

A quantitative look at mixed category constructions

Rob Malouf

Department of Linguistics and Oriental Languages
San Diego State University



Tenth International Conference on
Head-Driven Phrase Structure Grammar
18 July 2003

Parts of speech

- Syntactic categories are central to generative grammar, yet their nature is still unclear

- Three views of syntactic categories:

Traditional small, universal inventory of categories

Structuralist huge number of language specific categories,
organized into overlapping classes

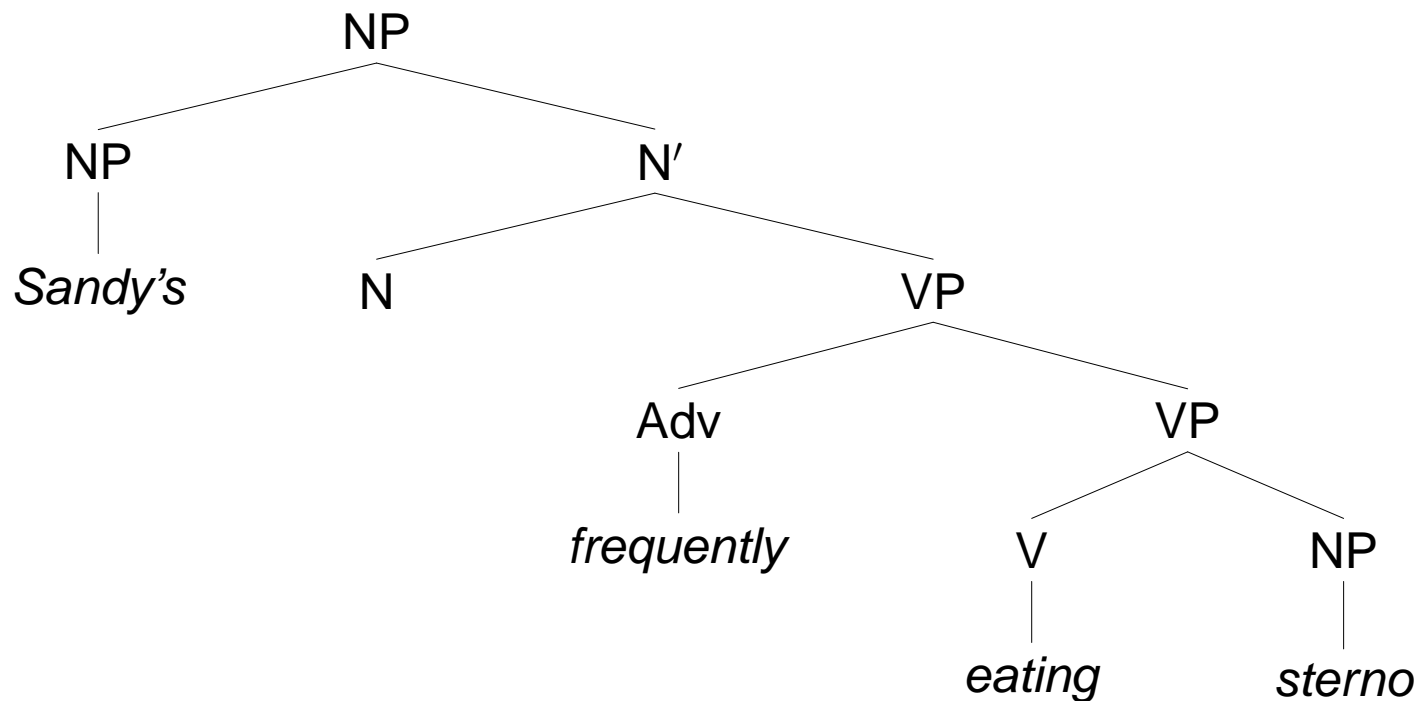
Functional small, universal inventory of prototypes

Parts of speech

- Mixed category constructions, which show properties of more than one basic part of speech, look like an empirical challenge to traditional categories:

(1) Pat worries about Sandy's frequently eating sterno.

