

# Nominal categories and the expression of possessors:

A crosslinguistic study of  
probabilistic tendencies and  
categorical constraints

Cathy O'Connor

Boston University

Joan Maling

Brandeis University & NSF

Barbora Skarabela

University of Edinburgh

April 4, 2009    University of Manchester

# 1. Introduction

# The Stochastic Generalization

Statistically noticeable but noncategorical patterns found in one language are often found in other languages in categorical and relatively inviolable form.

Bresnan, Dingare & Manning 2001  
Manning 2002

Givón (1979:28) contrasts a categorical restriction against indefinite subjects in Krio with the dispreference for them in English:

But are we dealing with two different kinds of facts in English and Krio? Hardly. What we are dealing with is apparently the very same *communicative tendency* -- to reserve the subject position in the sentence for the *topic*, the old-information argument, the "continuity marker." In some languages (Krio, etc.) this communicative tendency is expressed at the *categorical* level of 100%. In other languages (English, etc.) the very same communicative tendency is expressed "only" at the *noncategorical* level of 90%.

(cited by Manning 2003:316)

# The Stochastic Generalization

*So what?*

*Manning: To the extent that this is true, it has consequences for how we model grammars.*

*Also, it may provide insight into language learning and language change...*

# The Stochastic Generalization

Today: we'll present evidence that statistical patterns in usage of the English *s*-genitive are found in categorical form in a variety of Indo-European languages.

Along the way we'll ponder

- probabilistic versus categorical phenomena
- the factors driving speakers' choice of alternant and their relation to NP form classes

## 2. Probabilistic patterns in the English genitive alternation: ... a corpus study

This study: 10,000 examples of *s*-genitives and *of*-genitives from the Brown corpus (using a version POS-tagged by Fred Karlsson)

(Carried out as part of a project "Optimal typology of the DP: Markedness within the noun phrase" NSF award to Boston University, BCS-0080377, O'Connor, Anttila, Fong, Maling; 2000-2003)



Atlanta's mayor

The mayor of a New England town

Jean's car

The car of a neighbor

Rebecca's many virtues

The many virtues of walking

What drives the alternation?

Atlanta's mayor

The mayor of a New England town

Jean's car

The car of a neighbor

Rebecca's many virtues

The many virtues of walking

Three hypotheses:

# Hypothesis 1: Weight

Atlanta's mayor

***S-Genitive* possessor lighter**

The mayor of a New England town

***Of-Genitive* possessor heavier**

(Principle of end-weight; Stefanowitsch, also  
cf. Arnold et al., J.Hawkins, Wasow)

## Hypothesis 2: Discourse Status

Jean's car

***S-Genitive* possessor more accessible**

The car of a neighbor

***Of-Genitive* possessor less accessible**

Presumably a case of the widely observed 'old before new' tendency seen in many constructions (Deane, Anschutz).

# Hypothesis 3: Animacy

My niece's many virtues

***S-Genitive* possessor animate**

The many virtues of walking

***Of-Genitive* possessor inanimate**

The motivation for this tendency is controversial.

Important questions about these 3 factors  
(in this alternation and others):

Are they **independently** affecting  
the choice of alternant?

If they are independent,  
can we tell which one is **most important**?

The independence issue is tough:

Factors are confounded...

Her

second straight victory

Humans are  
often topical

Pronouns  
are light

Topics tend to  
get repeated  
and become  
discourse-old.

Pronouns  
index  
discourse-  
old entities



These confounds pose both methodological and conceptual problems.

- *Methodological*: how can we control for one factor when looking at the effects of another?
- *Conceptual*: How should we understand the reasons for these factors' co-occurrence? Why do they 'travel together'?

How about experimental studies?

Is that a way to control for the confounds?

How about experimental studies?

Is that a way to control for the confounds?

Experiments allow control of the 3 factors and others through direct manipulation. Rosenbach (2002, 2005) was the first to systematically control these three factors for the genitive alternation using experimental methods. (In addition, Rosenbach limited the stimuli to a few semantic types in order to control the relation between the head and modifier.)

**Table 1: Rosenbach's factors determining English genitive variation (Rosenbach 2003; Jäger & Rosenbach 2005)**

<b>Factors</b>	<b>s-genitive more likely</b>	<b>of-genitive more likely</b>
<b>animacy</b>	+animate possessor <i>the boy's eyes &gt; the eyes of the boy</i>	–animate possessor <i>the frame of the chair &gt; the chair's frame</i>
<b>topicality</b>	+ topical possessor <i>the boy's eyes &gt; the eyes of the boy</i>	–topical possessor <i>the headlamps of a car &gt; a car's headlamps</i>
<b>possessive relation</b>	+ prototypical possessor <i>the boy's eyes &gt; the eyes of the boy</i>	–prototypical possessor <i>the condition of the car &gt; the car's condition</i>

**Table 2: Preference for English s-genitive  
(Rosenbach 2002)**

[+animate]				[-animate]			
[+topical]		[-topical]		[+topical]		[-topical]	
[+proto]	[-proto]	[+proto]	[-proto]	[+proto]	[-proto]	[+proto]	[-proto]
<i>the boy's eyes/ the eyes of the boy</i>	<i>the mother's future/ the future of the mother</i>	<i>a girl's face/ the face of a girl</i>	<i>a woman's shadow/ the shadow of a woman</i>	<i>the chair's frame/ the frame of a chair</i>	<i>the bag's contents/ the contents of a bag</i>	<i>a lorry's wheels/ the wheels of a lorry</i>	<i>a car's fumes/ the fumes of a car</i>

**s-genitive > of-genitive**

**of-genitive > s-genitive**



**more s-genitive**

**less s-genitive**



Notice: Rosenbach found that **ANIMACY** was the most important factor.

It outweighed the importance of WEIGHT or TOPICALITY.

But she did not include any **pronouns** in her materials, because they were "categorical" -- not licensed in the **of**- genitive.

If she had asked subjects to choose between

**"her face"** and **"the face of her"**

they wouldn't have had much choice.

So she excluded pronominal possessors.

In our corpus, we did have actual examples with pronouns in the *of*-genitive, like

*to the west **of him***

*on the face **of it***

In our corpus, we did have actual examples with pronouns in the *of*-genitive, like

*to the west **of him***

*on the face **of it***

So we thought we could include pronouns, and maybe get another perspective on the Animacy vs. Discourse Status problem. By using the tool of **logistic regression**, we might be able to see how pronouns contribute to the probability of an expression showing up as an *s*-genitive or *of*-genitive.



But before we get to the logistic regression, corpus studies have another problem that experimental studies can avoid by their very design...

In a corpus, you have to decide which tokens represent real possibilities for alternation between the Of-genitive and S-genitive:

The glass of water

~~Water's glass~~

There are many, many such distractors, some obvious, some subtle. And these complicate the problem of exploring factor independence.

## 2a. Cleaning the sample

# First: exclude non-nominals.

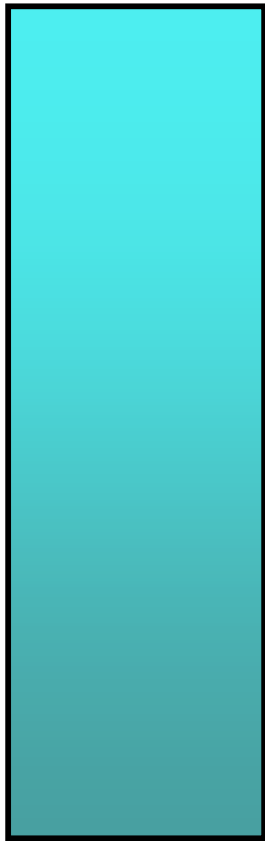
*A few examples:*

Verbal Of-NP: ~~*He thought of her.*~~

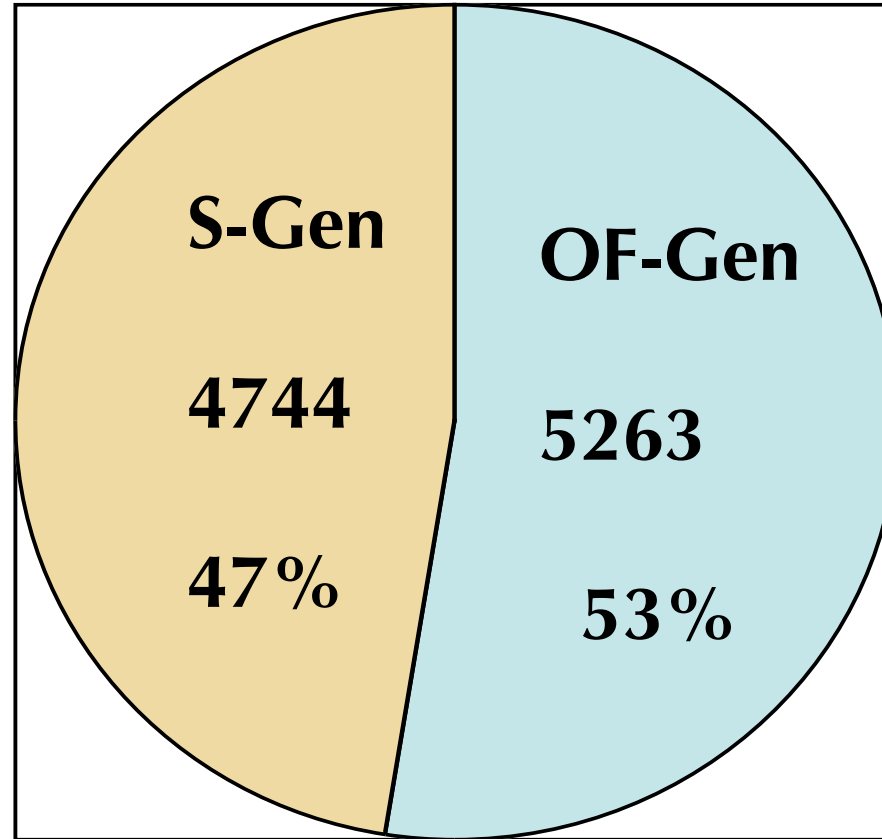
Adjectival Of-NP: ~~*bald and afraid of women.*~~

Contraction NP'S: ~~*Kate's all right.*~~

"All NP" sample, after removal  
of non-nominal examples



$N = 10,006$



Second: exclude all tokens of  
non-reversible constructions

*A few examples:*

**Partitives:**

*half of his stirrup guard*

**Measure and  
container phrases:**

*a drop of liquor*

*two saucers of water*

**Classifier phrases:**

*a grove of trees*

*a flight of wooden steps*

**Configuration and  
constitutive phrases**

*strips of skin*

*a...castle of pine boughs*

Second: exclude all tokens of  
non-reversible constructions

**'Sort' phrases** *the crassest kind of materialism*

**'Headless' Gens:** *that \_\_\_ of a frustrated gnome  
but only Kennedy's \_\_\_ did.*

**Indefinite Of-Gens:**

*a relative of the president*  
 $\neq$  *the president's relative*

**Nominal  
compounds:**

*dog-eared men's magazines*

**and many others...**

Third: exclude tokens where reversal substantially alters meaning or reference--'soft' non reversibles

## **Idioms, fixed phrases, and titles**

~~*bachelor of science*~~    *\*science's bachelor*

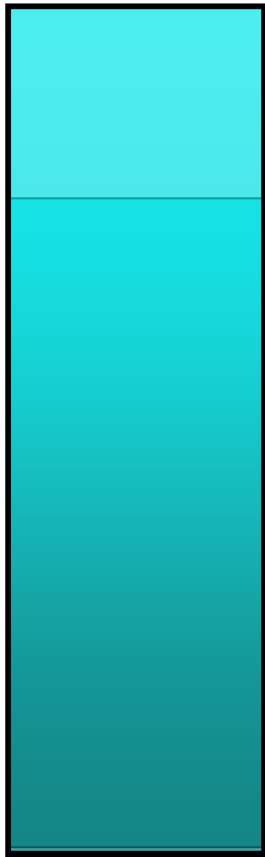
~~*Satan's L'il Lamb*~~    *#the L'il Lamb of Satan*

