Is it the size, or how you use it? Comparing the effects of subject length and predictability on contraction

Constantine Lignos
University of Pennsylvania, Children's Hospital of Philadelphia
Laurel MacKenzie
University of Manchester

NWAV42 ◆ 10/19/13

Overview

- Multiple factors condition auxiliary contraction (e.g., /ɪz/ ~ /z/), namely:
  - Subject length (size)
  - Predictability of the auxiliary (how you use it)
- What is their relative contribution?
- What causes these effects?

Analyzing auxiliary contraction

Auxiliary contraction

is

Yeah, **Salzburg's** nice. **Austria's** nice. **Europe is** nice! (sw_1151)

has

Oh, I'm sure **it's** been done. I'm sure **it has** been done. (sw_1060)

will

If I walk, **it'll** be ten degrees warmer, but **it will** last twenty minutes. (sw_1146)

The dependent variable

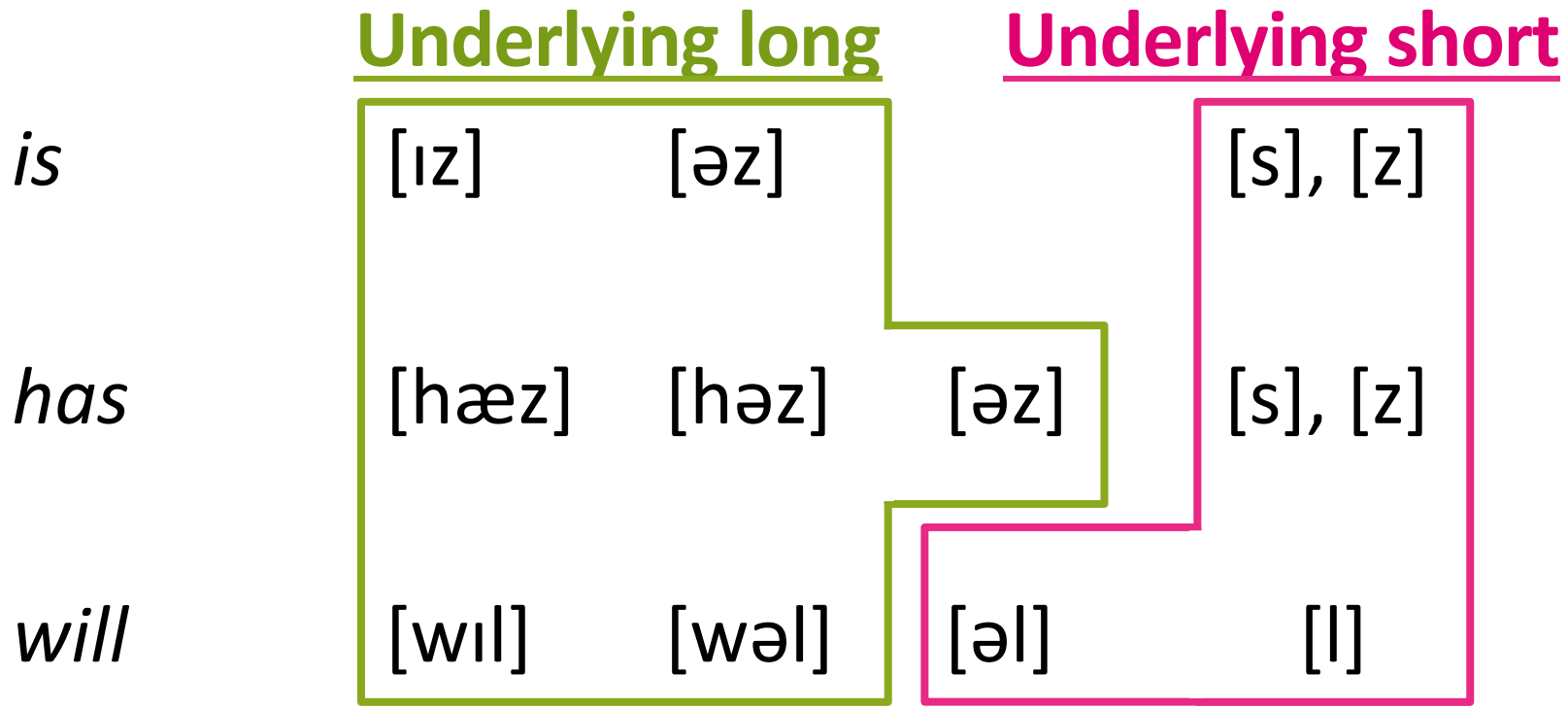
Underlyingly: a bipartite allomorphic alternation

“long”	/ɪz/	~	/z/	“short”
	/hæz/	~	/z/	
	/wɪl/	~	/l/	

(MacKenzie 2013)

The dependent variable

Surface forms can be reclassified:



(MacKenzie 2013)

Conditions on contraction

Subject length

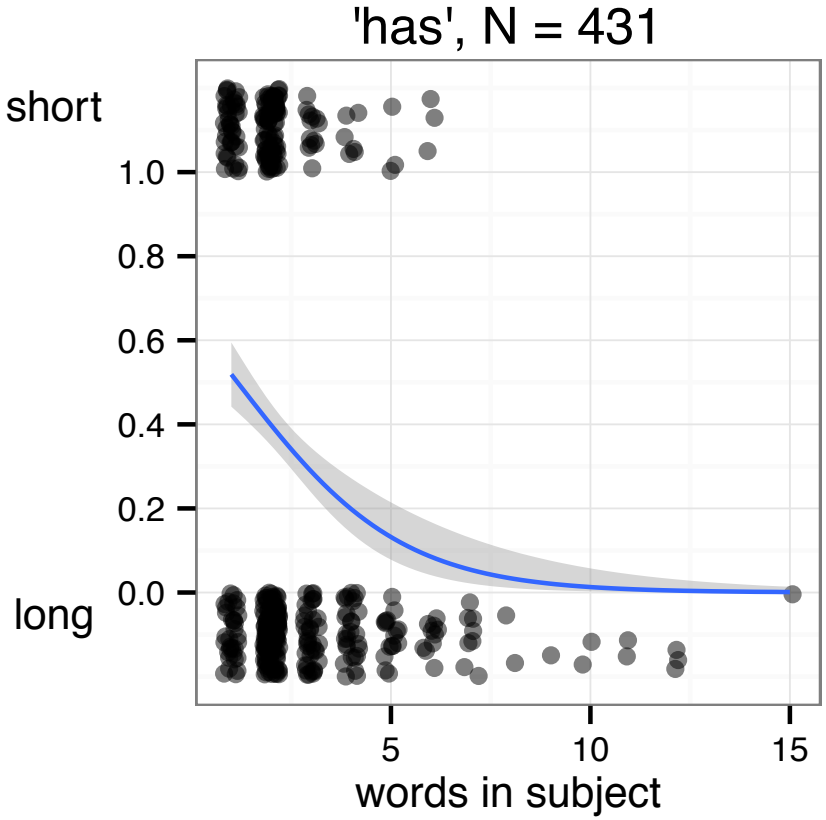
Orthographic word count:

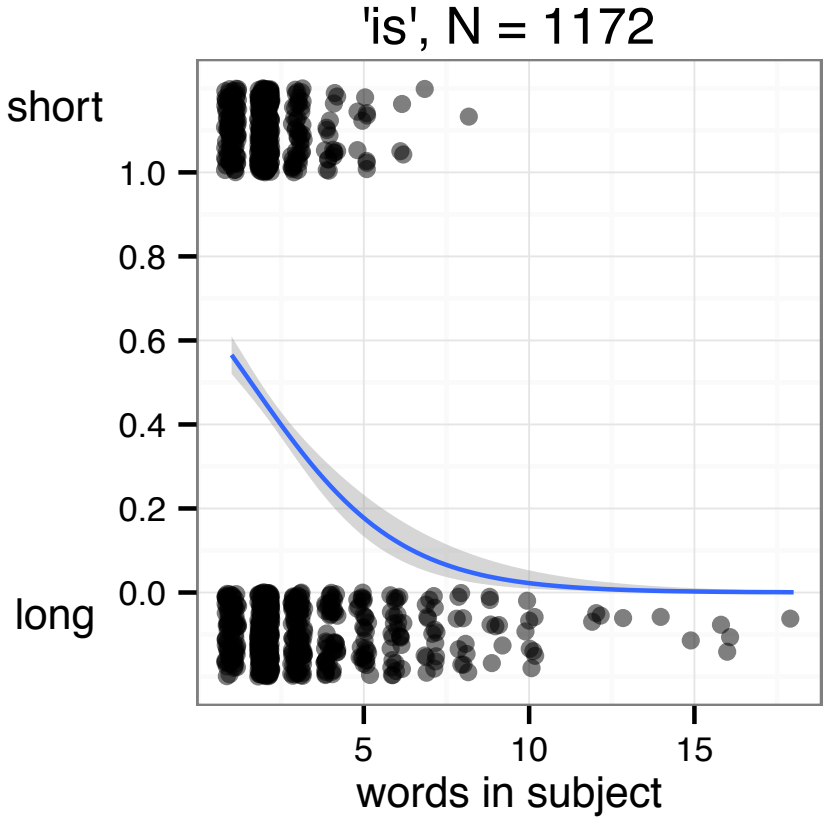
Salzburg's nice 1

The real estate out here's been pretty good 5

About the only thing I can do mechanically with a, a car is put gas in it 12

Subject length effect



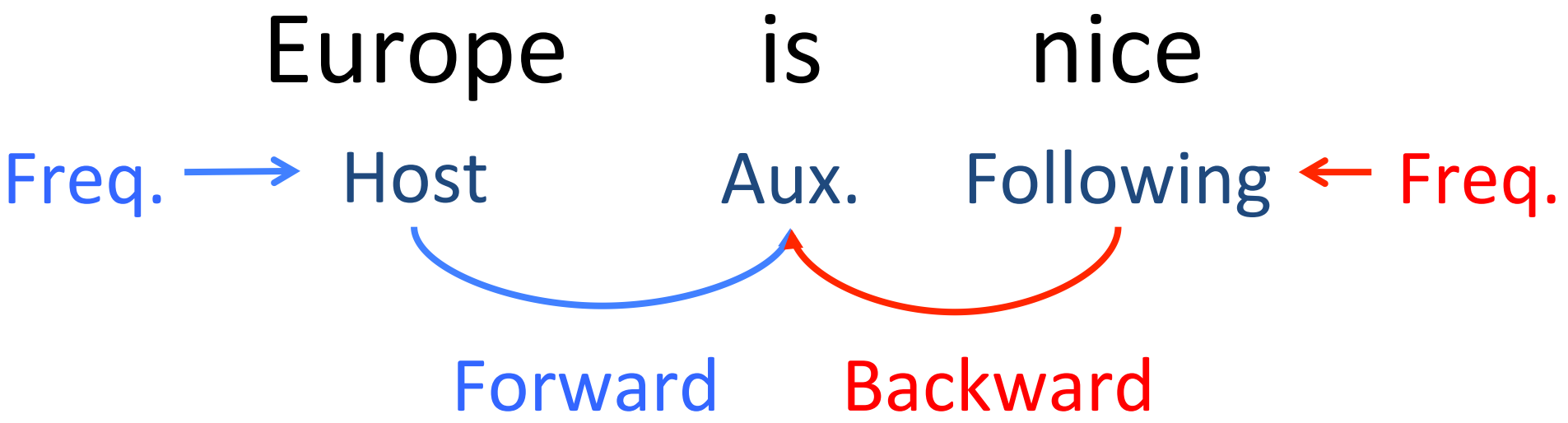


(MacKenzie, 2013)

Predictability

The more “predictable” a context is, the more likely contraction will occur there

(Frank & Jaeger, 2008)



Impact of predictability

"[Under our hypothesis] speakers should prefer a full form whenever the content conveyed by the form is unexpected in its context."

- They find:
  - Lower host freq. → less contraction
  - Lower forward prob. → less contraction

(Frank & Jaeger, 2008)

Why does predictability matter?