- But:



Mixed categories

- Another class of mixed categories are words which idiosyncratically participate in constructions characteristic of more than one part of speech
- For example, *near* looks like a preposition:
 - (2) a. Don't go **near the water**!
 - b. This place is dead, but **near the city** there's lots going on.
- But, *near* also looks like an adjective:
 - (3) a. The ferry reached **the near shore**.
 - b. When you **get near to** the east end of the trail, you come to a blind hairpin turn
- Ross (1972) places *near* somewhere in the middle on a continuum between prepositions and adjectives

Transitive adjectives

- Maling (1983) argues that *near* is an adjective:

- (4) a. Kim moved the lamp { nearer
*more near } (to) the bed.
- b. Chris didn't go { near enough
*enough near } (to) the water to get wet.

- And not a preposition:

- (5) a. Kim moved the lamp { *byer
more by } the bed.
- b. Chris didn't go { *into enough
enough into } the water to get wet.

- Unlike most adjectives, though, *near* optionally selects for a NP complement

Transitive adjectives

- Huddleston and Pullum (2002) point out more prepositional properties:
 - (6)
 - a. They pushed it **right under** the bed.
 - b. *They were **right enjoying** themselves.
 - c. *I believe the employees to be **right trustworthy**.
 - d. *The project was carried through **right successfully**.
 - (7) We found it **right near** the house.
- And more adjectival properties:
 - (8)
 - a. You have put it very/too near the pool.
 - b. *You have put it very/too in the pool.
 - c. It's gotten very/too wet.
- They remark: “It is thus highly exceptional in its syntax, combining a number of adjectival properties with those of the preposition.”

Transitive adjectives

- Newmeyer (1998) argues that while *near* shows both adjectival and prepositional properties, particular **uses** of *near* generally are one or the other
- When *near* takes a *to NP* complement, then it also takes adjectival degree modifiers:
 - (9) a. The gas station is **near to** the supermarket.
 - b. The gas station is **near enough to** the supermarket.
 - c. *The gas station is **right near to** the supermarket.
- When *near* takes a bare *NP* complement, then it also takes prepositional degree modifiers:
 - (10) a. The gas station is **near** the supermarket.
 - b. *The gas station is **near enough** the supermarket.
 - c. The gas station is **right near** the supermarket.

Transitive adjectives

- Newmeyer's claim is that *near* can be an adjective or a preposition, but not both
- It does show limited morphological mismatch:

(11) The gas station is nearer (to) the supermarket than the bank.
- Thus, *near* provides no evidence against the traditional part of speech theory of categories.
- Also no evidence for prototype categories

Transitive adjectives

- Newmeyer's argument leaves some open questions
- What about the adjectival degree modifiers like *very*?
 - (12) a. The gas station is **very near (to)** the store.
- Subtle and controversial grammaticality judgments:
 - (13) a. The gas station is **near enough (to)** the supermarket.
 - b. The gas station is **right near (to)** the supermarket.

Testing grammaticality

- To help resolve close grammaticality calls, we could look for natural occurrences in a corpus
- No examples in the North American News Text Corpus
- A Google search turns up 589 examples:
 - (14)
 - a. We camped **right near to** Acadia National Park, just outside of Bar Harbor.
 - b. It is located **right near to** Samurai, so two of the best rides in the park are right close together.
 - (15)
 - a. Having reached Barco, turn **right near to** the small church, thus arriving at the castle with its splendid angular tower.
 - b. Further on towards the dale of Hufield there is a child's grave on the **right near to** the burn.

Testing grammaticality

- Do a handful of Google citations constitute evidence for grammaticality?
- An observation: we find occurrences of *right near to* in unedited text (Google) but not in edited text (NA News Text Corpus)
- It is tempting to conclude that *right near to* is, strictly speaking, ungrammatical.
- But: the NA News Text Corpus is around 800 million words, while the corpus indexed by Google is on the order of 30 trillion words

Testing grammaticality

- We have a hypothesis: the difference in frequency of *right near to* in the two corpora reflects an underlying difference in their text types

	Text Type	
	Edited	Unedited
<i>right near to</i>	0	589
<i>right near \negto</i>	40	85,711

- Chi-squared test is not applicable
- Fisher's exact test strongly indicates independence ($P \approx 1$)
- There is no evidence that *right near to* is ungrammatical

Testing grammaticality

- It can be difficult to settle questions of grammaticality by studying a corpus
- Instead, we can use corpus data to directly examine properties of syntactic categories
- Is there evidence for distinct adjectival and prepositional uses of *near*?
- What is the relationship between the type of modifier (*very, right*) and the complement type (*NP, to NP*)?
- What is the relationship between the external distribution of *near* phrases and their internal distribution?