## **Deverbal nominals with argument constraints**

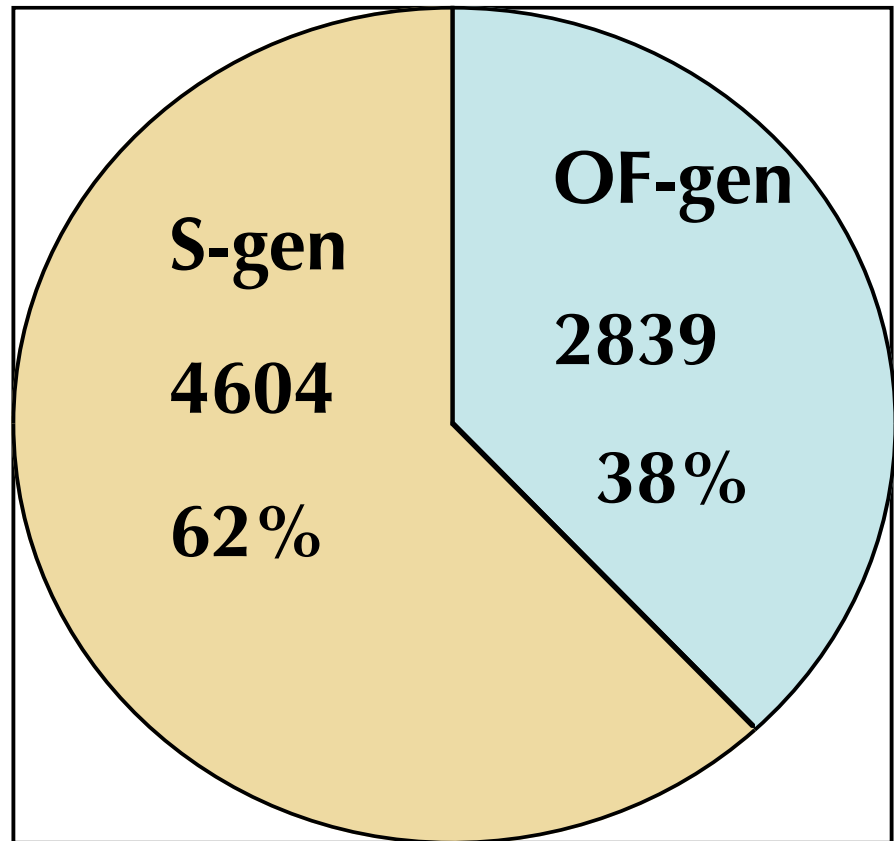
~~*fear of him*~~     $\neq$     *his fear*



Partially clean sample after removal  
of 'strict' and 'soft' non-reversibles



$N \equiv 70,506$



## 2b. Coding the sample

Today we'll just talk about our coding of modifiers-- the "possessor"

My sister's house

The house of my sister

Each was coded for weight, animacy, and discourse status.

## Coding for Weight:

Arnold et al., Wasow, and J. Hawkins suggest that the [orthographic] word is a reasonable measure of weight for most purposes.

It is also easily automated.

Each head and modifier were coded for weight in words, from 1 through >20.

# How to code for Discourse Status?

Even simple codes such as 'New', 'Inferable', and 'Old' are quite time-consuming, although they are clearly desirable.

With thousands of tokens, we chose instead to use **NP form as a proxy**: to exploit certain robust relationships between NP form and discourse status or accessibility.

Relying on previous research of Prince, Gundel et al., Ariel, *i.a.*, we coded possessors for **NP form** and for morphosyntactic definiteness.

# Coding for NP Form and Definiteness:

***Pronoun***

***Proper Noun***

***Kinship Term***

***Common Noun Definite***

***Common Noun Indefinite***

**Most accessible,  
most topical,  
discourse-old...**

**Least accessible,  
least topical,  
discourse-new...**

# Coding for Animacy:

## Elaborated Animacy Code

- *Human(oid)s*
- *Animals*
- *Human organizations*
- *Concrete objects*
- *Locations*
- *Temporal entities*
- *'Nonconcrete' entities*

## Simplified Animacy Code

***ANIMATE***

***ORG***

***INANIMATE***

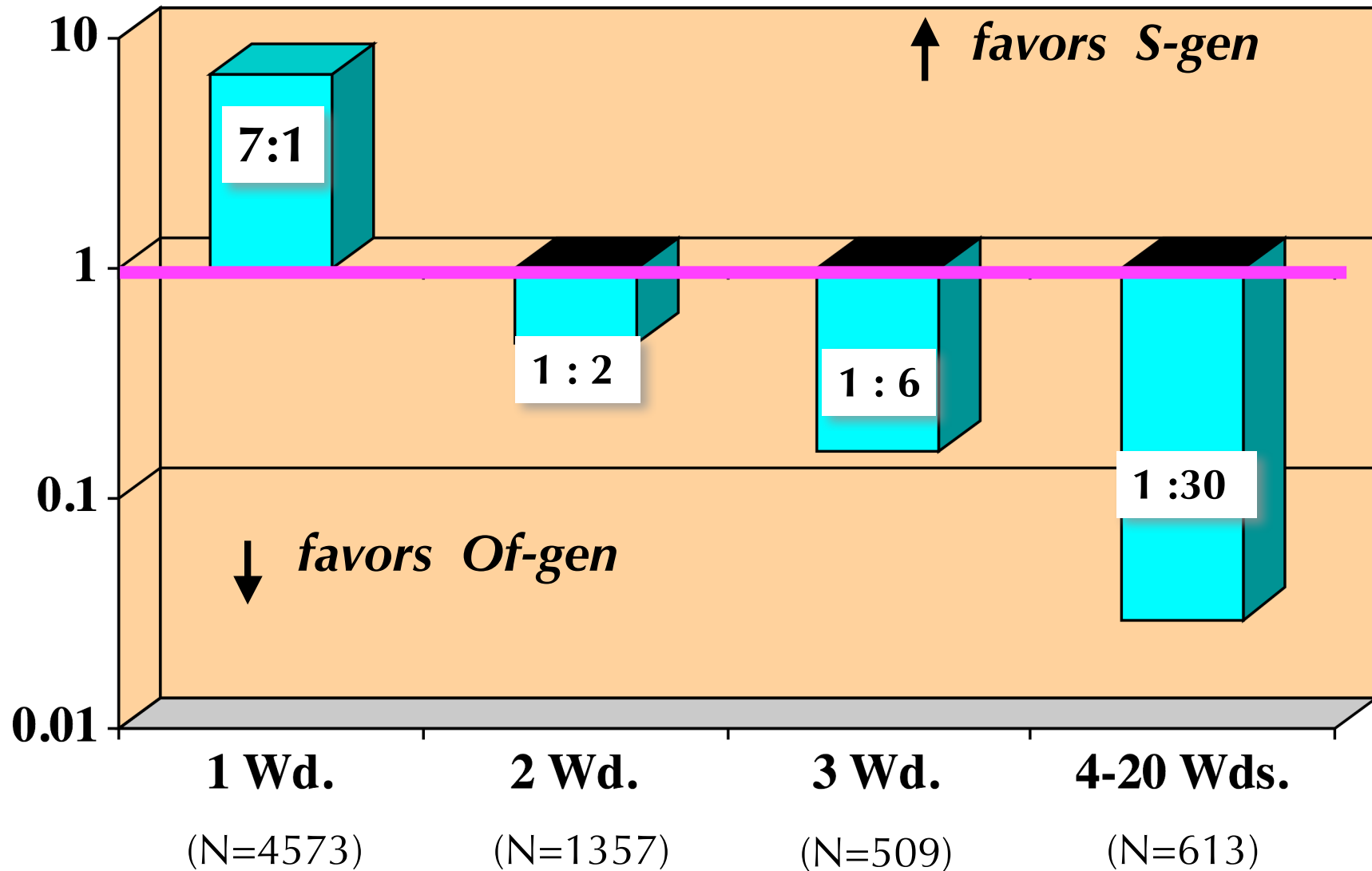
2c. Some results:

Visualizing one variable  
at a time

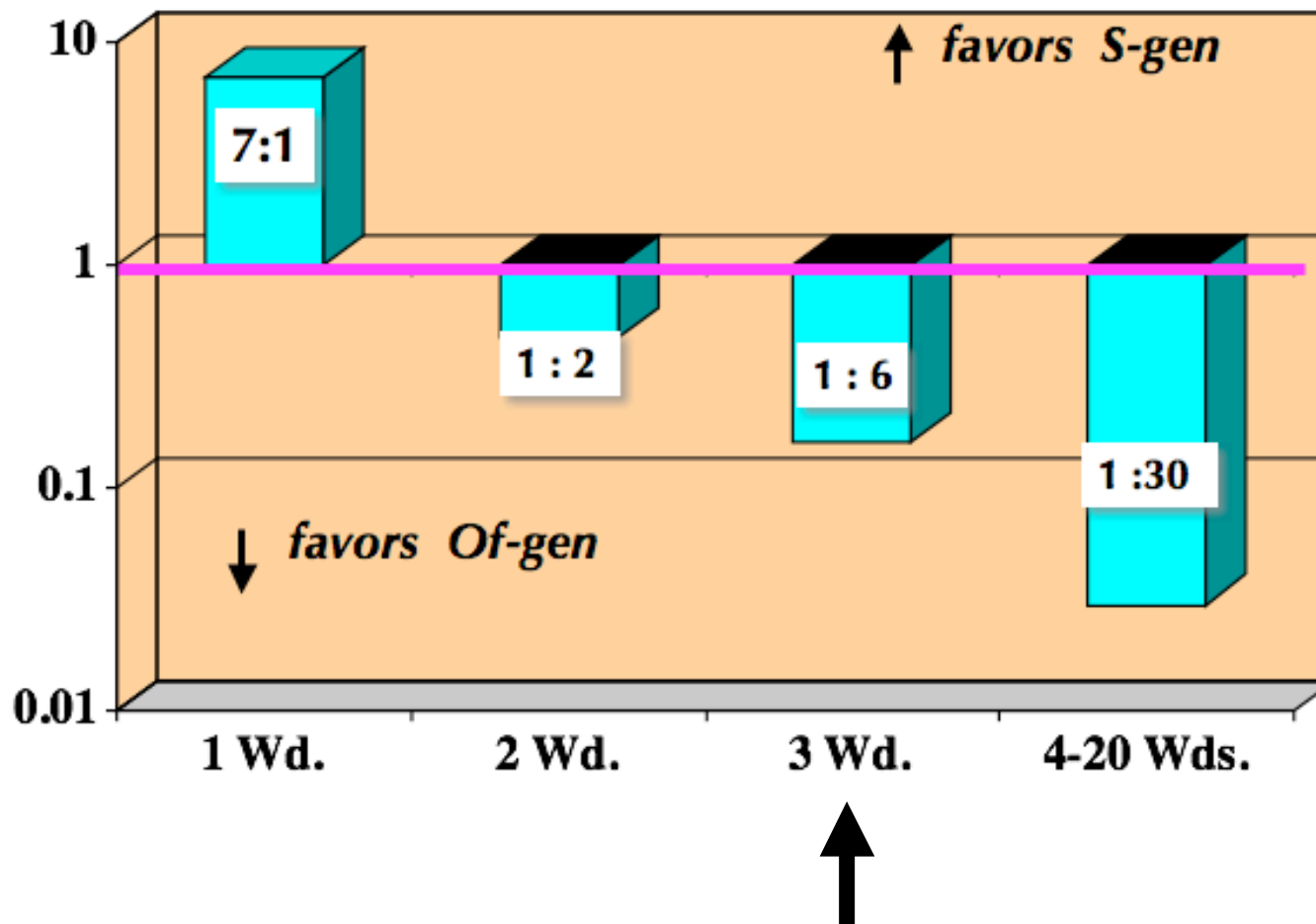


**Logistic regression** is a great tool, but it is sometimes hard to visualize what is going on. So first, we will show you some results in a form that is more accessible than the output from a regression.

