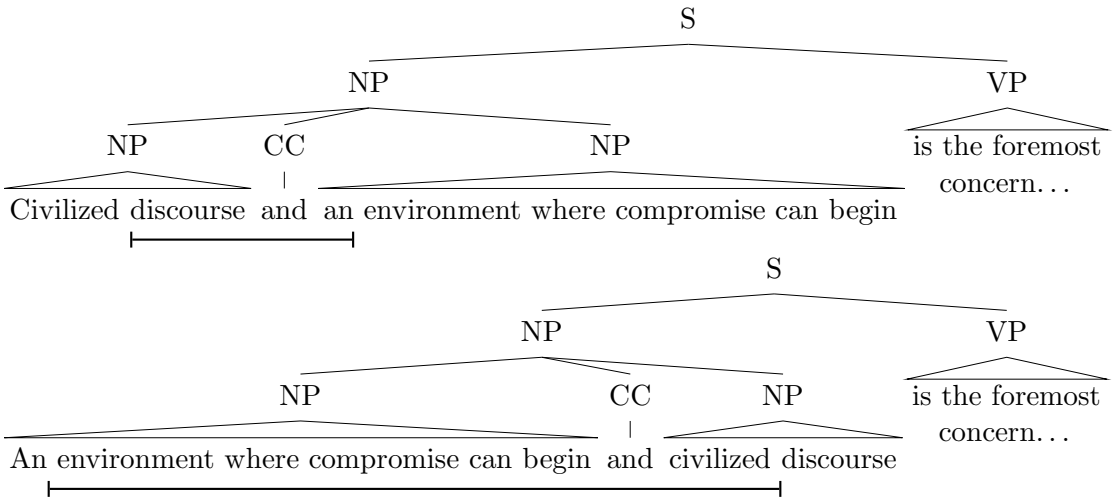
3.3 Conjunct position and theories of parsing complexity

- 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
- Sentence-initial coordinate NP are preverbal and leftmost; they provide a testbed in English for competing theories.

(12) Short-before-long sentence-initial coordinate NPs
   a. *Civilized discourse and an environment where compromise can begin* are lost in a hostile posture abetted by superficial media interviews.
   b. *Both the SUNY team and researchers at the National Magnet Laboratory in Cambridge, Mass.* are working with more potent magnetic brain stimulators.
   c. *A slowing economy and its effect on corporate earnings* is the foremost concern of many traders and analysts.

(13) Long-before-short sentence-initial coordinate NPs
   a. *Last week's uncertainty in the stock market and a weaker dollar* triggered a flight to safety, he said, but yesterday the market lacked such stimuli.
   b. *The state-owned industrial holding company Institute Nacional de Industria and the Bank of Spain* jointly hold a 13.94% stake in Banco Exterior.

– In the Hawkins (1994) Constituent Recognition Domain theory, 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.



