LSA 125 – Psycholinguistics and Syntactic Corpora

Today: *Accounts of syntactic variation (2)*

LSA Summer Institute 2009, UC Berkeley

Florian Jaeger

TAs: Judith Degen, Alex Fine, and Peter Graff

Today

- **Alternative audience design accounts**
- **Uniform Information Density** – a computational account of efficient language production
- **Alternatives to processing accounts**

- **Have you found a work group**?

Ambiguity Avoidance

- The last account of speakers' choices to be discussed today builds on ideas from Audience Design (Brennan & Williams, 1995; Clark & Murphy, 1982; Clark, 1992; Lockridge & Brennan, 2002)

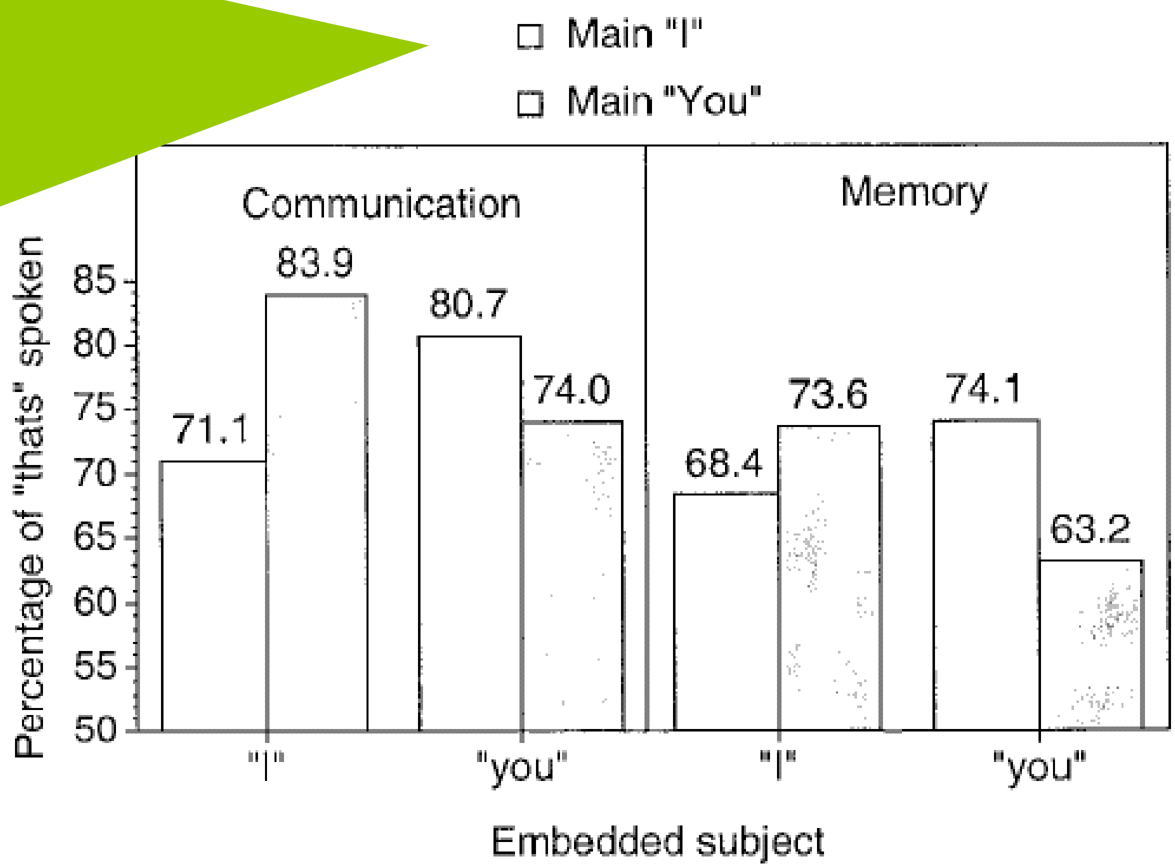
- Speaker may avoid ambiguity, or –more specifically, so called garden paths for comprehenders (Bolinger, 1972; Hawkins, 2004; Snedecker & Trueswell, 2003; Temperley, 2003)

A different type of Audience Design?