Generalized Linear Models

- Generalized Linear Models (GLMs) provide a mechanism for investigating the nature and magnitude of relationships among linguistic properties
- GLMs extend regression and ANOVA techniques to count and other type of data
- Much more informative than chi-squared test, which only tells you *if* there is a relationship
- GLMs are applicable to a wide range of experimental designs and sampling strategies
- Closely related to logistic regression (VARBRUL) and maximum entropy models
- Misclassification error may be a source of bias

Generalized Linear Models

- For a 2×2 contingency table:

	<i>X</i>	<i>not X</i>	Total
<i>Y</i>	50	100	150
<i>not Y</i>	500	1000	1500
Total	550	1100	

we model the count μ_{ij} in each cell by:

$$\log \mu_{ij} = \lambda + \lambda_i^X + \lambda_j^Y + \lambda_{ij}^{XY}$$

Generalized Linear Models

- Fitting a model means finding λ 's such that the counts are accurately predicted by

$$\log \mu_{ij} = \lambda + \lambda_i^X + \lambda_j^Y + \lambda_{ij}^{XY}$$

- The intercept parameter λ reflects the size of the corpus
- The *main effect* parameters λ^X and λ^Y reflect the marginal totals for X and Y
- The *association parameter* λ^{XY} captures any statistical dependence between X and Y

Generalized Linear Models

- For this example:

λ	6.908
λ^X	-0.693
λ^Y	-2.303
λ^{XY}	0.000

- In words:
 - ★ there is a small preference for $\neg X$ over X
 - ★ there is a larger preference for $\neg Y$ over Y
 - ★ there is no association at all between X and Y

Generalized Linear Models

- We search Google for sequences X near Y and record the counts:

	<i>to</i>	not <i>to</i>
<i>right</i>	580	85,420
<i>so much</i>	8	215
<i>very</i>	28,800	430,800
<i>so</i>	63,200	134,200
other	640,666	5,178,002

- Fitting the model yields the parameter estimates:

(Intercept)	15.46	right:to	-2.902
right	-4.109	so much:to	-1.201
so much	-10.09	very:to	0.171
very	-2.491	so:to	0.551
so	-3.658		
to	-2.090		

Interpreting the model

- Intercept and main effects are not interesting
- Association parameters:

right:to	-2.902	very:to	0.171
so much:to	-1.201	so:to	0.551

- This shows a **strong negative association** between *to* complements and the prepositional modifiers *right* and *so much*
- There is also a **positive association** (though less strong) between *to* complements and the adjectival modifiers *very* and *so*
- These results show that prepositional properties appear together (*right near the store*) and adjectival properties appear together (*so near to the store*)
- Mixtures of prepositional and adjectival properties are possible, but occur less frequently

Generalized Linear Models

- How does the external distribution relate to the internal distribution?
- Select 167 occurrences of *near to NP* and 161 occurrences of *near NP* from NA News Text Corpus
- Annotate each with syntactic role (NP modifier, VP modifier, predicative) and for presence of a degree modifier
- Results:

		NP mod.	VP mod.	Pred.
Deg. mod.	<i>near NP</i>	1	1	0
	<i>near to NP</i>	6	9	22
No deg. mod.	<i>near NP</i>	87	55	17
	<i>near to NP</i>	26	21	83