Odds of *s*-genitive over *of*-genitive  
by possessor **Weight** (n=7052)

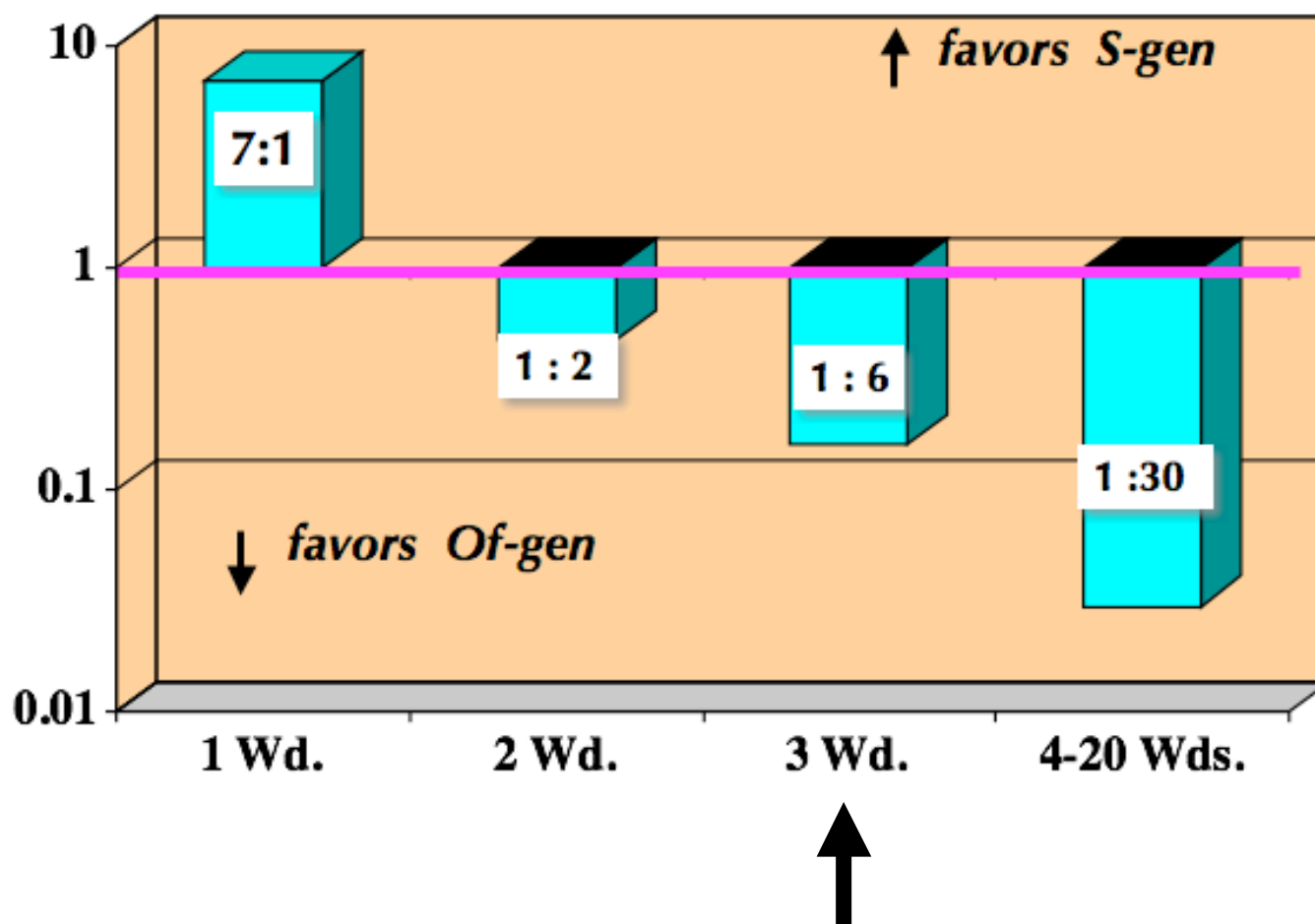


So what does this mean?

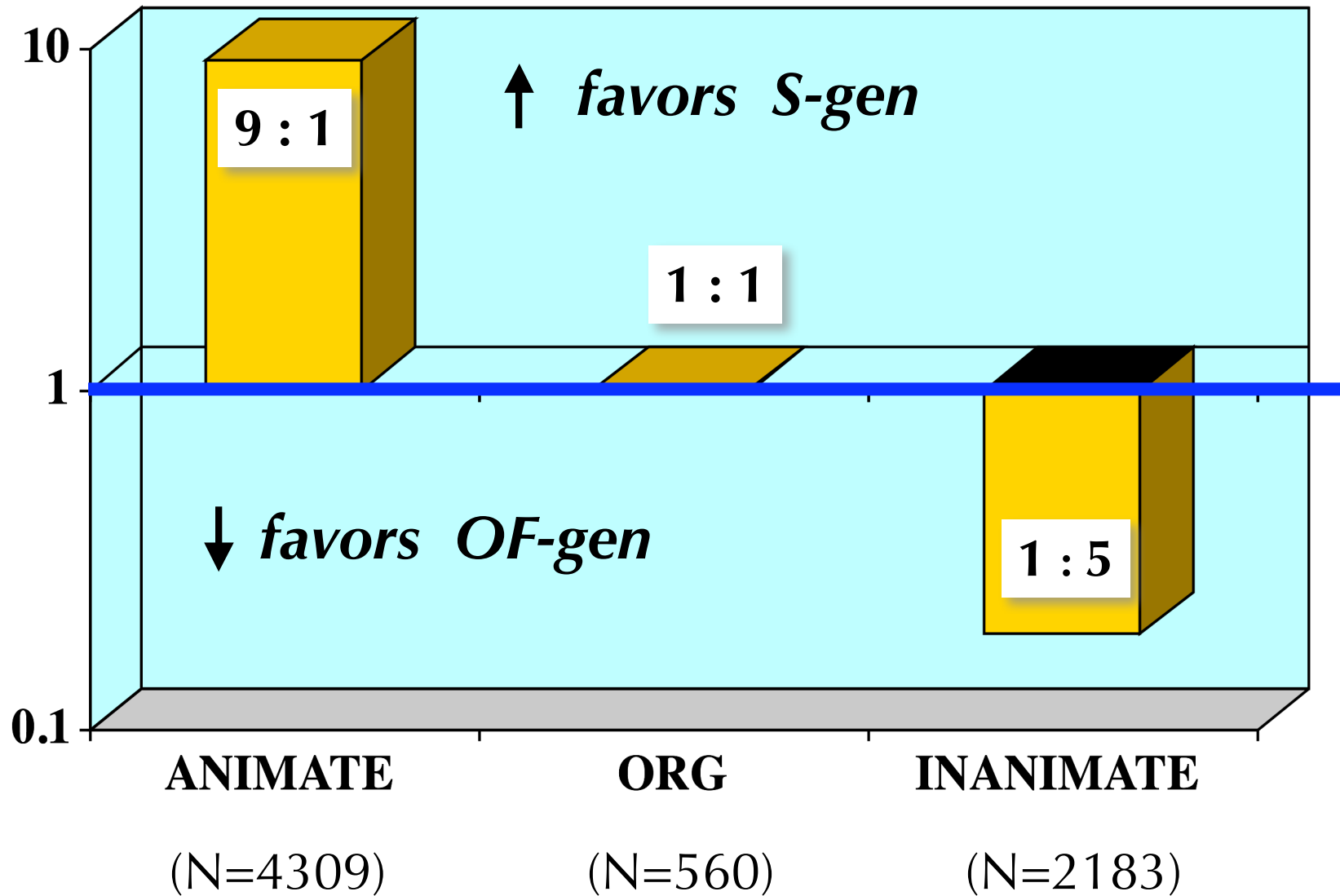


This says that in our sample of 7000 tokens from the Brown corpus, a 3-word-long possessor is six times more likely to end up as an *of*-genitive than to end up as an *s*-genitive:

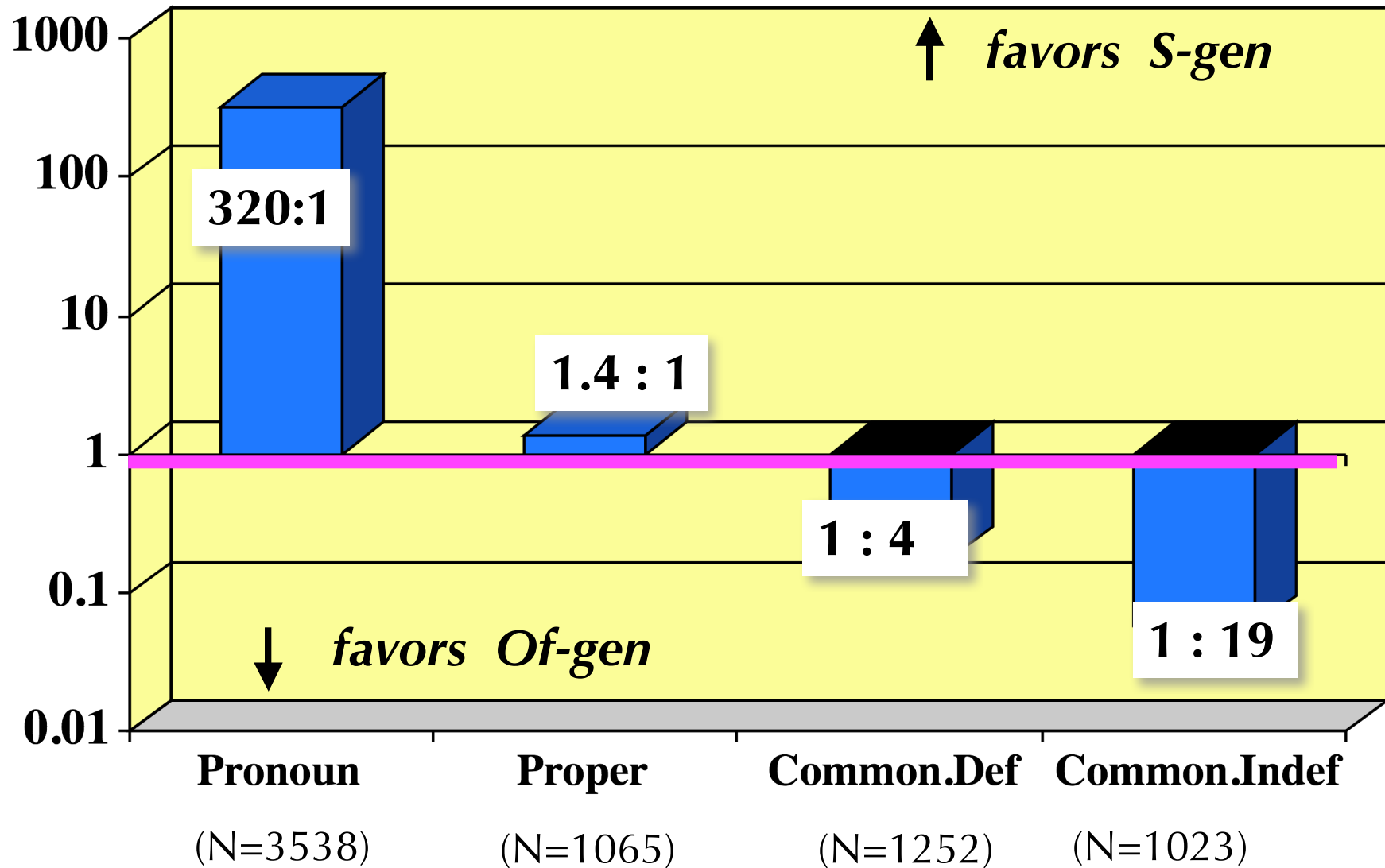
***The toys of the youngest children***



Odds of *s*-genitive over *of*-genitive  
by possessor **Animacy** category (n=7052)



Odds of *s*-genitive over *of*-genitive  
by NP form type (n=7052)



These graphs seem to show that the factors we've identified have an effect on speakers' choice of alternant.