(Ferreira & Dell, 2000:324)

[…] speakers (the *communication* group) looked at and spoke to their addressees when they produced the sentence. Addressees rated the clarity of speakers' productions with pen and paper on a 7-point scale. […] The group of speakers not paired with addressees (the *memory* group) performed the identical task, except that they did not speak to an addressee and no instructions about maximizing clarity were given.

- More complementizers when audience present.



Recall also Haywood et al. (2005)

- Higher rate of disambiguating cues if confederate is helpful

TABLE 1

Proportion of Target Responses Including That's *or Any Disambiguating Word*

Condition	Helpful confederate	Unhelpful confederate
Responses with *that's*		
Prime without *that's*		
Unambiguous (one referent) context	.13 (.20)	.17 (.26)
Ambiguous (two referent) context	.25 (.30)	.17 (.26)
Prime with *that's*		
Unambiguous (one referent) context	.50 (.38)	.33 (.32)
Ambiguous (two referent) context	.53 (.40)	.49 (.38)
Responses with any disambiguating word or words		
Prime without *that's*		
Unambiguous (one referent) context	.15 (.19)	.17 (.26)
Ambiguous (two referent) context	.29 (.30)	.18 (.26)
Prime with *that's*		
Unambiguous (one referent) context	.50 (.38)	.33 (.32)
Ambiguous (two referent) context	.60 (.38)	.50 (.38)

Different types of Audience Design

- Maybe speakers do design their utterances to their audience, but they do not bother to avoid ambiguity, or at least not most:
  - Real ambiguities, as in *cases that can create serious garden paths,* are rare (cf. Jaeger, 2006, submitted; see also **pragmatic ambiguity**, Wasow, 2002)
- Collateral signals (Jaeger, 2005): *that* in complement and relative clauses could *signal* production difficulty (Clark & Fox-Tree, 2002)

Table 6: Model improvement for each of the disfluency measures

	Fillers		Suspension/Restart	
	In NP	In NSRC	In NP	In NSRC
Coefficient in model	-0.02	0.89	-0.2	0.55
Change in -2log-LH	0	19.5	0.4	11.8
Significance level of χ^2	n.s.	$p < 0.001$	n.s.	$p < 0.001$

Type of Audience Design

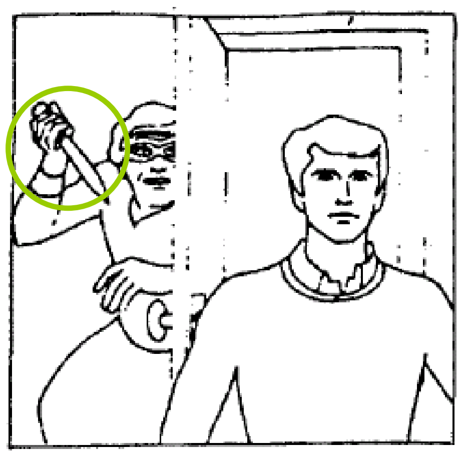
- Particular adjustments :
  - exaggerated speech to infants; speaking up to distant addressees
  - native speakers to nonnatives (Bortfeld & Brennan, 1997)
  - speakers with different conversational goals than addressees (Russell & Schober, 1999; Wilkes-Gibbs, 1986)

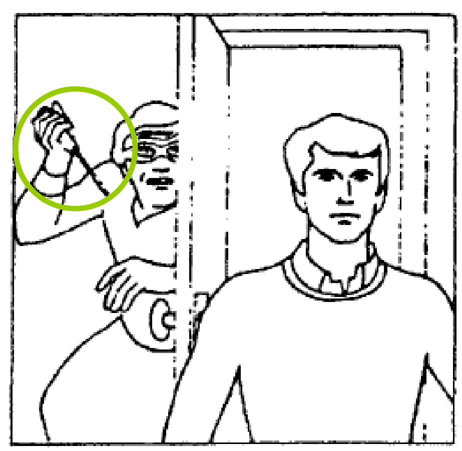
- Generic adjustments make speech easier to understand by the *average* listener (Brown & Dell, 1987)
  - pronouncing unpredictable words more clearly than predictable words (Lieberman, 1963)
  - Shortening given words (Bard et al., 2000; Fowler & Housum, 1987; Samuel & Troicki, 1998)

Mentioning of atypical instruments

(Brown & Dell, 1987; Dell & Brown, 1991; vs. Lockridge & Brennan, 2002)

- Speakers listen to story and see pictures
  - Instruments are either typical for action or atypical
- Speakers retell story to listener
  - Listener has access to pictures or not





Adolph hid behind the door and when the man entered the kitchen he stabbed him in the back. He wiped the blood off the icepick and rummaged through the drawers. Later police investigators found his fingerprints all over the icepick and had no trouble catching him.