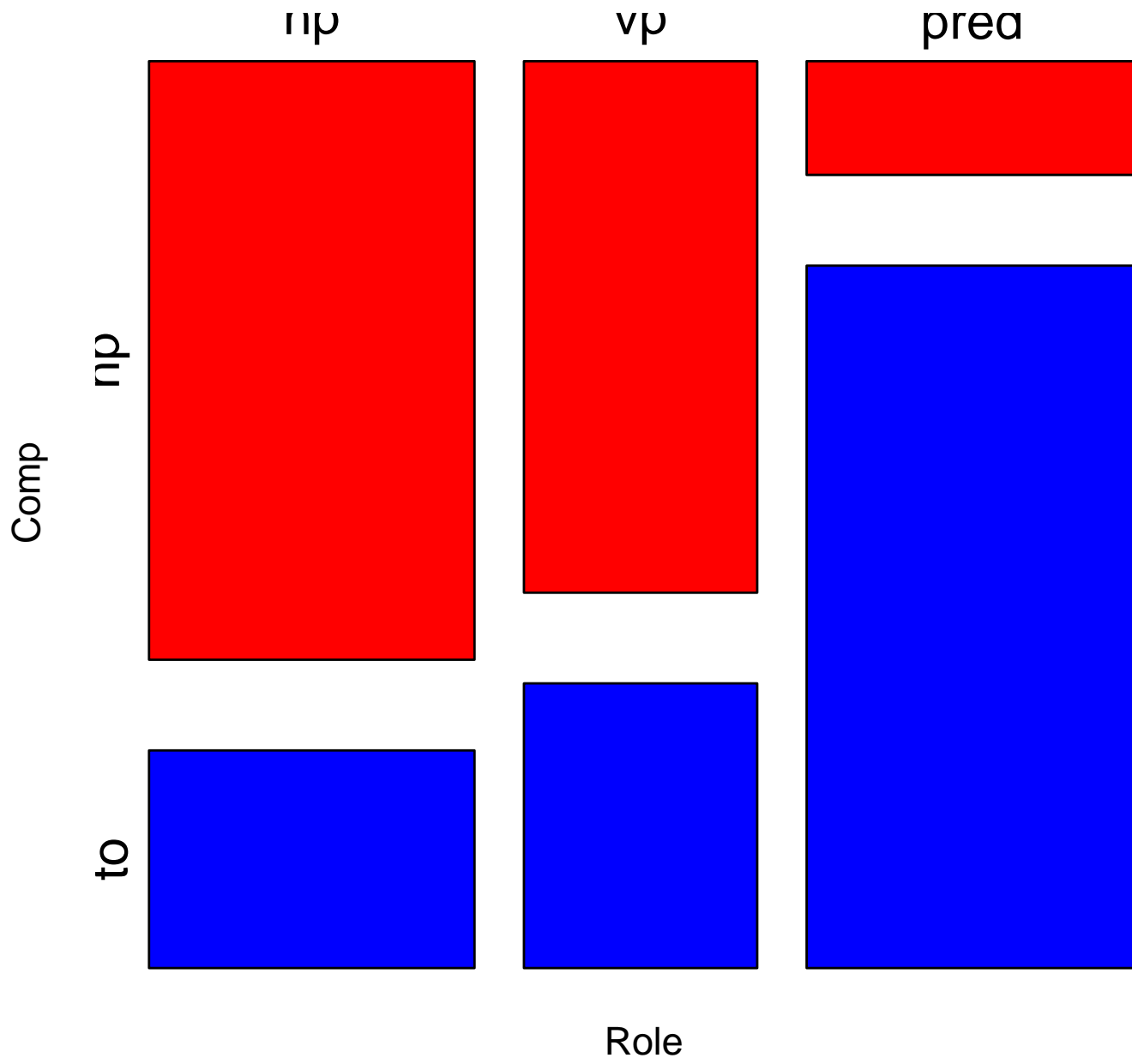
Generalized Linear Models

- The model:

(Intercept)	2.822	np:mod	—
np	1.645	vp:mod	—
vp	1.185	np:to	-2.814
mod	-4.534	vp:to	-2.563
to	1.599	mod:to	3.196

- **Strong association** between predicative contexts and *near to NP*, and between modifier contexts and *near NP*
- **Strong association** between (adjectival) degree modifiers and *near to NP*
- No independent association between context and modifier