We don't know, but maybe:

- We are highly rational in our communications, contracting in more predictable contexts to aid the hearer
- The architecture of our speech planning system is affected by predictability

Methodological challenges

Contractions can't be coded using Switchboard transcripts alone:

> Contractions are allowed, but be conservative. […] It is always permitted to spell out forms in full, even if the pronunciation suggests the contracted form.
>
> (Instruction manual for Switchboard transcribers)

Methodological challenges

- No explicit analysis of contraction, modeled mix of:
  - Phonological variation (/h/-deletion)
  - Morphological variation (short/long allomorphy)

- Poor estimates of predictability
  - Relatively small data set used for estimates

Open questions

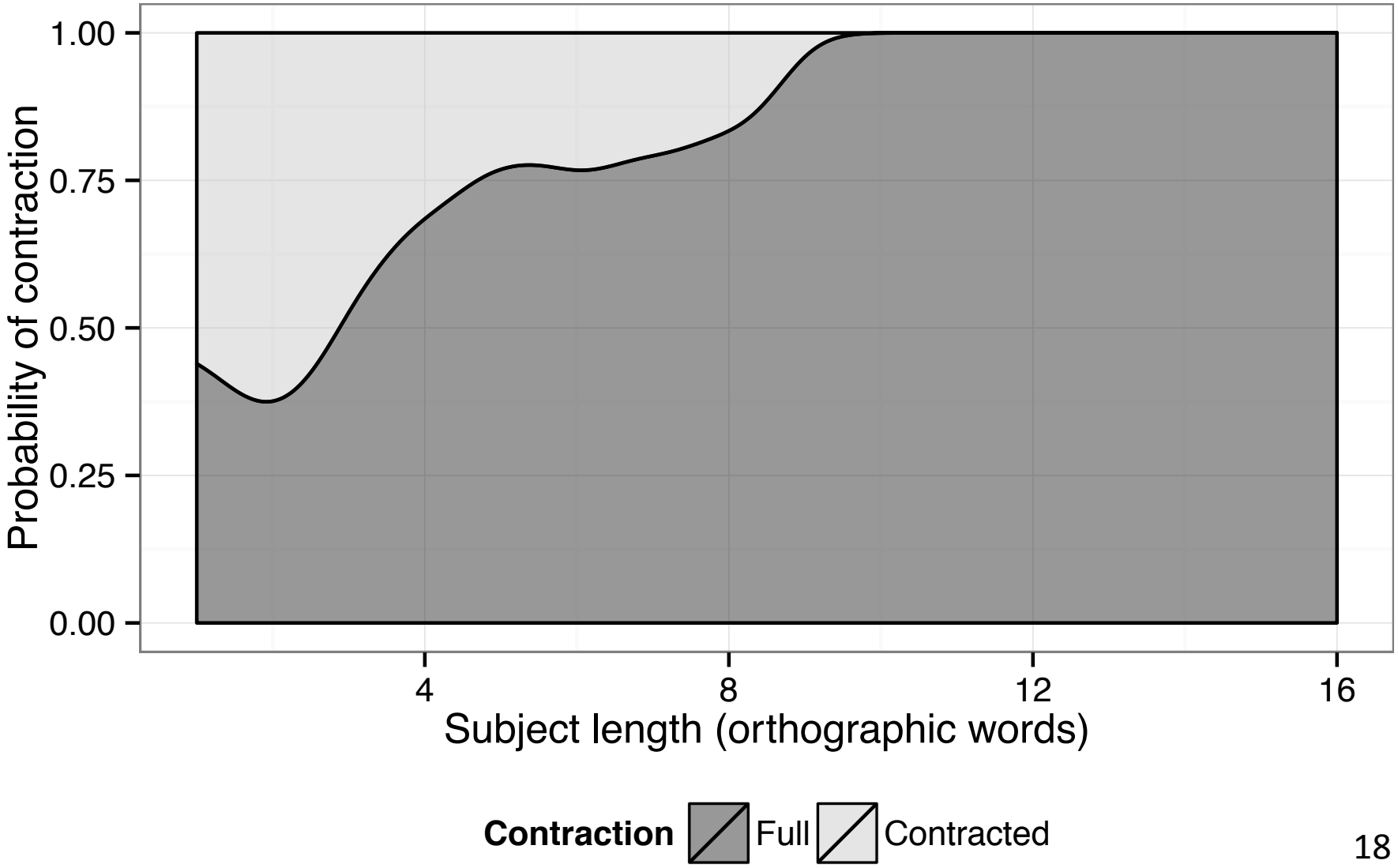
- What is the relationship between subject length and predictability?
  - Are they possibly measures of the same thing?
  - Are they equally important?
- What measure(s) best express subject length?
  - e.g., syllables, ortho. words, phon. words…

Size, or how you use it?