Results with confederate

- No audience design found if listener was confederate (=informed participant, as opposed to a real participant in the study):
  - Speaker showed same rate of within-clause mentioning of instruments, regardless of whether listeners had visual access to the instrument information.
  - Only feedback-based corrections were found
- Brown & Dell (1987) and Dell & Brown (1991) concluded that there is no generic audience design, but that production and comprehension are usually highly aligned anyway (→ only **grammaticalized audience design?**)

Results with real participant

(Lockridge & Brennan, 2002)

Table 2
Percentages of Explicit Mention for Typical and Atypical Instruments for Each Copresence Condition

Category	No Visual Copresence		Separate Display Copresence		Full Copresence	
	Typical	Atypical	Typical	Atypical	Typical	Atypical
Explicit mention						
Within clause						
After the verb	30.73	41.88	31.05	32.81	29.69	30.37
Before the verb	2.08	5.76	2.11	3.13	3.65	6.28
Incorporated	1.04	1.05	3.16	2.08	2.08	1.05
Total	33.85	48.69	36.32	38.02	35.42	37.70
Separate clause						
After the verb	2.60	2.09	3.16	2.60	1.56	3.14
Before the verb	3.13	5.24	3.68	5.73	3.13	3.66
Total, explicit mention	39.58	56.02	43.16	46.35	40.10	44.50
Implicit mention	52.08	35.08	50.00	41.15	49.48	41.36
Other	8.34	8.90	6.84	12.50	10.42	14.14
Total	100.00	100.00	100.00	100.00	100.00	100.00

Table 3
Percentages of Indefinite References (e.g., Using *a*, *some*) in First Mention of Typical and Atypical Instruments in the Three Copresence Conditions

Copresence Condition	Typical	Atypical
No visual copresence	19.7	30.5
Separate display copresence	19.4	22.3
Full copresence	21.3	25.0

Are languages designed for communicative success?

- Efficiently organized lexicon: **Frequency/probability ~ length of words** (Zipf, 1929; Mandelbrot, 1965; Manin, 2006; Piantadosi et al., 2009; Plotkin & Nowak, 2000)

- Efficient lexicon and grammar via processing pressures:
  - **Performance Grammar Correspondence Hypothesis** (Hawkins, 1994, 2004, 2007)
  - **Automatization, grammaticalization, training ~ reduction** (Bybee, 1998; Bybee & Thompson, 2000; Givon, 1979; Thomspon & Mulac, 1991, …)
  - …

- Efficient language production

Uniform Information Density

***Given a choice*, speakers prefer to keep the amount of information transmitted per unit time uniform.**

Jaeger (2006, in prep), Levy & Jaeger (2007), based on Genzel & Charniak (2002) and Aylett & Turk (2004)

Information Theory: Communication through a noisy channel is optimal if information is *uniformly* close to channel capacity. (Shannon, 1948)

→ If UID affects incremental online production …
Choice points at all levels of linguistic processing should be affected by the information density

‘Choices’ at many levels in production

Utterance level:	*Move the triangle to the left.*
	Select the triangle. Move it to the left.
Phrasal level:	*She gave {him the key/the key to him}*
	She already ate (dinner)
	She stabbed him (with a knife).
Word level:	*I read a book (that) she wrote.*
Morphological level:	*I’ve\have gone there.*
Phonological level:	*t/d*-deletion; final cluster reduction; vowel weakening
Phonetic level:	formant energies, F1/F2 ratio, speech rate

What does this view buy us?