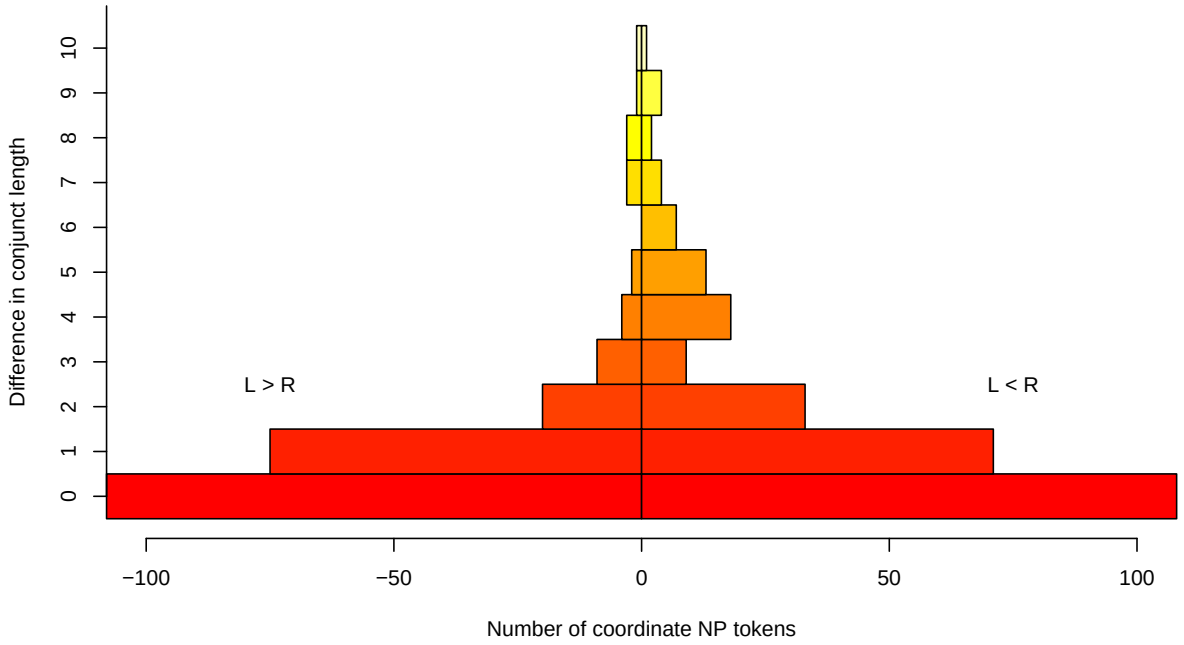


Figure 5: Distribution of sister NP conjunct size difference for sentence-initial positions, WSJ

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

  Both the SUNY team and researchers at the National Magnet Laboratory in Cambridge, Mass.

- When sentence-inital coordinate NPs are singled out, the L < R trend diminishes visibly (Figure 5). The preference for increasing conjunct weight is still highly significant overall, ($p < 0.001$) but for conjunct length difference ≤ 3, weight and ordering are no longer significantly correlated ($p = 0.20$).
- Appears to be an overlay of multiple weight-driven effects: Hawkins's domain minimization could be active everywhere, with a weaker center-embedding avoidance factor coming out in sentence initial positions

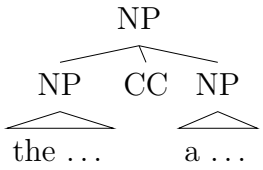
3.4 Discourse-driven information status and conjunct order

(14) Given before new in sentence-initial coordinate NPs (examples from WSJ)

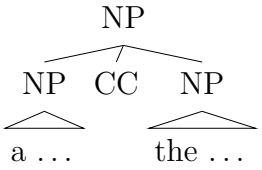
a. *Ray White in Utah and Walter Bodmer, a researcher in Great Britain*
b. *The latter two and Judge Daniel M. Friedman, 73*
c. *The city park and a street bearing the Rothschild name*

- If both weight and discourse status are active for sister NP conjunct, and given information status favors earlier expression, then we would expect that sometimes a [given,heavier] conjunct could precede a [new,lighter] conjunct (probably if the weight difference is small).
- Operationalize this by investigating NP coordinations where sister is initiated by *the* and the other by *a*.[14]

(15) 'Pre'



(16) 'Post'



- An average *greater* difference in R-L conjunct weight in the 'Post' (*A... the...*) condition than in the 'Pre' (*The... a...*) condition would support the discourse-factors-active hypothesis.
- Mean conjunct length difference is indeed significantly greater for *a* before *the* (Figure 6), suggesting that both information status and weight play a role.

[14] I assume here that definite/indefinite articles correlate reliably with given/new information status in discourse.

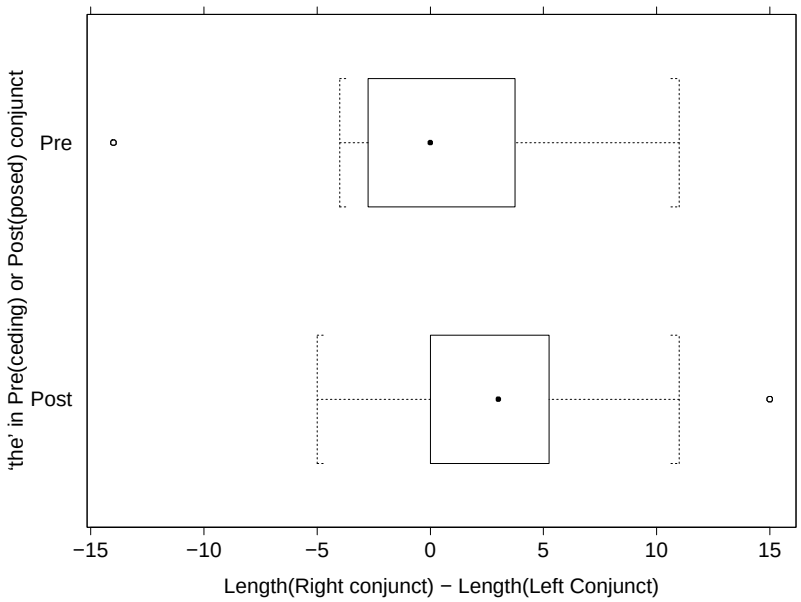


Figure 6: Difference in Length(R)-Length(L) for *the*-initial conjunct before/after *a*-initial conjunct. Center of box is median, edges of box are first and third quartiles ($p < 0.01$ by t-test; with outliers removed, $p < 0.02$)