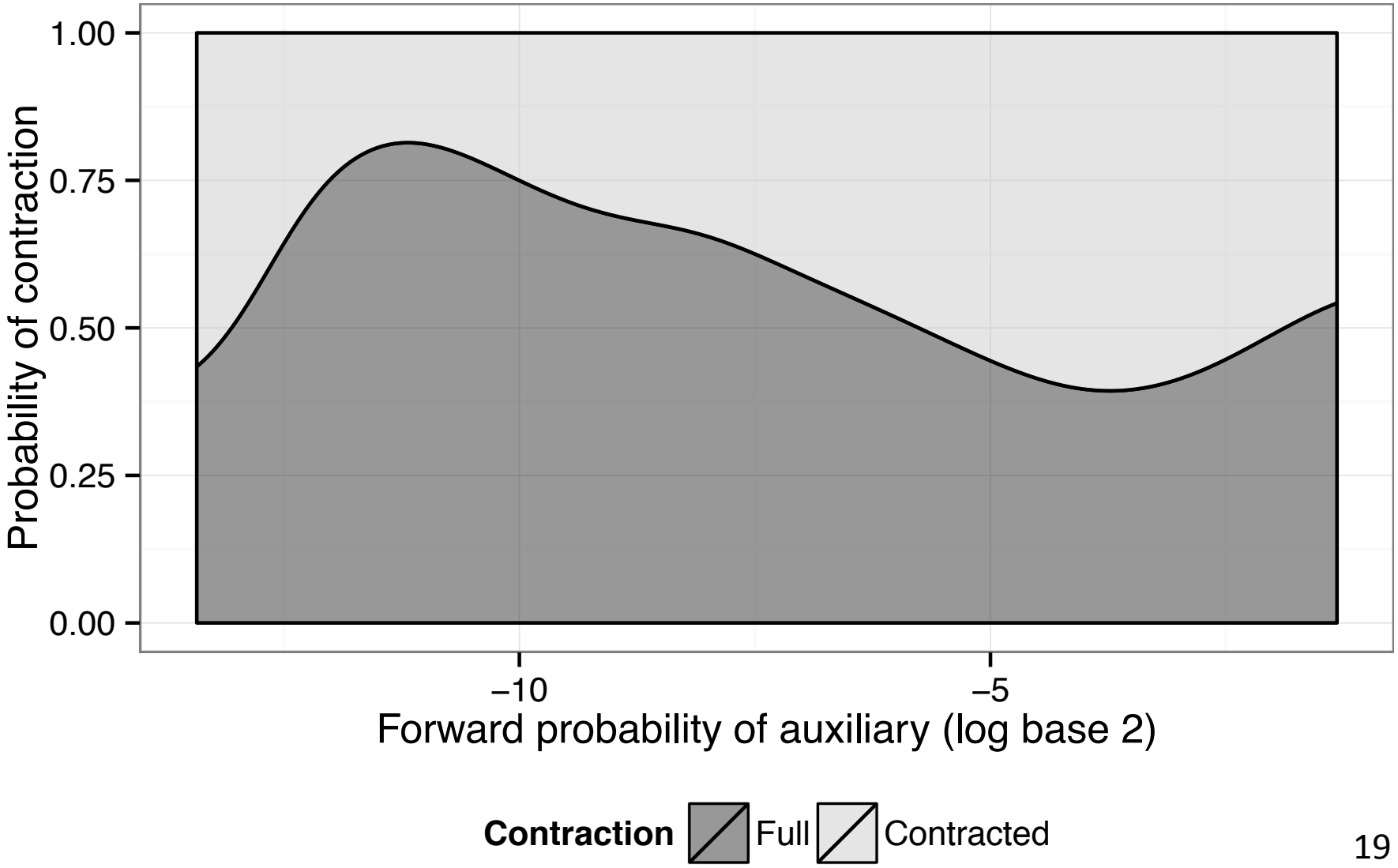
Data sources

- The Switchboard corpus (Godfrey et al., 1992)
- The Fisher corpus (Cieri et al., 2004)
  - 5-minute telephone conversations between strangers on a given topic
- The Philadelphia Neighborhood Corpus (Labov & Rosenfelder, 2011)
  - Sociolinguistic interviews carried out by Penn Linguistics students
- 1,092 tokens in contractible contexts with valid predictability information