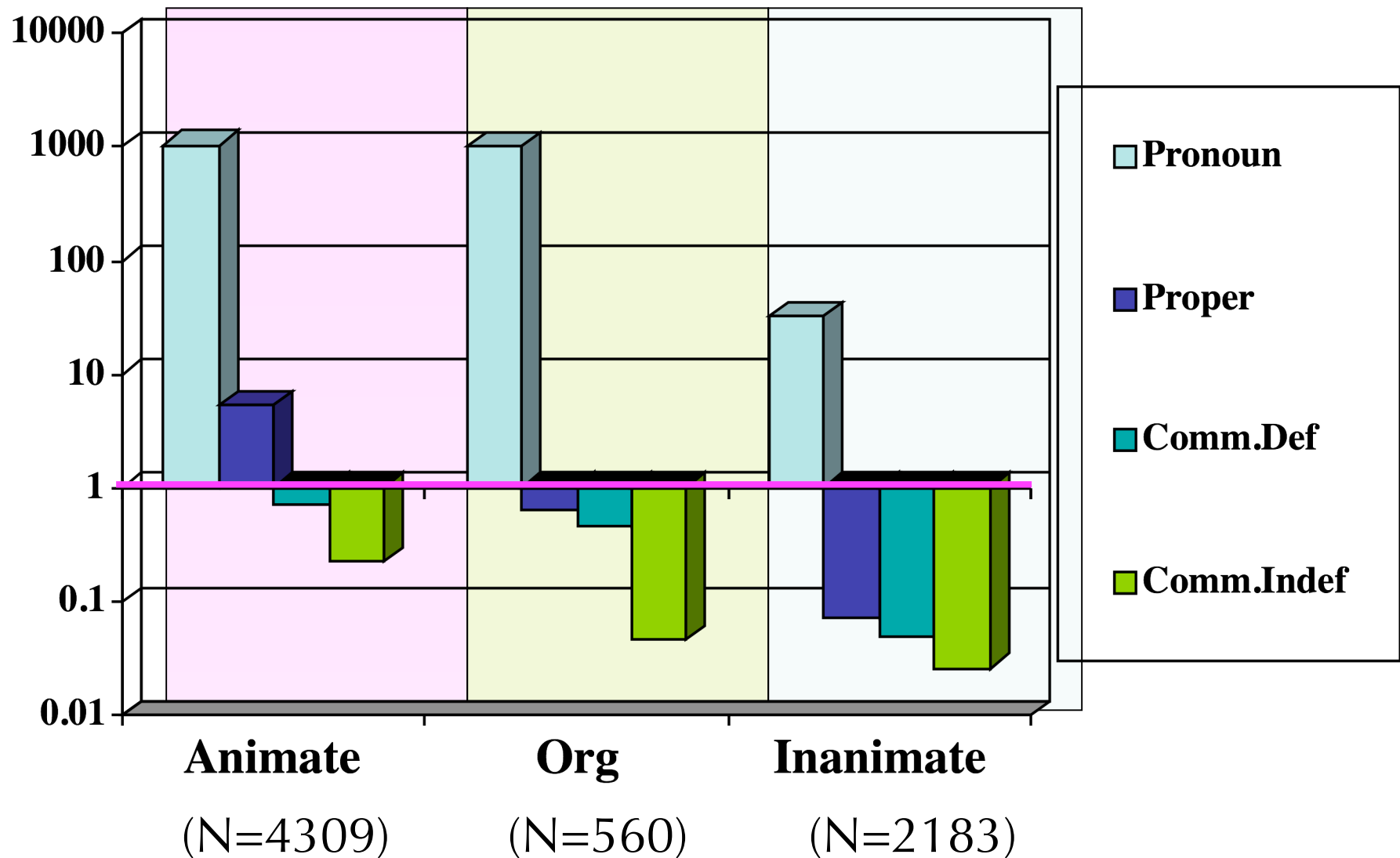
(Strictly speaking, these are not the regression results; they are hand-calculated odds ratios for the data--the data that were later subjected the logistic regression.)

But these graphs showed each factor one by one. They can't show us whether the same effects hold if we control for the other factors.

This is where the logistic regression comes in handy.  
But first, to mentally prepare for it, we can visually  
inspect the effects of one factor holding another  
constant...



# Odds of *s*-genitive over *of*-genitive by **NP form**, controlling **Animacy** (n=7052)



These patterns appear to suggest *independent* contributions by the different factors.

But a logistic regression can tell us whether the factors are statistically independent and what the magnitude of their contribution is.

(At least that is the hope.)

Logistic regression is a statistical procedure designed to predict binary categorical outcomes like this one:

Is a particular example more likely to be expressed as an *s*-genitive or an *of*-genitive?

Logistic regression is a statistical procedure designed to predict binary categorical outcomes like this one:

Is a particular example more likely to be expressed as an *s*-genitive or an *of*-genitive?

(It's been useful for actuaries and public health researchers, among others; those who need to use dichotomous categorical outcomes:

**Dead vs. Living;**  
**Employed vs. Unemployed;** etc.)

2d. Scenes from a  
logistic regression

We want to see to what extent a model with all three factors, Weight, Animacy, and NP Form, can predict the occurrence of a token as either an S-genitive or an Of-genitive.

# Possessor Variables

## **Possessor Animacy**

**(3 levels):**

Inanimate

Org

Animate

## **Modifier Weight (4 levels):**

1-word

2-word

3-word

4– >20-word

## **Modifier Expression Type (6 levels):**

Common Indefinite

Common Definite

Gerund

Kinship

Proper

Pronoun

Using all 7052 tokens, we ran a logistic regression.



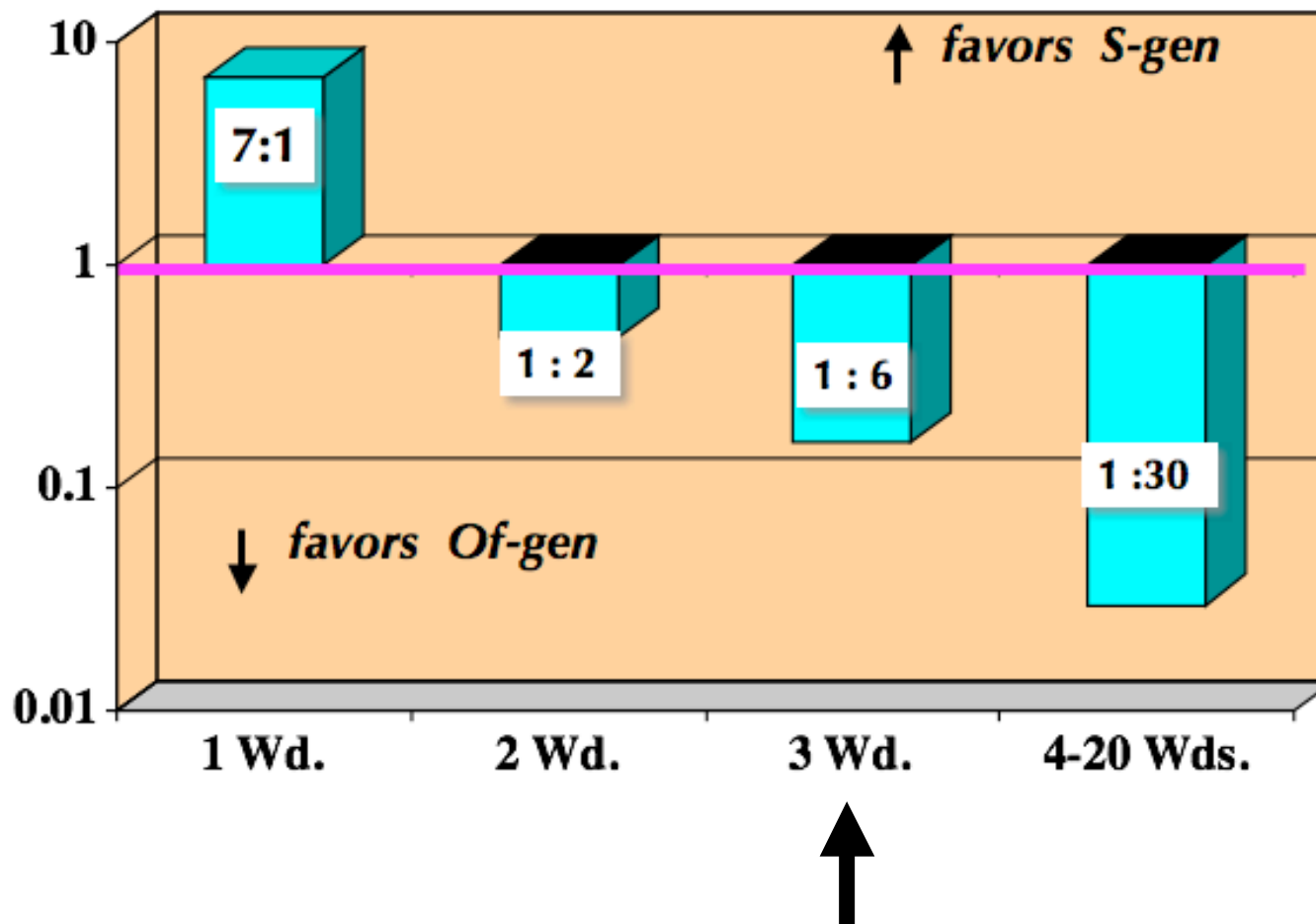
Using all 7052 tokens, we ran a logistic regression.

The logistic regression procedure examines every token and takes into account all of its coded features. It then constructs an equation. This regression equation tells us, for every token, what the probability is that it will be an S-genitive or an OF-genitive, compared to a reference case.

## Remember this?

A 3-word-long possessor is six times more likely to end up as an *of*-genitive than to end up as an *s*-genitive:

*The toys of the youngest children*



Here is the same type of information in a logistic regression output:

Predictor Variable	Wald Statistic	Odds Ratio (Exp (B))	95% CI Lower	95% CI Upper
<b>Mod.Wgt</b>	<b>125.49***</b>			
Mod. Wgt. 1 (2 wds)	13.01***	2.02	1.38	2.97
Mod.Wgt. 2 (3 wds)	59.52***	7.93	4.68	13.43
Mod.Wgt. 3 (4→20 wds)	90.22***	31.88	15.60	65.13

\*\*\*p<.0001

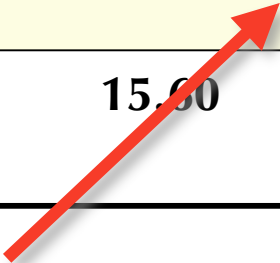
The Odds Ratio of 7.93 means that a token with a possessor that is 3 words long (*the shrieks of **excited kindergarten students***) is almost 8 times more likely to end up in the *Of*-genitive than is a token with a one-word possessor (***students'** shrieks*), **holding all other factors constant**.

Predictor Variable	Wald Statistic	Odds Ratio (Exp (B))	95% CI Lower	95% CI Upper
<b>Mod.Wgt</b>	<b>125.49***</b>			
Mod. Wgt. 1 (2 wds)	13.01***	2.02	1.38	2.97
Mod.Wgt. 2 (3 wds)	59.52***	<b>7.93</b>	<b>4.68</b>	<b>13.43</b>
Mod.Wgt. 3 (4→20 wds)	90.22***	31.88	15.60	65.13

\*\*\*p<.0001

The Confidence Interval indicates that we can be 95% certain that the true value of that particular Odds Ratio lies between 4.68 and 13.43.