3.5 Summary

- Coordinate NPs provide an English-internal testbed for differing theories of weight effects on linear order
- Overall English pattern of heavier after lighter also true of coordinate NPs
- Sentence-initial coordinate NPs suggest a complex interplay of domain minimization, center-embedding avoidance, and information status factors

4 Conclusion

- We have strong evidence for the cross-linguistic operation of a non-categorical *Conjoin Likes* at a variety of granularity levels, and a way to measure it
- Three separate results point to *Conjoin Likes* as operating more strongly in (formal) writing than in (conversational) speech, suggesting that parallelism is at least in part *structural* and *stylistic*;
- Coordinate NPs dramatically enrich the potential testbed for research on functional motivations for linear ordering of constituents;

- Investigation of specific syntactic position and information status indicates an interplay of memory-based, discourse-based, and center-embedding avoidance functional constraints operative

5 Further Work

- Written corpus search beyond the granularity permitted by 10^6 words, to see whether parallelism still looks like a structural preference when external governor, modifier content, modifiee content, and perhaps "topic" are simultaneously controlled for
- Other constituent types – Temperley's (p.c.) initial results suggest that L < R holds for English PPs, VPs, and clauses as well as for NPs, but no investigation yet of parallelism
- Investigate whether parallelism holds for *overall conjunct weight*
- Ambiguity avoidance as a motivation for apparent weight effects in NP coordination (Gibson and Schütze, 1999)

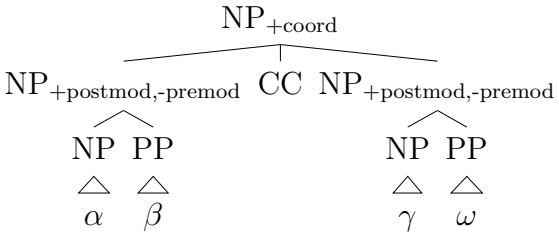
  Civilized discourse and an environment where compromise can begin
  An environment where compromise can begin and civilized discourse

- Application: coordinate scope resolution in syntactic parsing
  - Coordination, particularly NP coordination, is widely regarded as one of the most difficult aspects of probabilistic parsing (Collins, 1999), not only for English but apparently also for Chinese (Levy and Manning, 2003)

    (17) yǐqián bucéng yùdào-guò de qíngkuàng , wèntí
    before not-previously encounter.EXP conditions and problems
    'Problems and conditions heretofore unencountered'

    (18) a house with a garden and a brick wall

  - Incorporating the statistical interdependence of conjuncts could improve performance



- Closer analysis of weight effects in Chinese

– initial investigation suggests patterning is quite different

– Mean R-L length difference is slightly positive (0.13), much smaller than for WSJ (0.6)

– L<R preference seems to *decrease* with increasing weight differential (Figure 7; Chinese data highly sparse for weight differential above about 9).

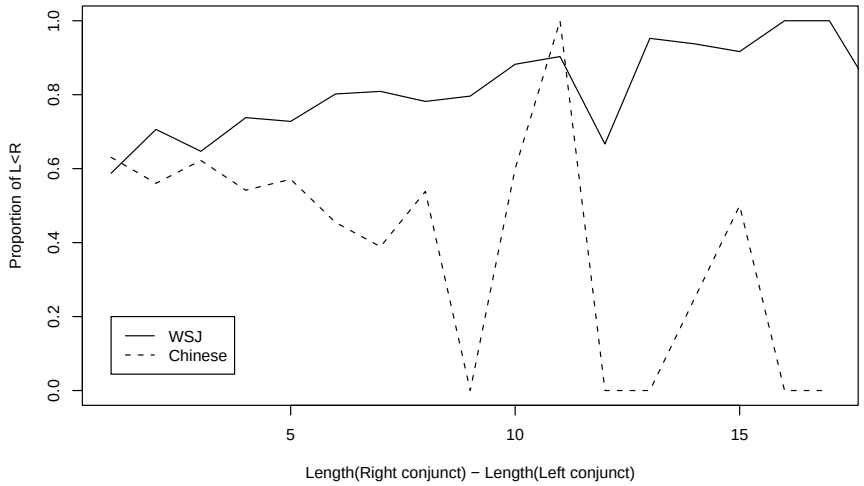


Figure 7: Comparison between WSJ and Chinese of L < R conjunct ordering preference by weight differential

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.

Bayer, S. (1996). The coordination of unlike categories. *Language*, 72(3):579–616.