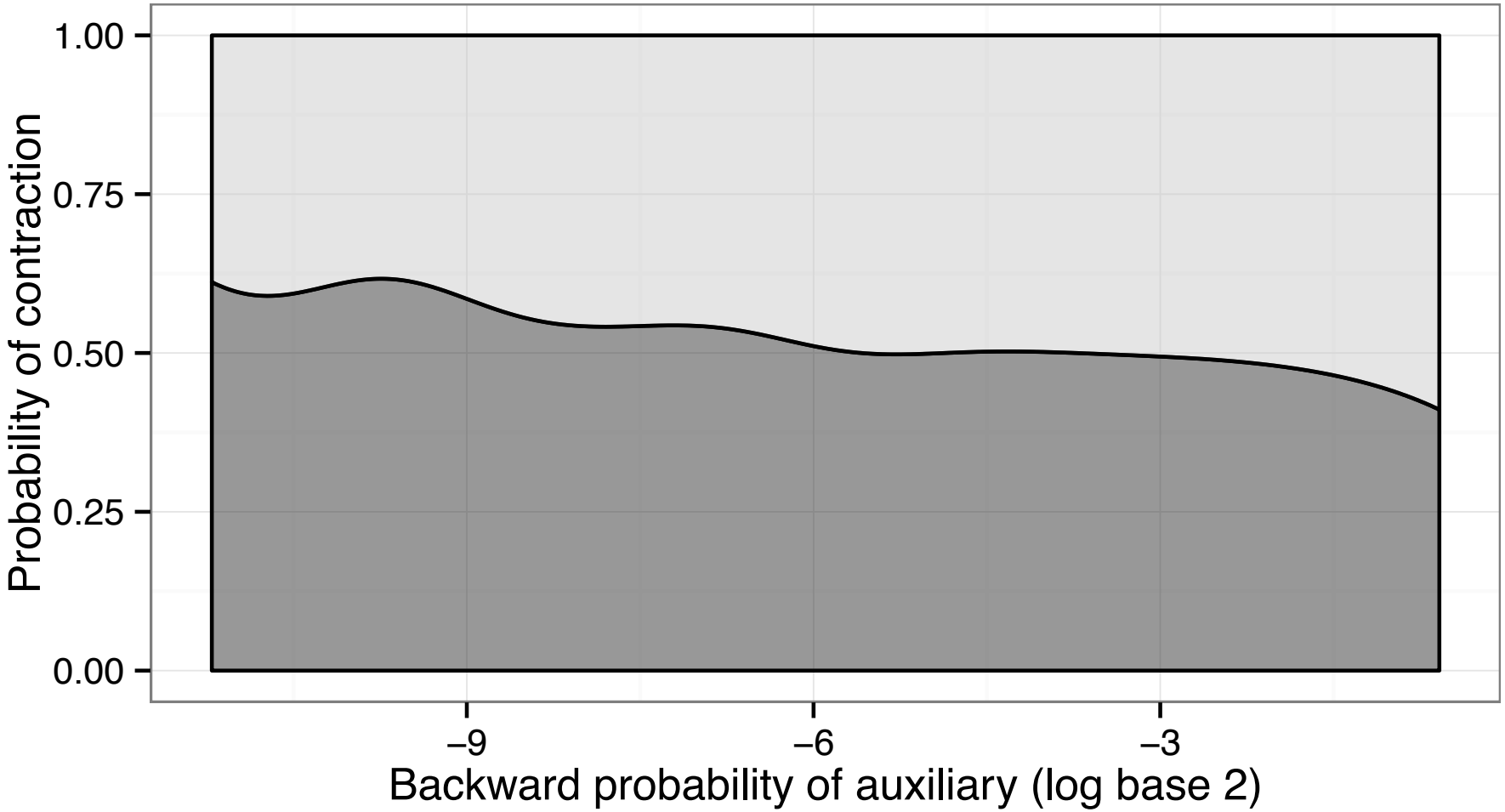
Number of words



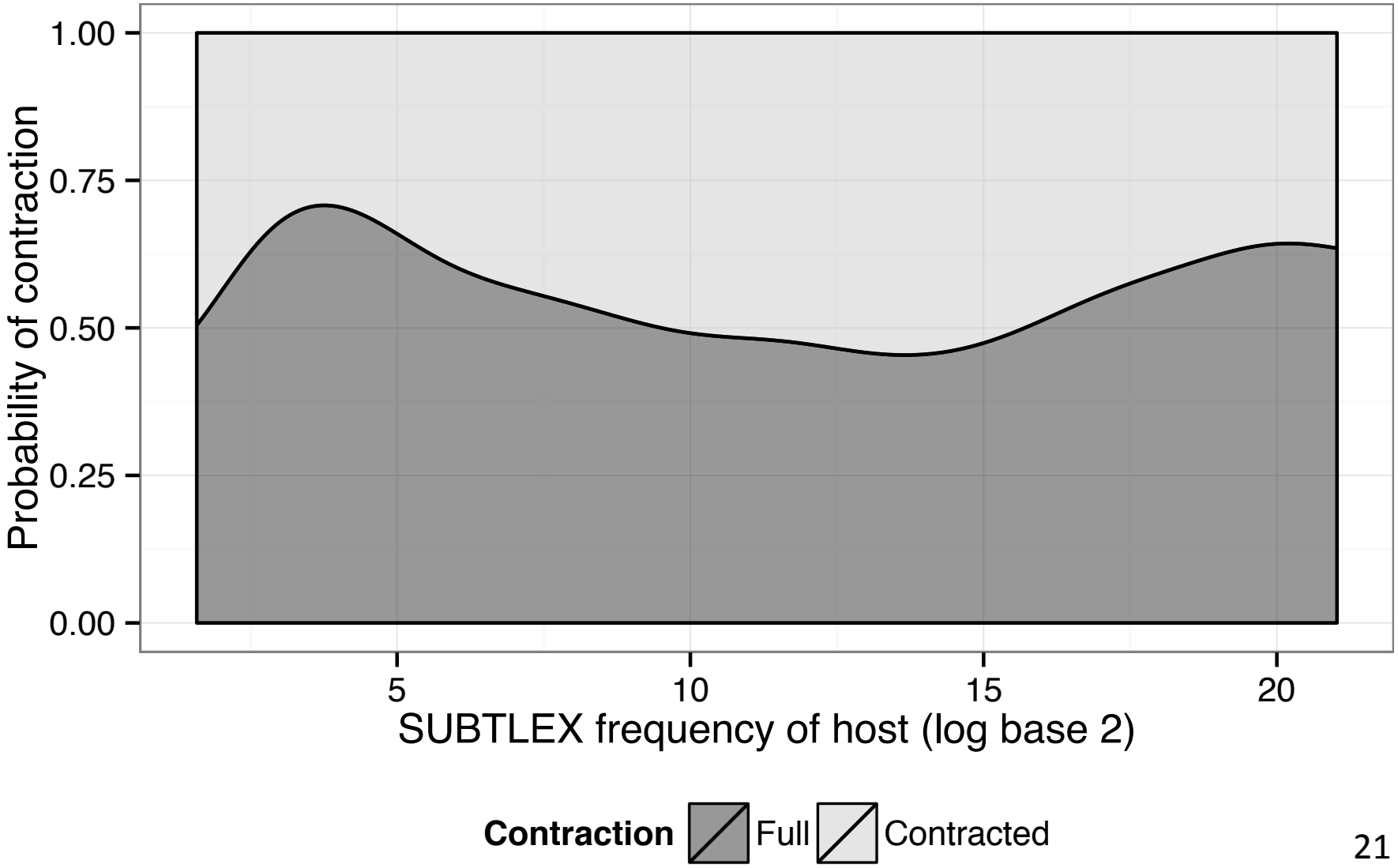
Forward probability



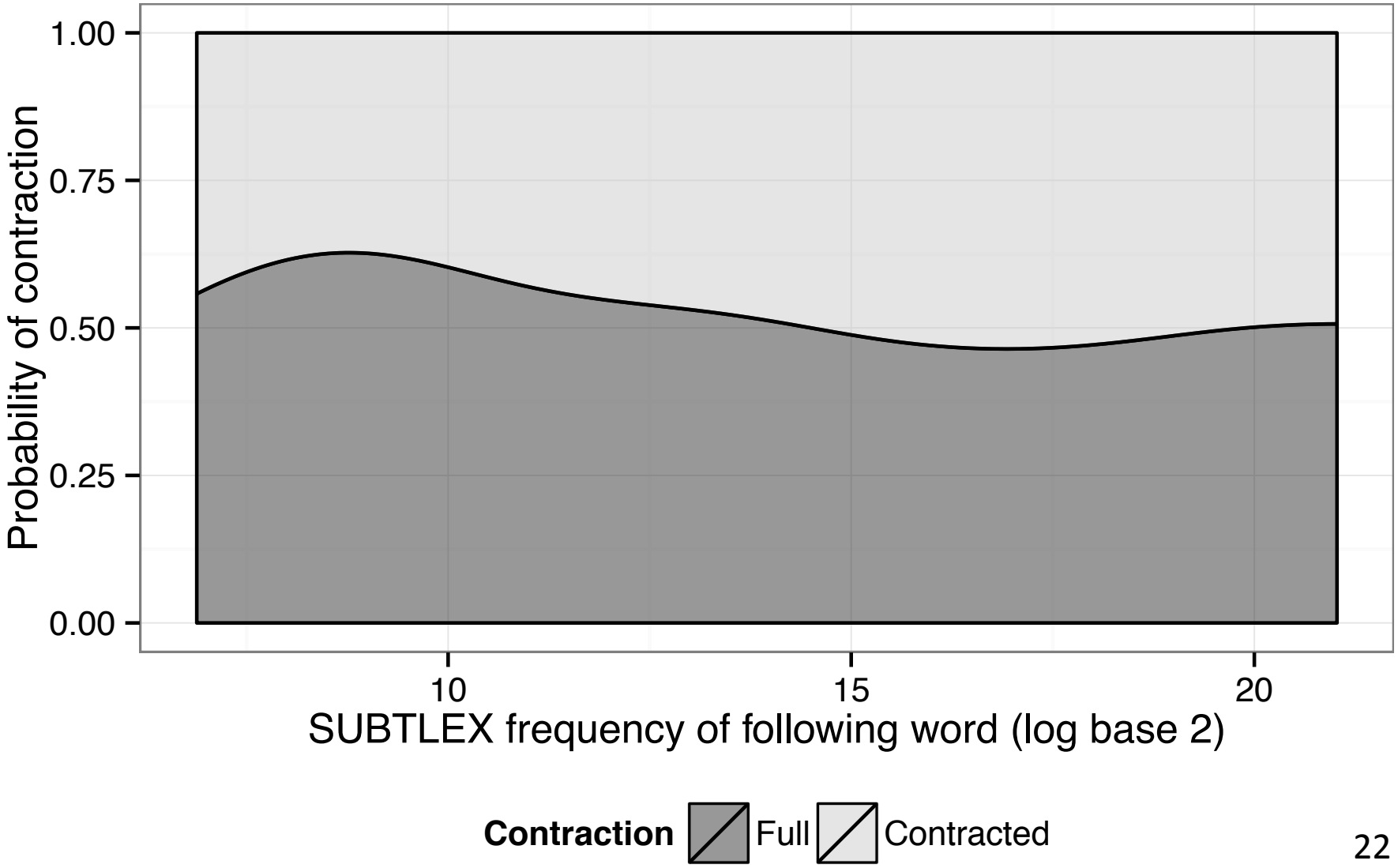
Backward probability



Host frequency



Following word frequency



Modeling methodology

- Mixed effects logistic regression using lme4.0
- Control fixed effects:
  - Subject-based: age***, gender*, education level***
  - Token-based: corpus, speaking rate, aux. identity, preceding segment (C or V) and stress
- Fixed effects of interest:
  - Subject size, frequency of host and following word (log), forward and backward probability of aux. (log)
- Random intercepts: preceding and following word, speaker dialect region, speaker identity

It's the size *and* how you use it…