- A **uniform account** that holds at all levels of production rather than being custom-tailored to specific phenomena.
  - UID is optimal in *several* ways [Levy & Jaeger, 2007]
- UID **generates novel predictions at many levels of production** -- e.g. let's look at phonetic/phonological production:
  - The information content (redundancy) of a word in its context should affect how we produce it.
  - Previous accounts have focused on the availability of *upcoming* material [e.g. *Principle of Immediate Mention*, Ferreira & Dell, 2000; also Bock & Warren, 1985; Levelt, 1981]

Evidence for efficient phonetic and phonological production

- **Predictability ~ word realization**
  - Duration of word/morpheme/syllable instances (Aylett & Turk, 2004; Bell et al., 2003, 2008; Jespersen, 1922; Pluymaekers et al., 2005a,b)
  - Phonetic realization of segments (van Son & Pols, 1998, 2002; van Son & van Santen, 2005)
  - Phonological realization (*t/d* deletion; vowel weakening; Bell et al., 2003)
  - Predictability ~ intonational accents of word instances (Brenier et al., 2006; Pan & Hirschberg, 2000; Watson et al., 2008)

Information can be defined in terms of probability: I(u) = -log p(u)= log 1 / p(u)
[Shannon 1948]

Evidence for UID as a general principle of efficient production?

Discourse level

Utterance level:

Phrasal level:

Word level:

Morphological level: ?

Phonological level: more information dense → less likely to have t/d-deletion & vowel weakening

Phonetic level: more information dense → pronounced longer, with more F1/F2 contrasts

Information Density & Auxiliary Contraction

Frank & Jaeger (2007-AMLaP; 2008-CUNY; 2008-CogSci; in prep)

Brain and Cognitive Sciences, University of Rochester

Morpho-syntactic Reduction

- What determines speakers' choice between contracted and full forms?
  - Unlike for *that*-omission, no meaning differences have been claimed [cf. Bolinger, 1972; Dor, 2005; Yaguchi, 2001]

Contracted

… and I'm never happier than when I am a kangaroo …

Same information spread over more time/words

Full

… and I am happier than when I am a kangaroo …

→ **UID prediction:** Speakers should use full form if information conveyed by contractible element is high

How to estimate the information carried by a contractible element

and I AM never …

w-2 *w-1* *w*

$$I(\text{AM} \mid \text{context})$$

$$=$$

$$-\log p(\text{AM} \mid \text{context})$$

$$\approx$$

$$-\log p(\text{AM} \mid \text{“}\textit{and I}\text{”})$$

$$=$$

$$-\log [\, p(\text{“}\textit{am}\text{”} \mid \text{“}\textit{and I}\text{”}) + p(\text{“}\textit{'m}\text{”} \mid \text{“}\textit{and I}\text{”})\,]$$

Information theoretic definition of Shannon information content

Use trigram model to estimate probability (backoff)

Data

- Extraction of utterances from a large corpus of spontaneous AE speech (Switchboard, Godfrey et al., 1992 ~800,000 sentences in 650 dialogues)

  - **HAVE: e.g. *'d* vs. *had* (>2,400 contractible cases)**
  - **NOT: *n't* vs. *not* (> 5,000 contractible cases)**
  - **BE: e.g. *'s* vs. *is* (> 9,000 contractible cases)**