Predictor Variable	Wald Statistic	Odds Ratio (Exp (B))	95% CI Lower	95% CI Upper
<b>Mod.Wgt</b>	<b>125.49***</b>			
Mod. Wgt. 1 (2 wds)	13.01***	2.02	1.38	2.97
Mod.Wgt. 2 (3 wds)	59.52***	7.93	4.68	13.43
Mod.Wgt. 3 (4→20 wds)	90.22***	31.88	15.60	65.13



\*\*\*p<.0001

Our model looked great. We showed it to Michael, our stats consultant from Public Health. He gave us some bad news.

Our model had blown up.

Predictor Variable	Wald Statistic	Odds Ratio (Exp (B))	95% CI Lower	95% CI Upper
Mod.Expression Type	317.93***			
Exp.Type1 (Comm.Indef)	308.7***	1975.8	847.4	4606.8
Exp.Type2 (Comm.Def.)	204.7***	403.77	177.4	918.5
Exp.Type3 (Gerund)	.584	21271.6	.000	2.69E+15
Exp.Type4 (Other)	106.6***	335.9	107.6	1651.0

Why did our model blow up?

<i>Modifier Exp. Type</i>	<b>S-Gen</b>	<b>OF-gen</b>	
<b>Common Indefinite</b>	56	967	1023
<b>Common Definite</b>	251	1001	1252
<b>Gerund</b>		51	51
<b>Other</b>	34	40	74
<b>Kinship</b>	35	13	48
<b>Proper</b>	619	446	1065
<b>Pronoun</b>	3528	10	3538



This set includes examples like

3064(bg2-388). Every bone and muscle in his body showed , but he did not give the appearance of starving.

4904(bj1-676) They were reluctant to appoint sheriffs to protect the property , thus running the risk of creating disturbances

Michael: Ohh, so these things are categorical! You can't include things that are categorical. That's like including men in a study of pregnancy rates.

Me: These are not categorical! You CAN have gerund possessors that are s-genitives! I even found one once! **"Walking's many virtues."** I found it in an airline magazine.

Michael: Well, the regression doesn't see it that way. It doesn't know what to do. Hence the confidence intervals that stretch off to infinity.

Michael: But you have other problems. Look at these pronouns. **3528 to 10!** That's categorical. You can't do that.

<i>Modifier Exp. Type</i>	<b>S-Gen</b>	<b>OF-gen</b>	
<b>Common Indefinite</b>	56	967	1023
<b>Common Definite</b>	251	1001	1252
<b>Gerund</b>		51	51
<b>Other</b>	34	40	74
<b>Kinship</b>	35	13	48
<b>Proper</b>	619	446	1065
<b>Pronoun</b>	<b>3528</b>	<b>10</b>	3538

Me: No, that's **variable**. You see, in linguistics, we value these rare cases. They indicate that there's a real alternation here, but that these extragrammatical factors may be playing a role in the actual usage patterns. Just look at these examples!

91(ba-91). But questions with which committee members taunted bankers appearing as witnesses left little doubt that they will recommend passage of it.

2231(bg-783). this gentleman here ... informs me that Germany is just on the other side of him.

Michael: Well, whatever.

But you have other problems too.  
You have collinearity.

You said you wanted to investigate the  
independence of your three factors? You have a  
giant black hole where they're all confounded:

Michael: Look. You have 3028 pronouns in the model.

And 87% of them are human.

And of all the human referents you have here, 73% of them are expressed as pronouns.

And all those human pronouns are one word long.

That's like a black hole. It's dragging your model into total collinearity.

Michael: And don't forget that over 99% of your pronouns are in the S-genitive construction.

So in Public Health this would be like doing a study of mortality, where the S-genitive is death, and the OF-genitive is survival, and you have 3500 subjects. And your subjects all smoke, have high blood pressure, and never exercise. And at the end of the study 3490 of them are dead and 10 are alive.

So big deal. Those 10 survivors tell you nothing.

Me: OK OK. We'll take out the Pronouns.

## **Moment of reflection:**

- ***Public Health Categorical vs. Linguistics Categorical***  
**Avoid categorical oppositions in order to keep the model from blowing up.**



## Moment of reflection:

- ***Public Health Categorical vs. Linguistics Categorical***  
**Avoid categorical oppositions in order to keep the model from blowing up.**

*But what does this mean for linguistic data, where good examples are strong evidence of the nature of the alternation's underlying grammar, even if they are statistically rare?*

*And notice that unexpected little chunks of near-categorical cases popped up in the midst of this study of what is clearly a case of probabilistic patterns.*

2e. Our redone  
regression model

No pronouns, no gerunds, 3340 tokens in the model

	Predicted		
Observed	S-Gen	OF-Gen	% correct
S-Gen	667	259	
OF-Gen	197	2245	
			87.2%

This model is pretty good. It increases the correct predictions from 72% (baseline) to 87%. The Naglekerke  $R^2$  (another index of predictive power) is .618.

**Chi-square = 1867.789; df = 7; \*p<0.001**

So do we know more about the questions we asked earlier?

- 1. Are the factors independently affecting the choice of alternant?**
- 2. If they are independently affecting speakers' choice of alternant, can we tell which one is most important?**

**Are the factors independently affecting the choice of alternant?**

Animacy and NP Form appear to be independent of Weight. Their correlations are low ( $r = -.16$  and  $-.18$ ) and they have independent large contributions to the model.

Rosenbach (2005) established this experimentally.

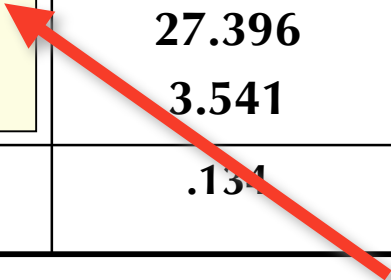
This independence is important because one hypothesis about the role of weight in processing suggests that Animacy and Expression Type effects are at least partially epiphenomena of weight effects.

What about Animacy and NP Form Type? Are they independent?

They are at least partially independent in their effects (( $r=.415$ ) and high independent Wald scores).

**Which is the "most important"?**

<b>Predictor</b>	<b>Wald Statistic</b>	<b>Odds Ratio: Exp (B)</b>	<b>95% CI Lower</b>	<b>95% CI Upper</b>
<b>Mod.Weight</b>	<b>196.6***</b>			
Mod.W.(1)	24.8***	<b>2.138</b>	1.586	2.883
Mod.W.(2)	100.1***	<b>7.327</b>	4.960	10.824
Mod.W.(3)	138.1***	<b>29.162</b>	16.615	51.185
<b>Express.Type</b>	<b>166.7***</b>			
Express.Type(1)	165.9***	<b>9.799</b>	6.924	13.867
Express.Type(2)	7.5**	<b>1.485</b>	1.121	1.967
<b>Mod.Animacy</b>	<b>568.9***</b>			
Mod.Animacy(1)	566.3***	<b>27.396</b>	20.859	35.983
Mod.Animacy(2)	76.4***	<b>3.541</b>	2.667	4.701
<b>Constant</b>	<b>-2.008***</b>	<b>.134</b>		



So we, like Rosenbach, had to remove pronouns from our study.

And like Rosenbach, we found that animacy is the "most important factor".

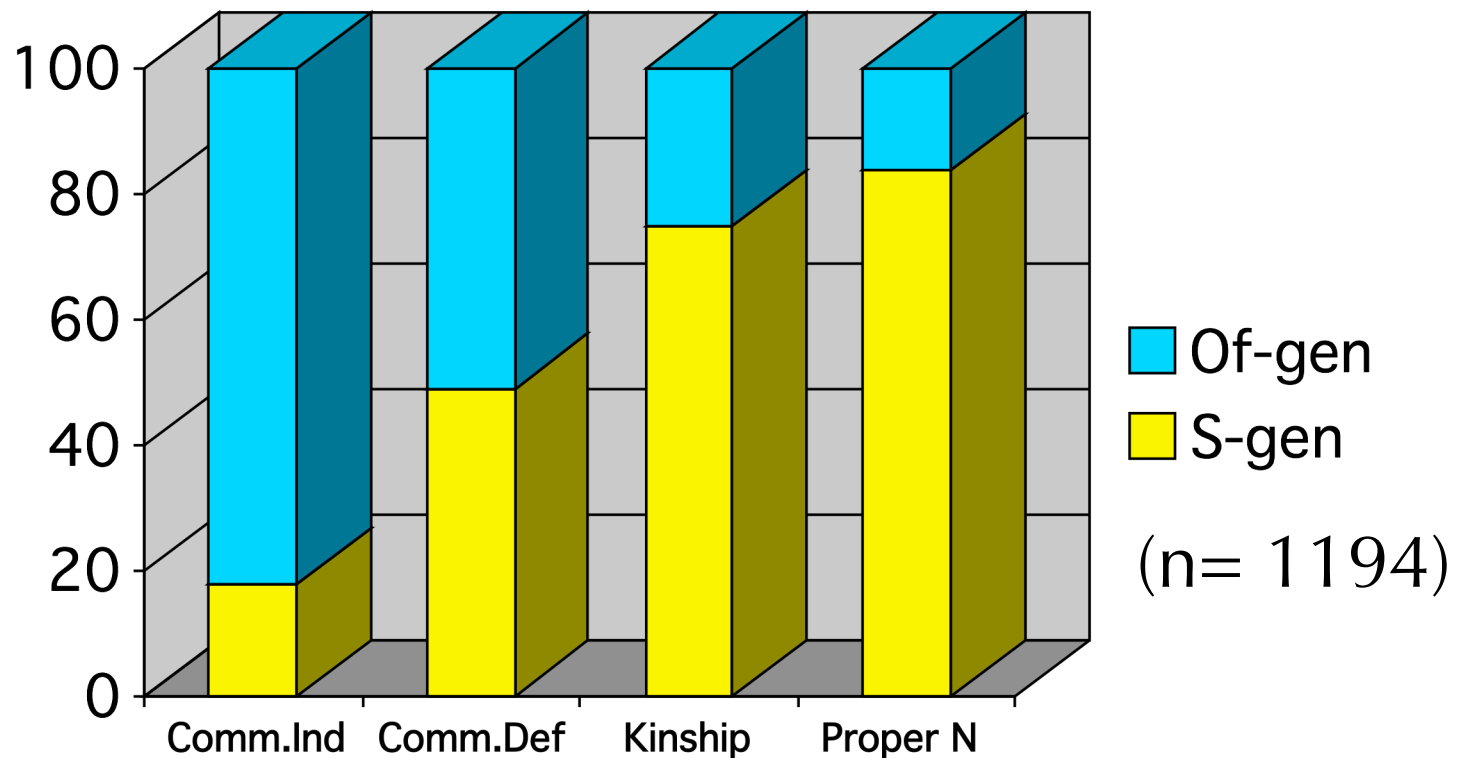


If we go back to the data, can we see any further evidence that our discourse status factor is doing any work?