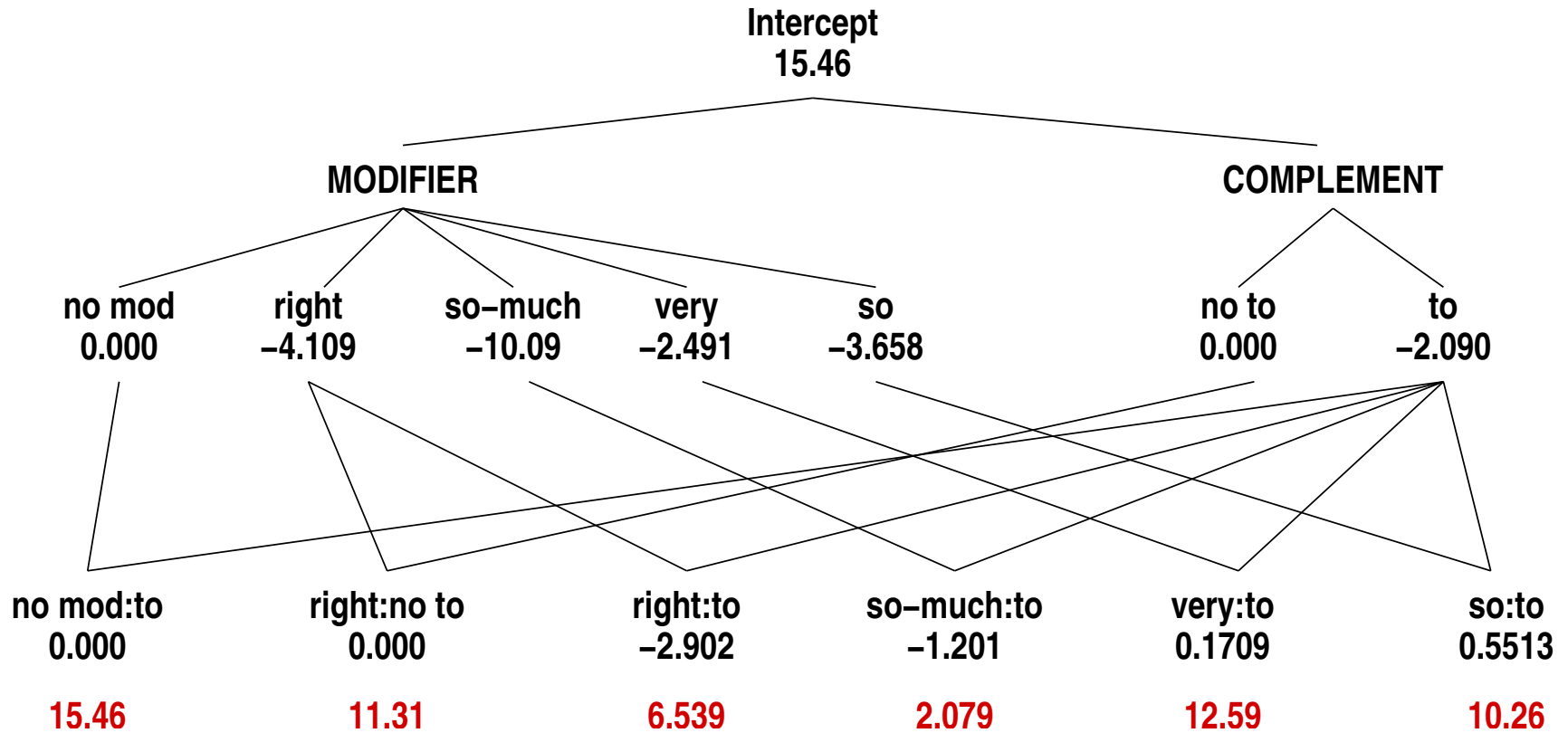
NANC



Some conclusions

- Newmeyer was right (sort of)
 - ★ Both experiments show a clear preference for strictly adjectival or strictly prepositional uses of *near*
- Huddleston and Pullum were right (sort of)
 - ★ Mixed uses of *near* are well attested, and are not even particularly rare
- Neither approach to syntactic properties is able to capture all the facts
- Results are consistent with a **prototype** model (Ross and many others)
- What is the correct formal representation for syntactic categories?

Hierarchical models



Conclusions

- Simply eyeballing a corpus may not shed light on problematic grammaticality judgments
- Quantitative analysis may reveal regularities that are not obvious from patterns of grammaticality judgments or from casual inspection of a corpus
- With *near*, we see evidence for a prototype effect: mixed uses are possible, but dispreferred
- Speculation: A combination of inheritance hierarchies with generalized linear models may be a way of capturing these prototype effects in a formal grammar