Analysis

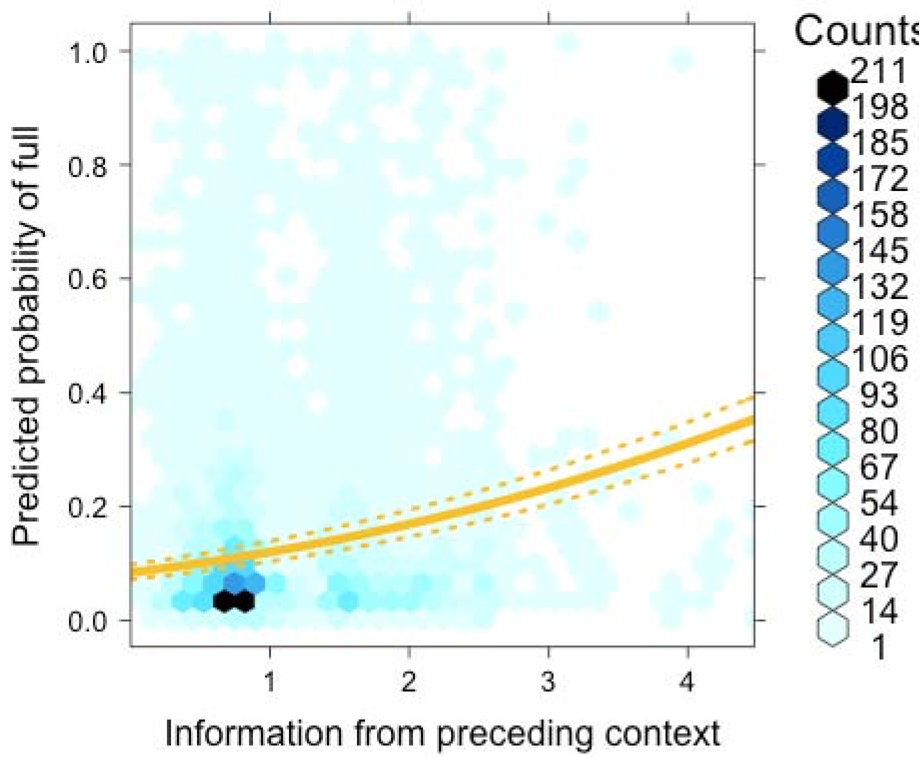
- Multilevel logit analysis to analyze when speakers' choose **full over contracted** forms depending on the information carried by it.

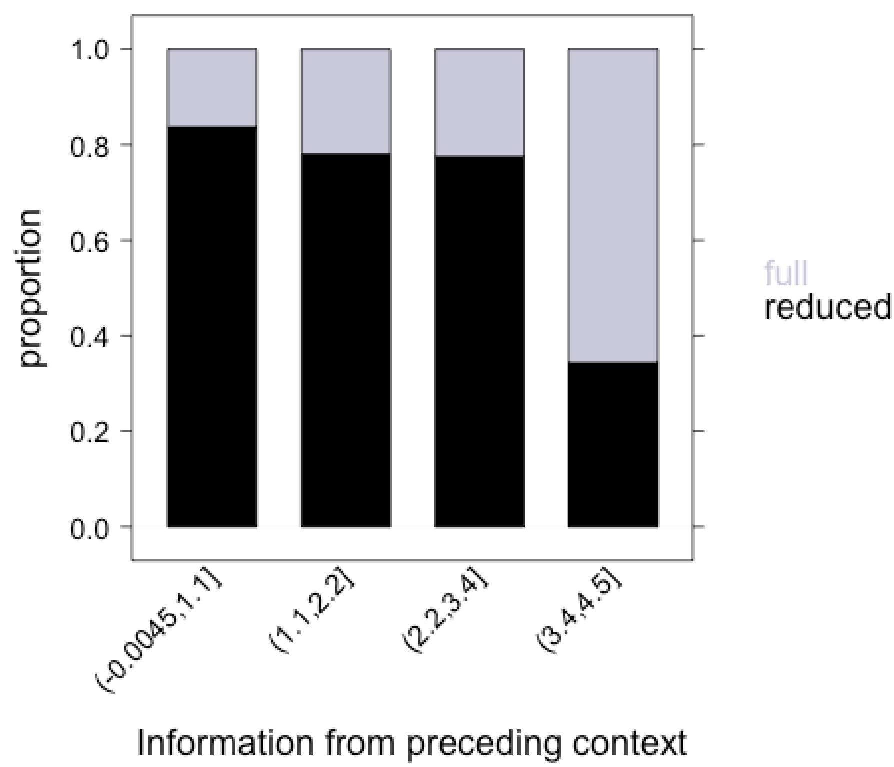
$$\text{logit}[\text{p}(full)] = \ln\frac{\text{p}(full)}{\text{p}(reduced)} = -\beta\log\text{p}(HAVE \mid w_{i-1}) - \beta\log\text{p}(HAVE \mid w_{i+1}) + X_{Controls}\beta_{Controls} + Zb$$

- Simultaneously controlling for:
  - Complexity of *upcoming* material
  - Complexity of *host*
  - Speech rate and fluency
  - Social effects
  - Random effects for individual differences between subjects and elicitation sessions