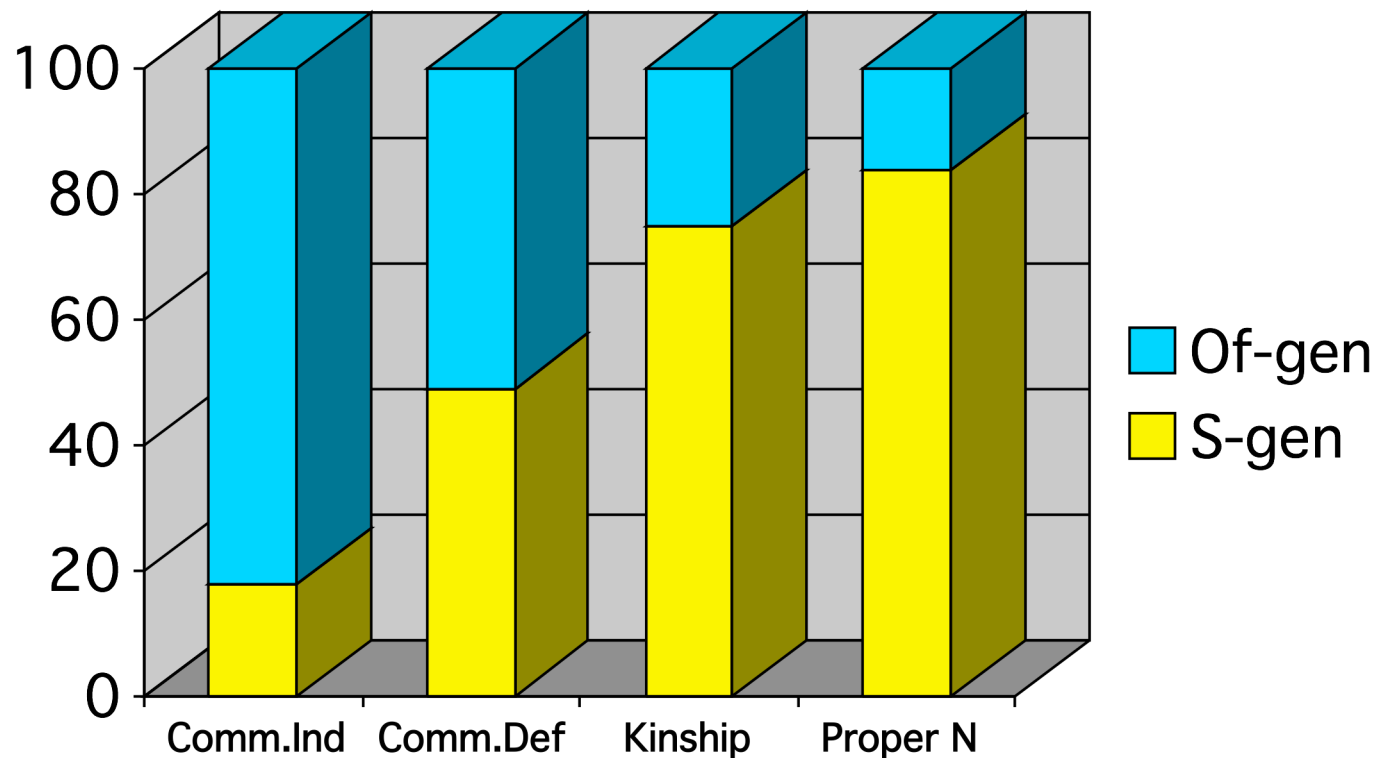
Could the discourse status factor be applicable beyond pronouns?

Let's control for animacy by looking only at Animate possessors:

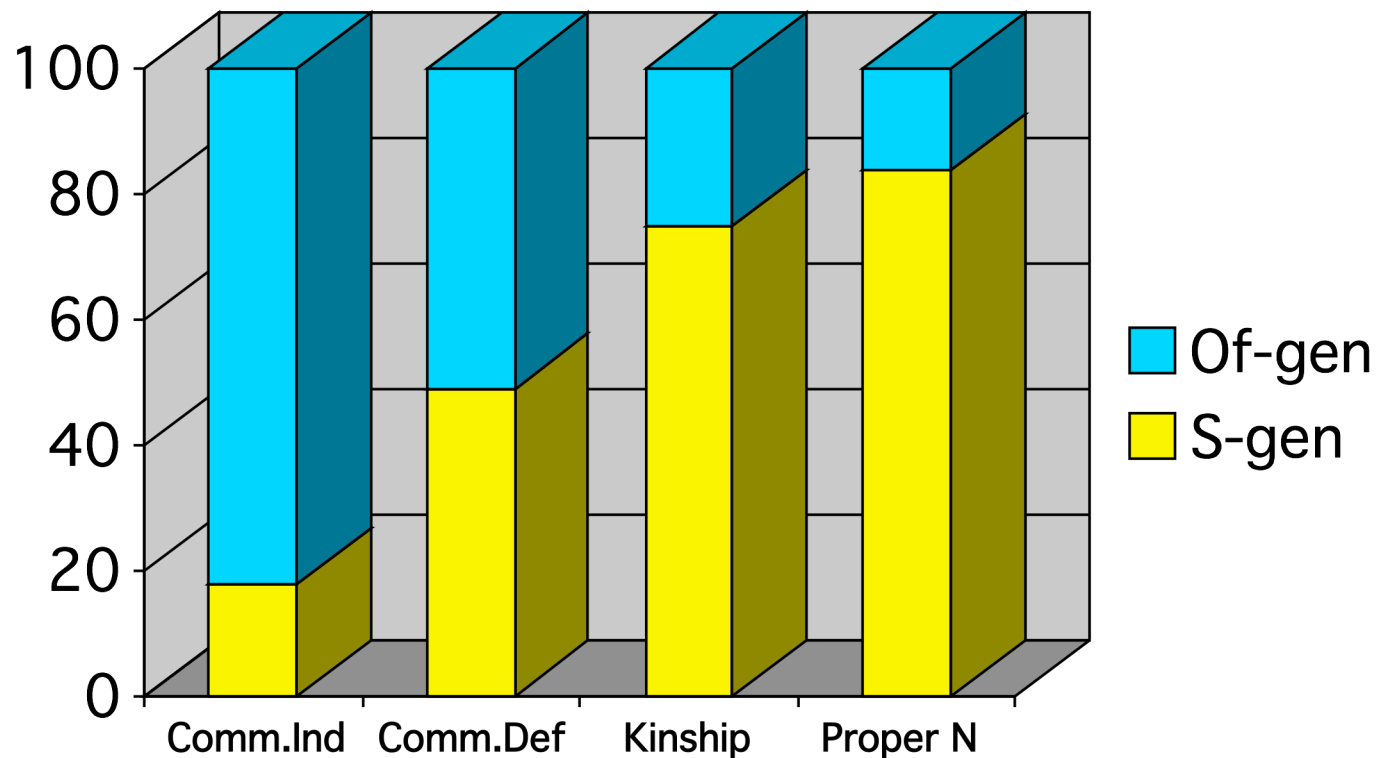
NP form types without pronouns–  
Preferences for *s*-genitive vs. *of*-genitive  
**all Animate possessors**



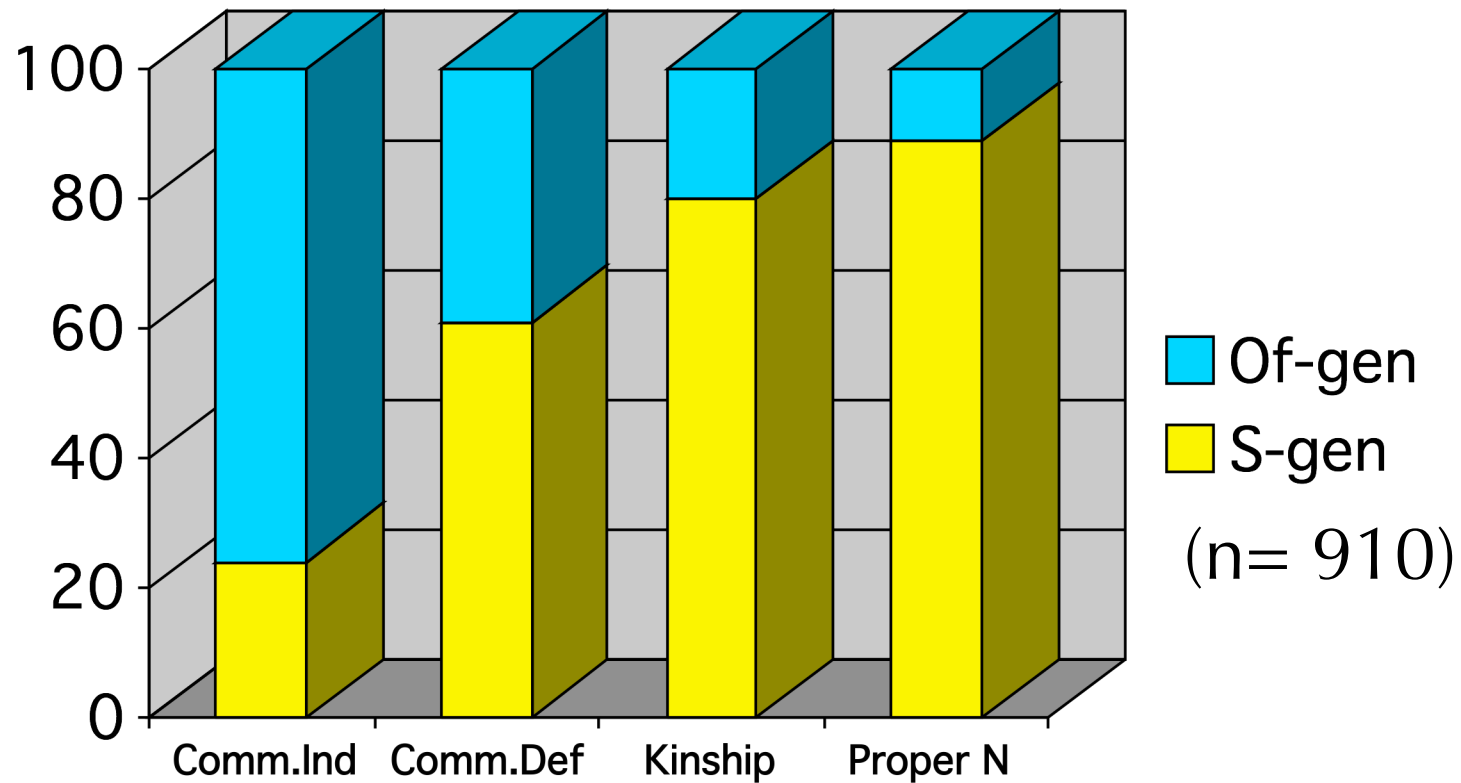
But what about weight? You're including all sorts of phrases here, including possessors that are 3 or more words long.



OK, so let's control for weight. We'll only look at examples with possessors that are 1 or 2 words long.