Predictor	Estimate	Std. Error	P(>|z|)	
Intercept	-0.812	0.766	0.289	
Number of words	-0.440	0.074	3.34E-09	***
Host freq.	0.063	0.036	0.078	.
Following word freq.	0.070	0.033	0.036	*
Forward prob.	0.242	0.060	5.71E-05	***
Backward prob.	0.098	0.059	0.095	.

(Significance codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1)

…but size matters much more.

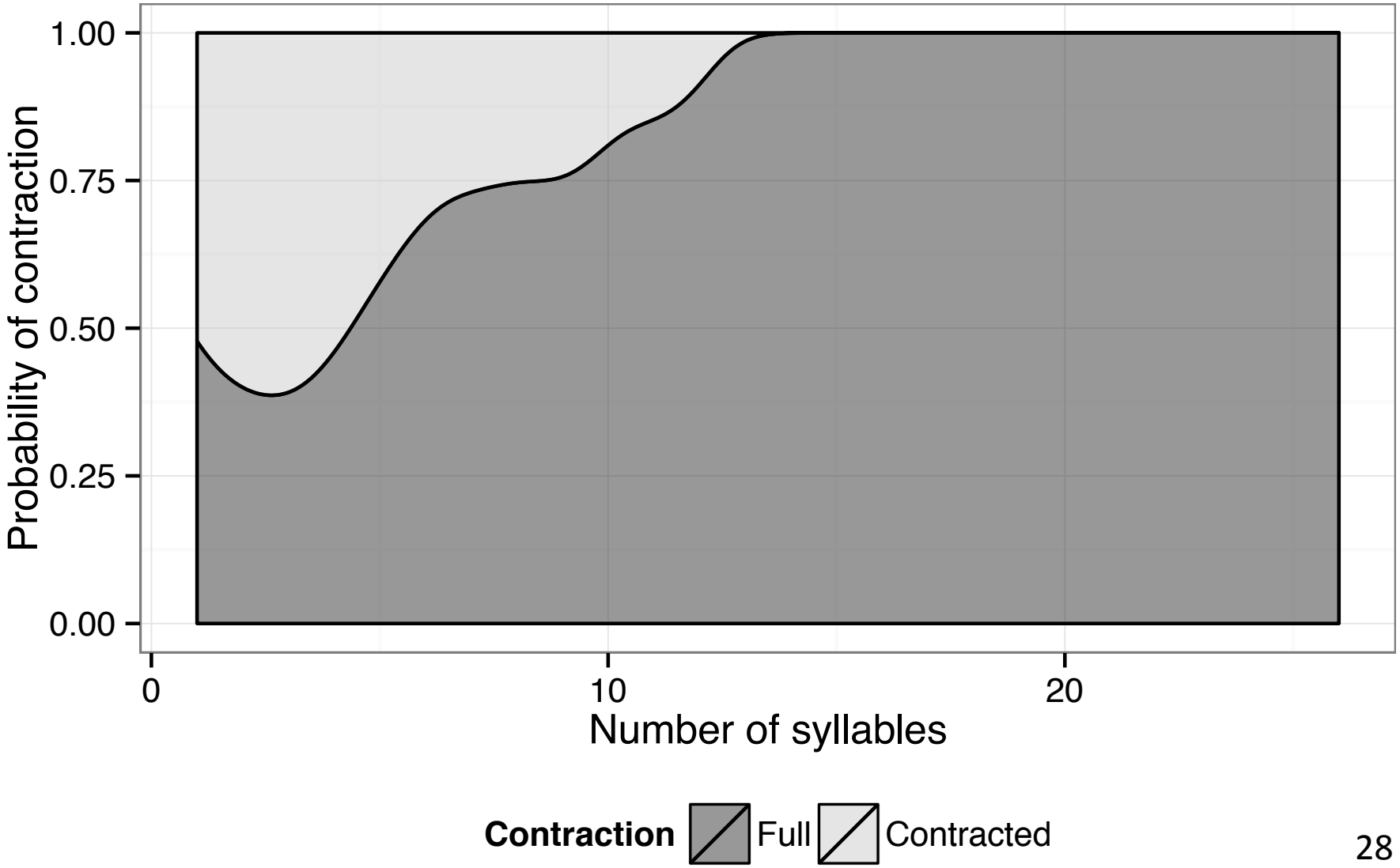
- *Impact*: effect of predictor across meaningful range
- Number of words
  - *Mary* (1) vs. *the county that I live in now* (7)
  - Increase of 6 words → **14.1x** less likely to contract
- Forward probability
  - *Communism is* (.364) vs. *work is* (.00787)
  - 46x (2^5.5) decrease → **3.8x** less likely to contract

What is size, really?