BE contraction

I *'m/am* *not interested …*





Redundancy given following context

and I AM never …

w-2 *w-1* *w* *w+1*

$$I(\text{AM} \mid \text{context})$$

$$=$$

$$-\log p(\text{AM} \mid \text{context})$$

$$\approx$$

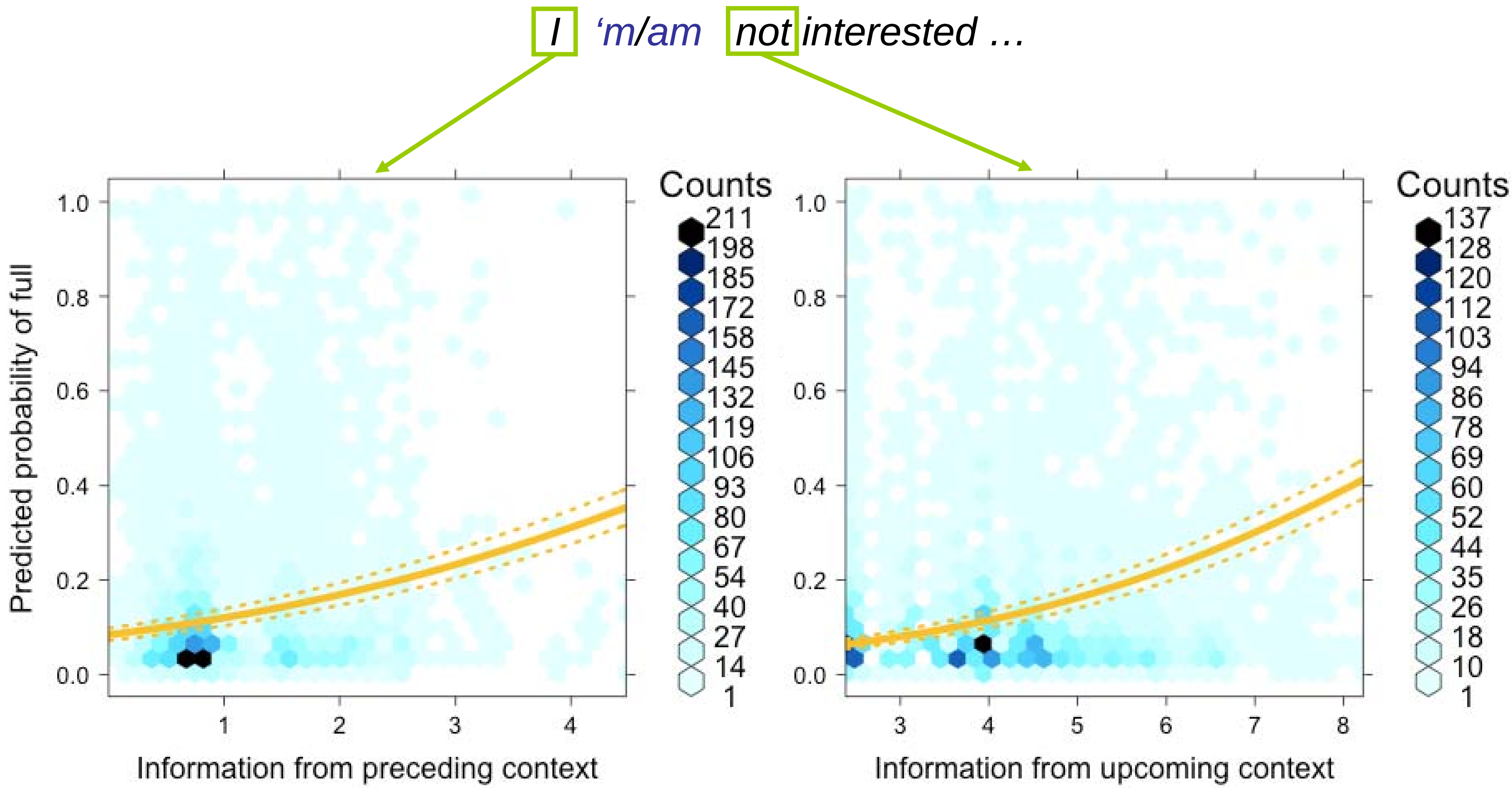
$$-\log p(\text{AM} \mid \text{“}\textit{never}\text{”})$$

$$=$$

$$-\log [\, p(\text{“}\textit{am}\text{”} \mid w{+}1{=}\text{“}\textit{never}\text{”}) + p(\text{“}\textit{'m}\text{”} \mid w{+}1{=}\text{“}\textit{never}\text{”}) \,]$$