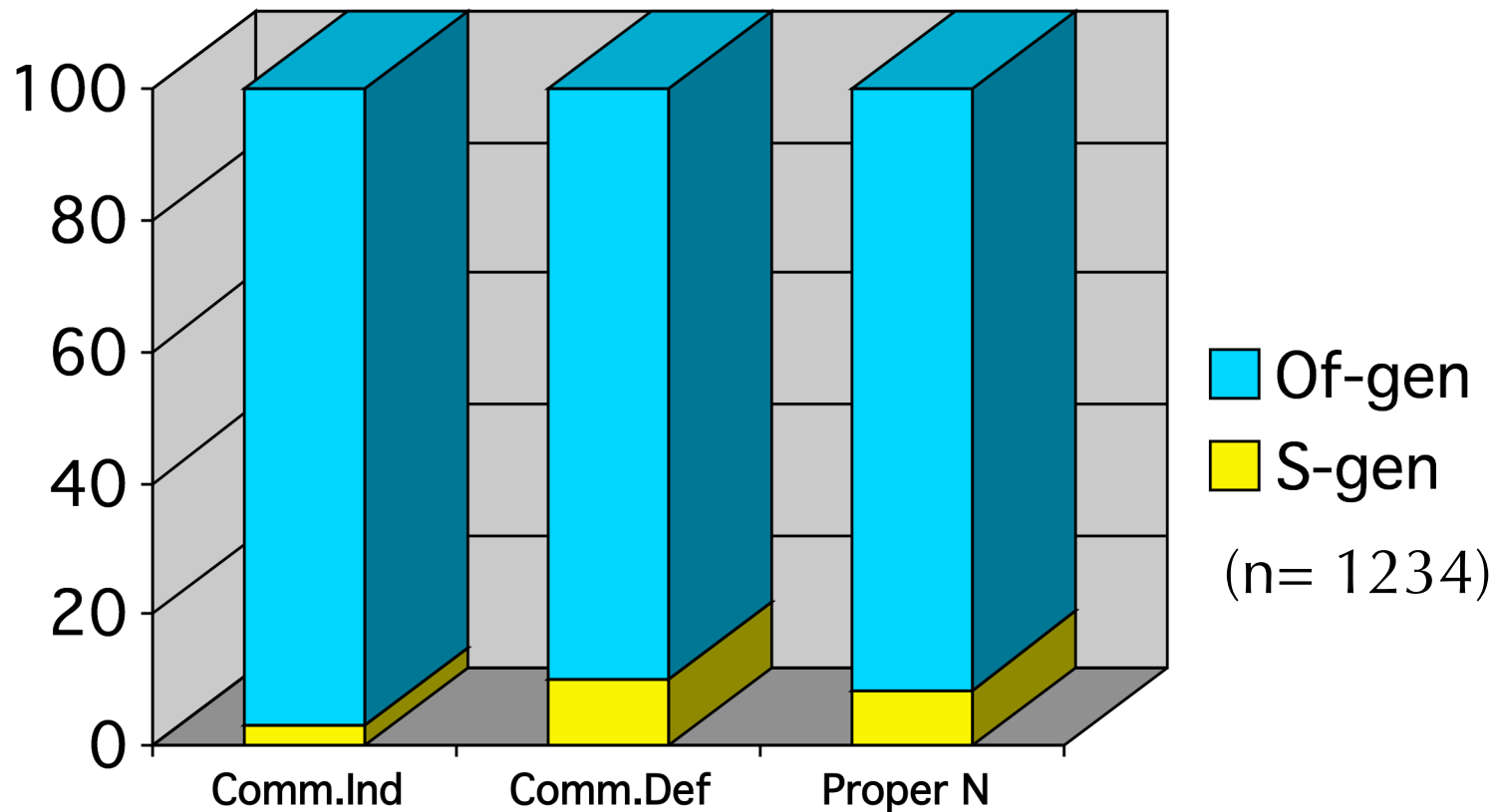


NP form types without pronouns–  
Preferences for *s*-genitive vs. *of*-genitive  
**all Animate possessors**  
**Weight = 1–2 words only**



Does this effect of NP Form type hold if we look at all the examples with Inanimate possessors?  
(Again, we'll control for weight with 1-2 word possessors.)

NP form types without pronouns–  
Preferences for *s*-genitive vs. ***of***-genitive  
**all Inanimate possessors**  
**Weight = 1–2 words only**



## Interim Summary:

Animacy pops out as the strongest factor in Rosenbach's experiments, and in our logistic regression, but that is only after we have removed the overwhelmingly accessible and almost categorical pronouns. And we still see effects of NP type, our proxy for discourse status, even after we control for animacy.



And notice that the discourse status factor is far more active among ANIMATE possessors, even after pronouns are removed.

So overall, our hierarchy of NP form types does seem to consistently display a scalable affinity with the prenominal possessor position.



### 3. Categorical instantiations of probabilistic factors

# The Stochastic Generalization

Statistically noticeable but noncategorical patterns found in one language are often found in other languages in categorical and relatively inviolable form.

Bresnan, Dingare & Manning 2001  
Manning 2002

# The 'Mono-Lexemic' Possessor Construction (Skarabela, O'Connor, Maling 2004)

A survey of 30 Romance, Slavic and Germanic languages revealed that 19 of them had a highly constrained possessive construction that alternated with the standard adnominal genitive after the head noun.

By and large, this prenominal possessor

- is not phrasal
- cannot be modified
- cannot appear with a determiner
- consists of no more than one word
- is limited in some languages to certain NP form types

# The 'Mono-Lexemic' Possessor Construction (Skarabela, O'Connor, Maling 2004)

The construction seems to grammaticize the tendencies associated with **animacy** and **discourse status** we saw in the English data,

in the vehicle of very restricted **weight** (one unit).

# Icelandic

**Pronoun:**

*mitt hús*

**“my house”**

*þeirra bíll*

**“their car”**

**Proper  
noun:**

*Siggu hús*

**“Sigga's house”**

**Kinship  
term:**

*mömmu bíll*

**“Mommy's car”**

**Common  
Noun:**

*\*systurinnar hús*

**“sister.the's house”**

*\* yfirmannsins hús*

**“boss.the's house”**

# German

**Pronoun:**

*mein* haus

"my house"

**Proper  
noun:**

*Franks* haus

"Frank's house"

**Kinship  
term:**

*Mutters* haus

"Mama's house"

*Omas* haus

"Grandma's house"

\**Bruders* haus

"brother's house"

**Common  
Noun:**

\**(des)Bosses* haus "(the) boss's house"

# Czech

**Pronoun:**

*moje kniha*

“my book”

**Proper  
noun:**

*Milanova kniha*

“Milan’s book”

**Kinship  
term:**

*bratrova kniha*

“brother’s book”

**Common  
Noun:**

*kamarádova kniha*

“friend’s book”

*učitelova kniha*

“teacher's book”

*klukova kniha*

“boy’s book”



# Bosnian/Croatian/Serbian

**Pronoun:**

njegova kuća

"his house"

**Proper  
noun:**

Milanova kuća

"Milan's house"

**Kinship  
term:**

mamina kuća

"Mommy's house"

bratovljeva kuća

"brother's house"

**Common  
Noun:**

prijateljeva kuća

"friend's house"

zubareva kuća

"dentist's house"

delfinova igrack

"dolphin's toy"

# Russian

**Pronoun:**

ego kniga

"my book"

**Proper  
noun:**

Mashina kniga

"Masha's book"

**Kinship  
term:**

bratova kniga

"brother's book"

djadina kniga

"uncle's book"

\*kuzenina kniga

"fem.cousin's book"

**Common  
Noun:**

\*drugova kniga

"friend's book"

Pronoun	Proper N	Kinship	Common animate	Common Inanimate
Bosnian/Croatian/Serbian				*
Czech				*
Russian		*	*	*
Icelandic	*	*	*	
German	*	*	*	

So Icelandic, German, and Russian seem to limit the MLP to pronouns and proper names, and a few kinship terms that act like proper names.

But this issue of kinship terms acting as proper nouns is actually somewhat complicated:

## A note on the indexical complexity of kinship terms as referring expressions...

Joan to her cousin:

*"Hey Cousin! It's great to see you."*

Joan to her friend:

*##"That's Cousin's car."*

- The Speaker's use of a Kinship term as a vocative is not sufficient for that term to become a Proper Noun-like referring expression.

A note on the indexical complexity of kinship terms as referring expressions...

Joan's mother talking about her own sister:  
*"Sis is making pancakes."*

Joan's mother to her own sister:  
*"Sis, will you hand me that ladle?"*

Joan to her mother re: her mother's sister:  
*## Sis's car is outside."*

A note on the indexical complexity of kinship terms as referring expressions...

Father to two young sons:

*"Mommy is making pancakes for you."*

Same father to wife:

*"Mommy, will you meet me for lunch?"*

Wife: #\$\$%&@

George H.W. Bush, (re: wife Barbara), to press

*"Mommy and I are going to Camp David."*

Pronoun	Proper N	Kinship	Common animate	Common Inanimate
Bosnian/Croatian/Serbian			*	
Czech			*	
Russian	*		*	*
Icelandic	*		*	*
German	*		*	*
Romance	*	*	*	*



# Spanish

**Pronoun:**

*su casa*

“his/her house”

**Proper  
noun:**

\* *Sylvia casa*

“Sylvia's house”

**Kinship  
term:**

\* *mami casa*

“Mommy's house”

\* *hermano casa*

“brother's house”

**Common  
Noun:**

\* *maestro casa*

“teacher's house”

Possible  
responses to this  
proposal:

Pronoun	Proper N	Kinship	Common animate	Common Inanimate
Bosnian/Croatian/Serbian				*
Czech				*
Russian		*	*	*
Icelandic		*	*	*
German		*	*	*
Romance	*	*	*	*

"But these aren't the same thing! The Possessive Adjective in Slavic is not related to these Germanic and Romance possessors! And the PA is an adjective, not a nominal!"

One response:

Corbett (1987) showed that the Slavonic PA has some characteristics of a nominal possessor, including the ability to act as an antecedent to pronouns (and even relative pronouns in some of these languages):

Example:

Upper Sorbian (Corbett 1987:303, ex. (22-23))

Słysetaj...Wićazowy hłós, kotryž je zastupił  
(they)hear Wicaz's voice, who is gone.in

To je našeho wučerjowa zahrodka. Wón wjele w njej džěła  
that is our teacher's garden. he a.lot in it works

(Does the Mono-Lexemic Possessor "construction" in these Indo-European languages stem from a single genetic source?

Probably not, but that is not our focus.)

All of these instances appear to make use of at least two of the three factors we explored in our corpus study of English:

The English preference for **light prenominal possessors** here is categorical: WEIGHT = 1.

Czech provides evidence of the monolexemic nature of this construction:

(a) kniha *Milan-a*      *Kunder-y*  
book Milan-gen.      Kundera-gen.  
*“(a/the) book of Milan Kundera”*

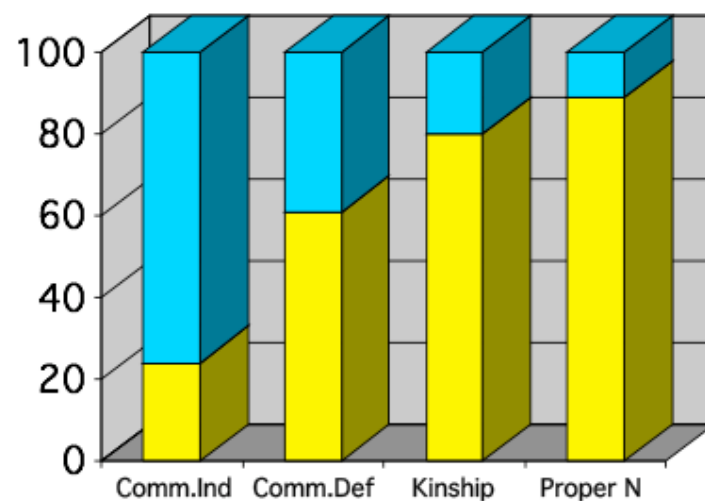
(b) *Milan-ova*      or      *Kunder-ova*      kniha  
Milan-poss.adj. book or      Kundera-poss.adj.      book  
*“Milan’s book”*      or      *“Kundera’s book”*

(c) \* *Milan-ova*      *Kunder-ova*      kniha  
Milan-poss.adj. Kundera-poss.adj. book  
*“Milan Kundera’s book”*

NP Form Types: The English probabilistic pattern,  
that different NP form types have scalar affinities  
with the prenominal possessor

here becomes categorical:

Pronoun	Proper N	Kinship	Common animate	Common Inanimate
Bosnian/Croatian/Serbian				*
Czech				*
Russian	*	*	*	*
Icelandic	*	*	*	*
German	*	*	*	*
Romance	*	*	*	*



What about animacy, the factor that emerged as "most important" in our regression?

Pronoun	Proper N	Kinship	Common animate	Common Inanimate
Bosnian/Croatian/Serbian				*
Czech				*
Russian		*	*	*
Icelandic		*	*	*
German		*	*	*
Romance	*	*	*	*

Here animacy seems subordinated to NP form type: While no language allows **inanimate common nouns**, all seem to allow **inanimate pronouns** and some allow **inanimate proper nouns**.