Bayer, S. and Johnson, M. (1995). Features and agreement. In *Proceedings of the 1995 ACL*, pages 70–76. Association of Computational Linguistics.

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

Collins, M. (1999). *Head-Driven Statistical Models for Natural Language Parsing*. PhD thesis, University of Pennsylvania.

Dalrymple, M. and Kaplan, R. (2000). Feature indeterminacy and feature resolution in description-based syntax. *Language*, 77(4).

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.

Ingria, R. J. P. (1990). The limits of unification. In *Proceedings of the 28th Annual Meeting of the ACL*, pages 194–204. Association for Computational Lingusitics.

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

Levy, R. (2001). Feature indeterminacy and the coordination of unlikes in a totally well-typed HPSG. MS., Stanford University.

Levy, R. and Manning, C. (2003). Is it harder to parse Chinese, or the Chinese Treebank? In *Proceedings of ACL 2003*.

Levy, R. and Pollard, C. (2001). Coordination and neutralization in HPSG. In *Proceedings of HPSG 2001*. CSLI.

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.

Przepiórkowski, A. (1999). *Case Assignment and the Complement-Adjunct Dichotomy: A Non-Configurational Constraint-Based Approach*. PhD thesis, Universität Tübingen, Germany.

Sag, I. (2002). Coordination and underspecification. In *Proceedings of HPSG 2002*.

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.

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.

Wasow, T. (2002). *Postverbal Behavior*. CSLI.

Xue, N., Chiou, F.-D., and Palmer, M. (2002). Building a large-scale annotated Chinese corpus. In *Proceedings of COLING*.

Appendix: Contingency tables for modifier attachments by external governor

†of (479)	Right	
Left	HasPP	NoPP
HasPP	35	15
NoPP	51	378

*between (263)	Right	
Left	HasPP	NoPP
HasPP	5	15
NoPP	18	225

†by (220)	Right	
Left	HasPP	NoPP
HasPP	24	20
NoPP	34	142

†in (216)	Right	
Left	HasPP	NoPP
HasPP	4	2
NoPP	13	197

†with (198)	Right	
Left	HasPP	NoPP
HasPP	29	13
NoPP	38	118

†for (186)	Right	
Left	HasPP	NoPP
HasPP	12	13
NoPP	21	140

†to (182)	Right	
Left	HasPP	NoPP
HasPP	22	8
NoPP	20	132

†are (142)	Right	
Left	HasPP	NoPP
HasPP	14	2
NoPP	18	108

†as (139)	Right	
Left	HasPP	NoPP
HasPP	14	2
NoPP	10	113

Table 7: Contingency tables of NP conjunct modification by PP in WSJ, controlling for external governor of coordinate NP. *: $p < 0.05$; **: $p < 0.01$; †: $p < 0.001$

you/'from, due to' (24)	Right	
Left	HasMod	NoMod
HasMod	0	1
NoMod	1	22

zai/'at' (21)	Right	
Left	HasMod	NoMod
HasMod	3	3
NoMod	2	13

**shi*/'be' (19)	Right	
Left	HasMod	NoMod
HasMod	6	2
NoMod	0	11

**dui*/'toward' (16)	Right	
Left	HasMod	NoMod
HasMod	5	0
NoMod	2	9

**wei*/'for' (15)	Right	
Left	HasMod	NoMod
HasMod	6	3
NoMod	0	6

**you*/'have' (14)	Right	
Left	HasMod	NoMod
HasMod	5	0
NoMod	1	8

Table 8: Contingency tables of NP conjunct premodification in Chinese, controlling for external governor of coordinate NP. *: $p < 0.05$; **: $p < 0.01$; †: $p < 0.001$

of (113)	Right	
Left	HasPP	NoPP
HasPP	3	6
NoPP	35	69

between (24)	Right	
Left	HasPP	NoPP
HasPP	1	0
NoPP	1	22

by (12)	Right	
Left	HasPP	NoPP
HasPP	1	0
NoPP	1	10

in (84)	Right	
Left	HasPP	NoPP
HasPP	1	1
NoPP	13	69

with (90)	Right	
Left	HasPP	NoPP
HasPP	2	4
NoPP	13	71

for (71)	Right	
Left	HasPP	NoPP
HasPP	1	2
NoPP	21	47

to (64)	Right	
Left	HasPP	NoPP
HasPP	1	6
NoPP	13	44

are (39)	Right	
Left	HasPP	NoPP
HasPP	1	0
NoPP	2	36

as (25)	Right	
Left	HasPP	NoPP
HasPP	2	0
NoPP	11	12

Table 9: Contingency tables for NP conjunct modification by PP, Switchboard corpus.