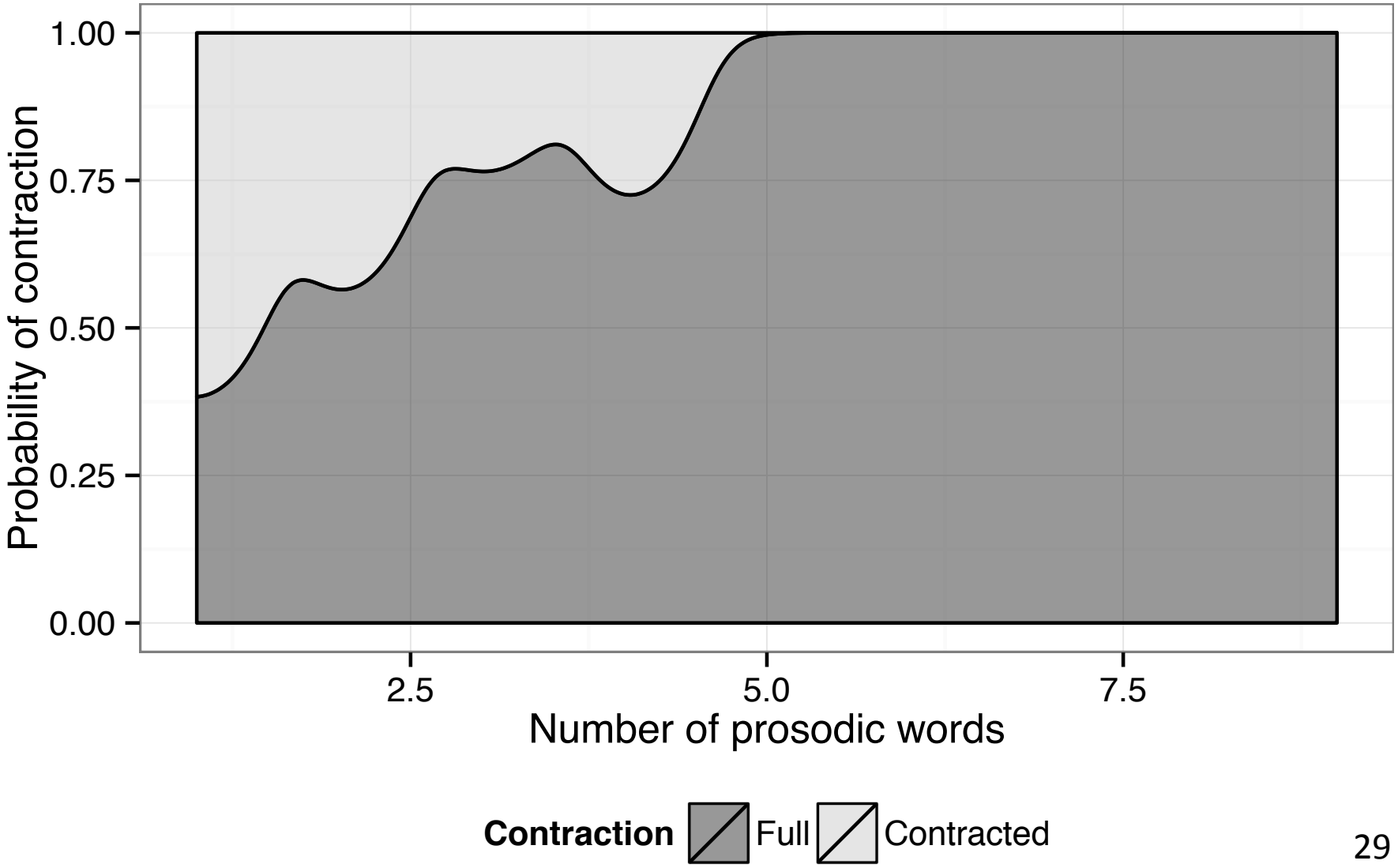
Operationalizing length

- Baseline: orthographic words
- Phonological
  - Number of syllables
  - Number of prosodic words, function words
    (following Selkirk, 1984, 1995)
  - Some combination of these?
- Syntactic
  - Height (maximal embedding) of subject parse
    (syntactic parses from Penn Treebank (Marcus et al., 1993))