# Inanimates allowed:

## Spanish

su color

"its color"

## German

seine Farbe

"its color"

Berlins Straßen

"Berlin's streets"

## Icelandic

hennar litur

"its color"

Reykjavíkur götur

"Reykjavik's streets"

## Czech

jeho barva

"its color"

\* **Berlinova ulice**

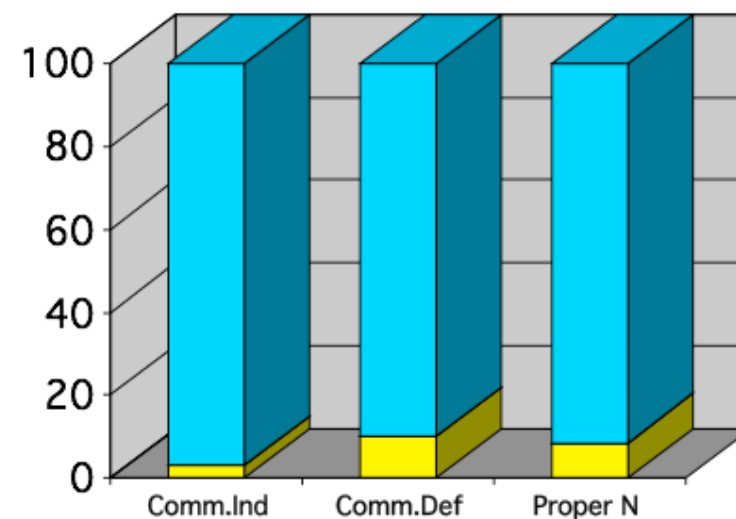
"Berlin's streets"

(sounds like a person)



But no **common noun Inanimates** are allowed...

Pronoun	Proper N	Kinship	Common animate	Common Inanimate
Bosnian/Croatian/Serbian				*
Czech				*
Russian	*		*	*
Icelandic	*		*	*
German	*		*	*
Romance	*	*	*	*



paralleling our  
English probabilistic  
patterns.

While the historical sources are important and interesting, here we want to emphasize the tension between the categorical and the probabilistic, the general and the particular.

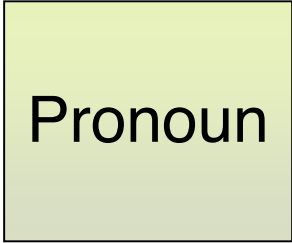
The Mono-Lexemic Possessor configuration suggests that there is something highly adaptive about the prenominal, single-word, discourse old, animate possessor.

In each case we've looked at, the MLP construction alternates with an unmarked possession construction, a post-nominal genitive that has no constraints on length, definiteness, or animacy.

So the MLP construction can potentially join the ranks of other non-canonical constructions that display 'conventional pragmatics' or specified value ranges for information status. Many unrelated languages have preposing and postposting constructions with fixed information status values for the pre- or post-posed element. (See Birner & Ward 2005 i.a.)

The next step would be to evaluate carefully just exactly what the information status requirements are for interpreting these one-word possessor expressions.

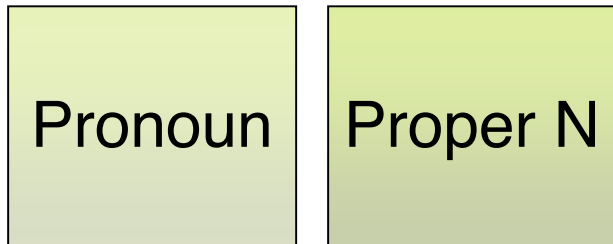
Pronouns, clearly, are discourse-old in Prince's terms, and are used to index activated discourse entities in Gundel's terms. In Ariel's terms they are high accessibility markers.



Pronoun

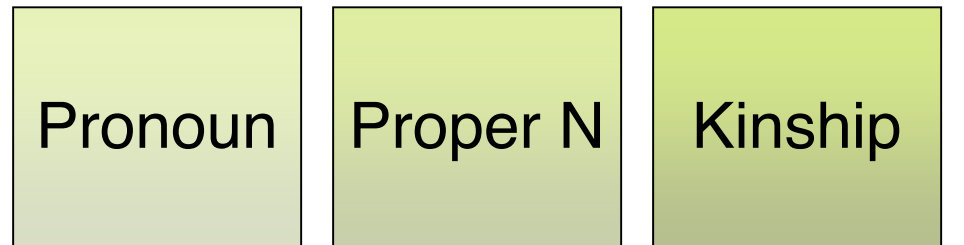
So if the MLP construction requires a highly accessible possessor, pronominal possessors will by their nature be welcomed in that construction.

Proper nouns are hearer-old in Prince's terms. They generally are used when the speaker believes the hearer is familiar with the referent.



So some MLP constructions may allow possessors expressed as proper nouns, since the referent of a proper noun meets the condition of accessibility.

Kinship terms also have a feature that anchors their information status in a conversation.



They are relational. A term like *brother* requires a "possessor" as an argument. So when speakers hear an MLP like ***brother's car***, they have to figure out "Whose brother?"

And the MLP, with its one-word slot, does not provide any place to index the possessor explicitly. So there is interpretive work to be done. What are the possibilities?

# Interpretive evidence: Czech

In isolation:

That is **bratrovo auto**. "That is brother's car"

Most available reading? "*my (Speaker's) brother's car*"

# Interpretive evidence: Czech

With some context:

Kamarádka má několik bratrů v Praze.

Nedávno si půjčila **bratrovo auto**.

"My friend has several brothers in Prague.

Recently, she borrowed ***brother's car***."

Most available reading?    *"my (speaker's) brother's car"*



With even more context:

*Moje spolubydlící Jana není z Prahy, ale má tam bratra a bratrance, a tak tam často jezdí. Navíc nedávno si tam její bratr koupil velký byt, kam si může přivést kamarády. Zajeli jsme do Prahy párkrát spolu, ale já vždycky přespáváme u **bratrancovy přítelkyně**, protože bydlí v centru.*

"My roommate Jana is not from Prague. But she has a brother and a cousin there. She often goes to visit them. Moreover, recently her brother has bought a big apartment there and she can bring along her friends. The two of us went there together a couple of times, but we always stay at **cousin's girlfriend**, because she lives downtown."

With even more context:

"My roommate Jana is not from Prague. But she has a brother and a cousin there. She often goes to visit them. Moreover, recently her brother has bought a big apartment there and she can bring along her friends. The two of us went there together a couple of times, but we always stay at **cousin's girlfriend**, because she lives downtown."

What did Czech speakers say?

4 of 10 L1 Czech speakers: it is the speaker's cousin,

2 of 10 L1 Czech speakers: it is probably Jana's cousin,

4 of 10 L1 Czech speakers: it is Jana's or the speaker's cousin.

Interpretive evidence: Russian

In context:

"My friend Abi had no way to get to the airport, so he borrowed *mamaina mashina*"  
"mom's car"

Available interpretations:

the speaker's mother

Abi's mother

No other interpretation

## Czech common nouns: quasi-relational

Pronoun	Proper N	Kinship	Common Animate
---------	----------	---------	-------------------

Some common nouns are quasi-relational, like "teacher" (teacher of whom?), or even "dentist". Speakers seem to treat these in the same way as kinship terms, by trying to anchor them via the speaker or via an accessible discourse entity.

# Czech common nouns: quasi-relational

In isolation:

*To je zubařovo auto.*

this is dentist-PA car

“That is \_\_ dentist’s car.”

*To je kadeřníkovo auto.*

this is hairdresser-PA car

“That is \_\_ hairdresser’s car.”

=> speaker's dentist/ hairdresser

## Czech common nouns: non-relational

*Koupila jsem neteři kolo,*

*ale ona se stejně vozí na **klukově koloběžce***

“I bought a bike for my niece,

but she rides **boy’s scooter** anyway.”

Native speakers were puzzled: "But who is the boy??"

(Cannot be construed as a compound, "boy's scooter")

# What about other Slavic languages?

Some of these languages allow common nouns in the MLP construction. Do they also require an explicit discourse-old antecedent to anchor them?

## Interpretive evidence: Bosnian/Croatian/Serbian

We took the kids to Marine World. They have a big central pool, with animal trainers doing several different acts, all going on at the same time: a seal doing tricks for food, and a big turtle floating around the pool, and other things.

The kids loved it, but after a while it got kind of crazy. The seal got bored and started hitting the water with its flipper, splashing people. The turtle dove out of sight, and then **delfinova igracka** (dolphin's toy) hit someone in the audience. So we left.

Native speakers find this fine. In B/C/S, an **inferrable discourse entity** can anchor a possessive adjective.



# Interpretive evidence: cf. Czech

We took the kids to Marine World. They have a big central pool, with animal trainers doing several different acts, all going on at the same time: a seal doing tricks for food, and a big turtle floating around the pool, and other things.

The kids loved it, but after a while it got kind of crazy. The seal got bored and started hitting the water with its flipper, splashing people. The turtle dove out of sight, and then **delfinova igracka** (dolphin's toy) hit someone in the audience. So we left.

In Czech, on the other hand, an **inferrable discourse entity cannot** anchor a possessive adjective.

# Summary

- The 3 big factors driving speaker choice of genitive alternant in English can be seen operating across languages, some only remotely related. What is probabilistic in English is categorical in some other languages, as 'the stochastic generalization' predicts.
- But pursuing this idea brings us face to face with some messy puzzles: our notions of "categorical" and "probabilistic," straightforward-sounding in Manning's generalization, turn out to be riddled with particularities and contingencies.

- Each version of the "mono-lexemic possessor" construction has specific idiosyncrasies. In what sense are these diverse configurations instances of the same "construction"?

This problem space is like a biological symbiosis...



The leaf-cutter ants live on fungi. They cultivate patches of it, growing it on pieces of leaves they cut.



They make sure their fungus garden doesn't get infected with a persistent mold that could kill it. They can do this because of a bacterium they have on their own bodies.



This symbiosis is more than the sum of the parts. It requires us to think about at least five independent organisms-- ants, fungi, molds, leaves, and bacteria. But the object of our consideration is a *system* in which these five interact. Which is most important?



In some sense the system is the level of organization we want to understand. Are the same principles working in other symbioses? And what about the details of each environment?





We need a way to think about the larger forces operating across languages while at the same time taking into full account the minute specifics of each language.



For example, in the Slavic languages the syntactic structure of the noun phrase does not require an overt determiner. Does this open the door for use of common nouns in the single-lexeme MLP construction?





In German, like English, the syntax of NPs requires an overt article if the head is a singular common noun.

In the pre-nominal slot, a common noun possessor cannot be monolexemic because the syntax requires an overt article.

In German, *des Bosses Haus* (*the boss's house*) is no longer used. Is it because of a persistent pressure from the one-lexeme requirement?

By this logic, in Icelandic, the suffixal definite article should allow the language to get around the problem, and extend the MLP construction to common nouns, but it still does not allow common noun possessors.





Like students of a symbiotic system, we are forced to continually attend both to the larger functional pressures and the local ecological details that either facilitate or resist those pressures.

Thank you!

# Acknowledgements

This research was supported by NSF grant BCS-008037, "Optimal Typology of Determiner Phrases". The support of the NSF Linguistics Program is gratefully acknowledged. No endorsement of this research is implied.

Many thanks to faculty colleagues Arto Anttila, Vivienne Fong; graduate research assistants Gregory Garretson, Marj Hogan, Mary Hughes, Barbora Skarabela and undergraduate research assistants: Amy Rose Deal and John Manna. Thanks also to Joan Bresnan, Annie Zaenen, and Tom Wasow for discussions of animacy, and to Boston University students in LS 751, Spring 2002, for discussions of some of this material.

Many thanks also to Michael Winters and Tim Heeren for statistical consulting. Thanks for data judgments to Vera Dumancic, Masha Polinsky, Marion Rheiner (and her mother).