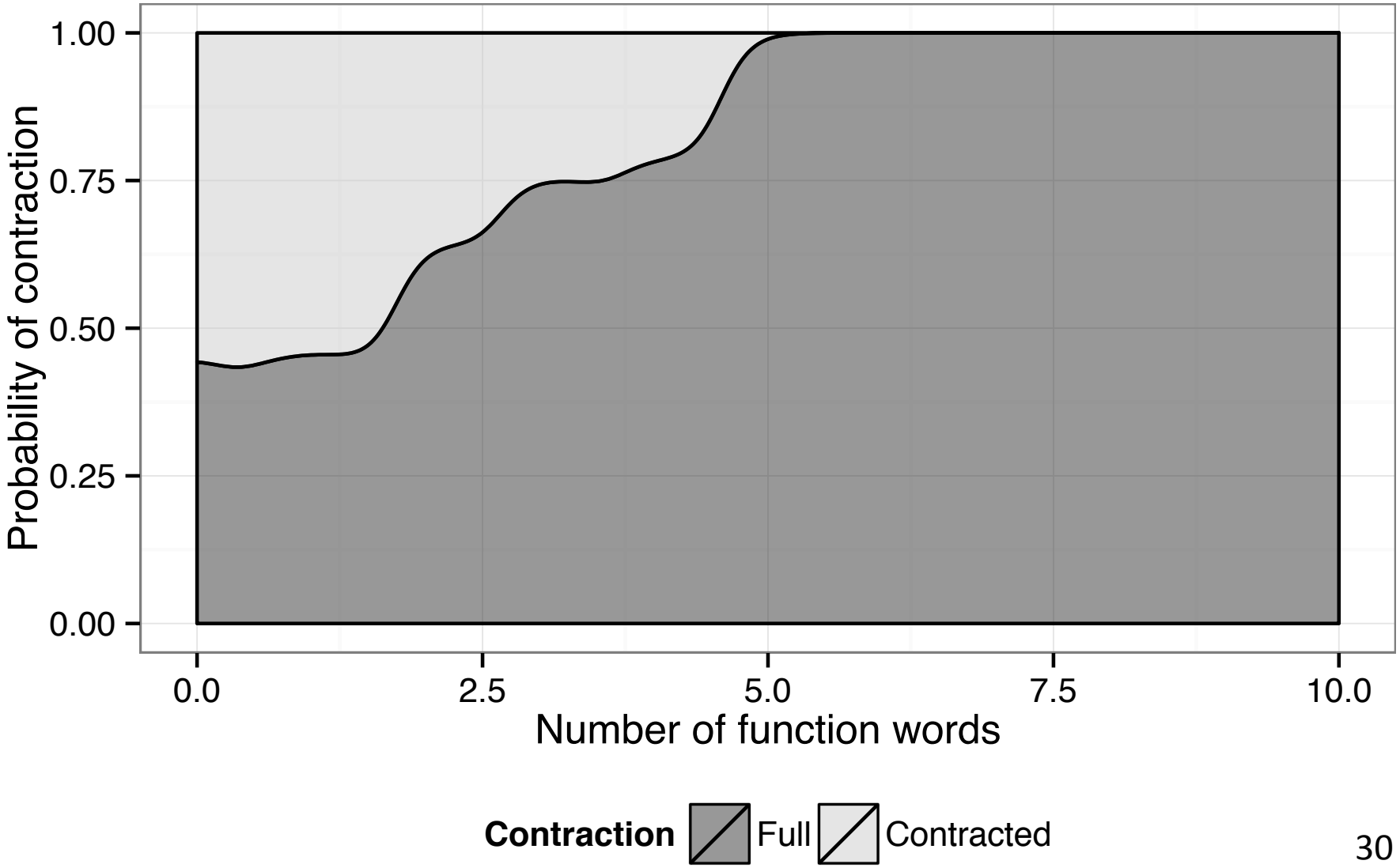
Number of syllables



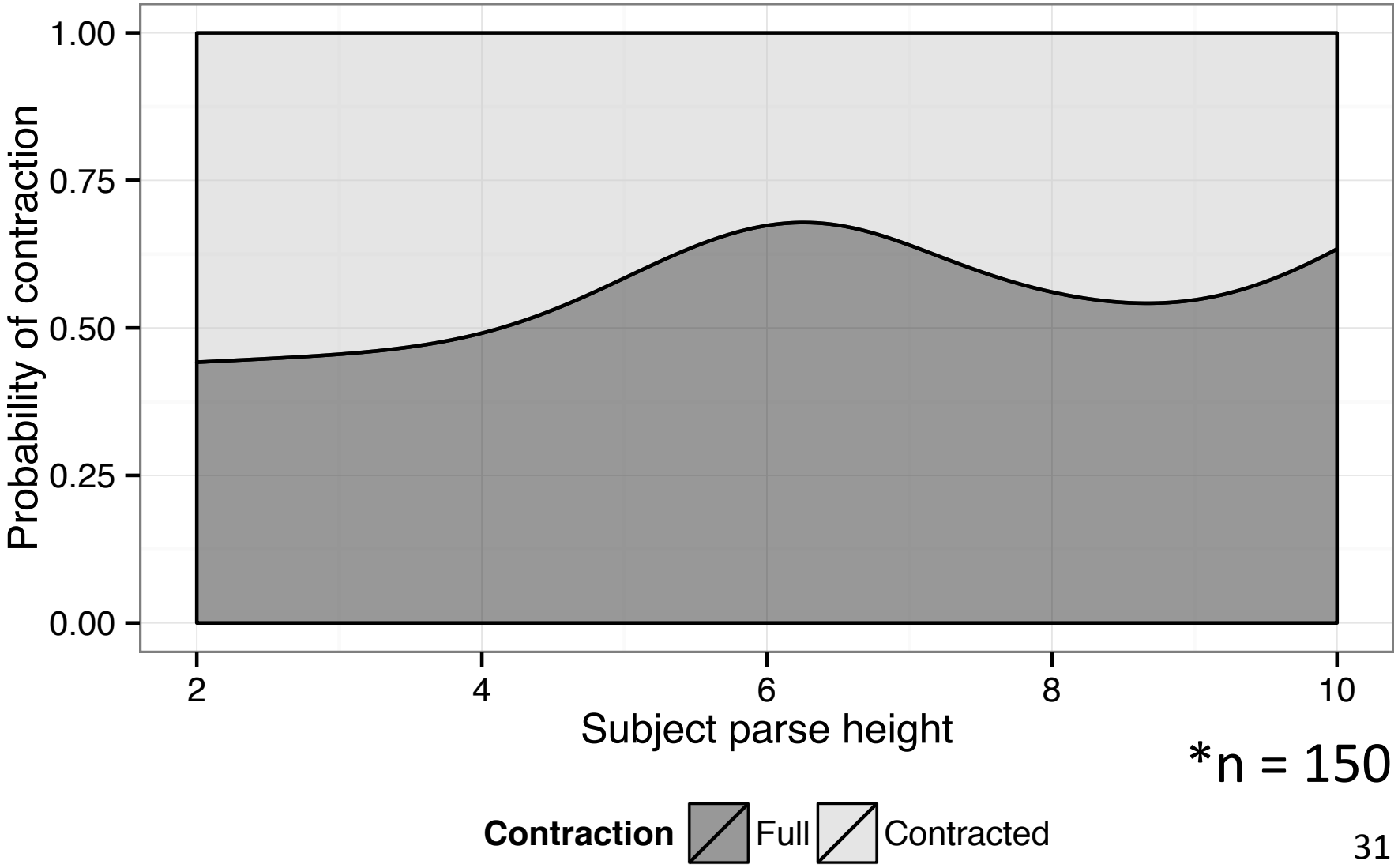
Number of prosodic words



Number of function words



Subject parse height*



*n = 150

Comparing measures of length

- High level of correlation across length measures makes dissociation difficult
- Replacing number of orthographic words with other predictors results in worse fits (AIC/BIC)
- Splits of words into separate predictors does not produce better fits (Chisq. LL ratio test)
  - e.g., n. function words + other words, n. single syllable function words + other words

Conclusions

- It's the size **and** how you use it…
- …but it's mostly the size
- Size and predictability are independent
- Number of words is the best predictor of subject size and is not reducible to other predictors

Extensions

- Possible confounds of forward predictability
  - Real relationship may be between auxiliary and subject head, not last word of subject
- Explanation from speech planning and/or prosodic phrasing
- Connection to other length-conditioned variable phenomena (e.g., heavy NP shift, dative alternation)