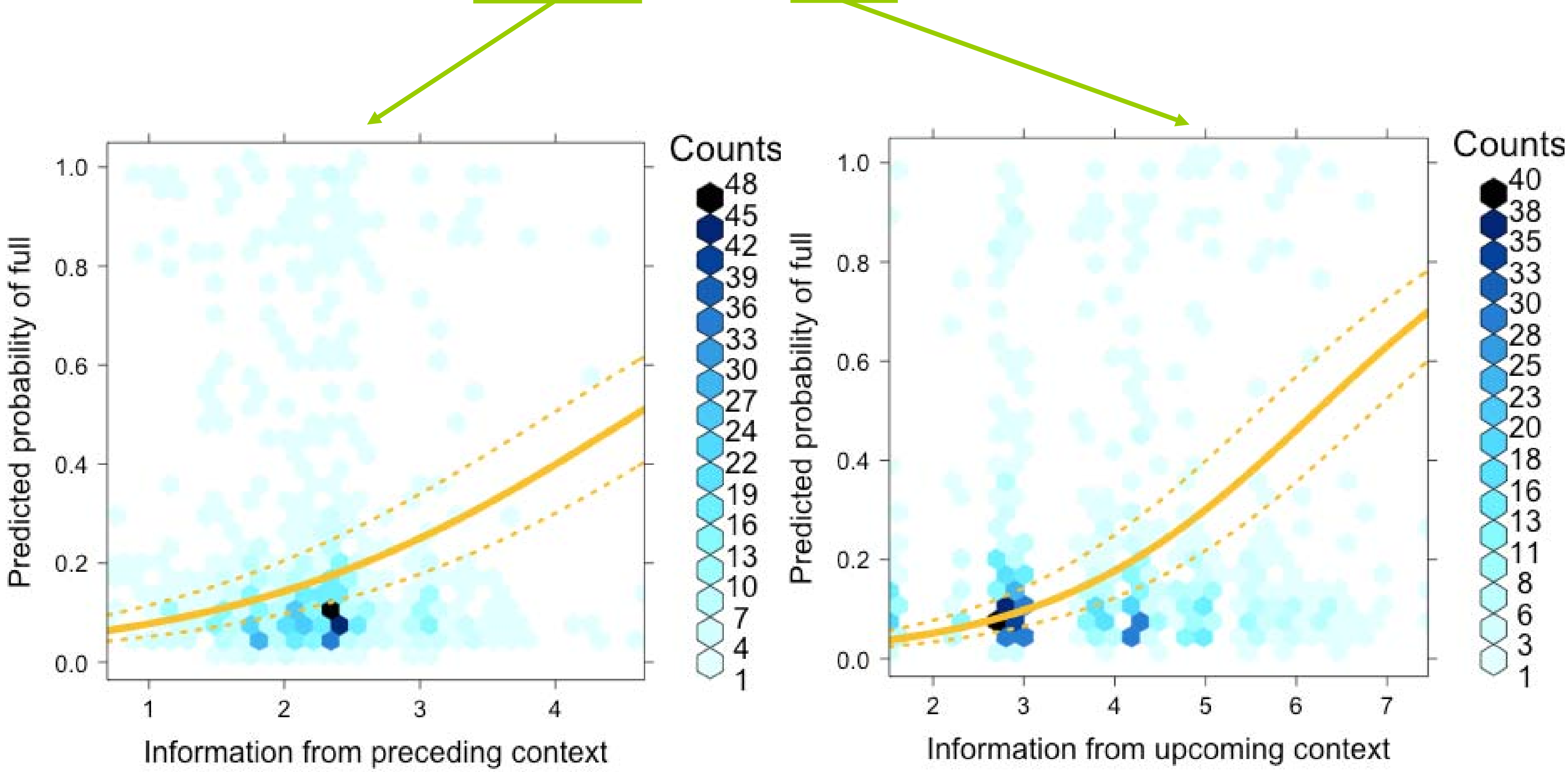
BE contraction

I 'm/am not interested …



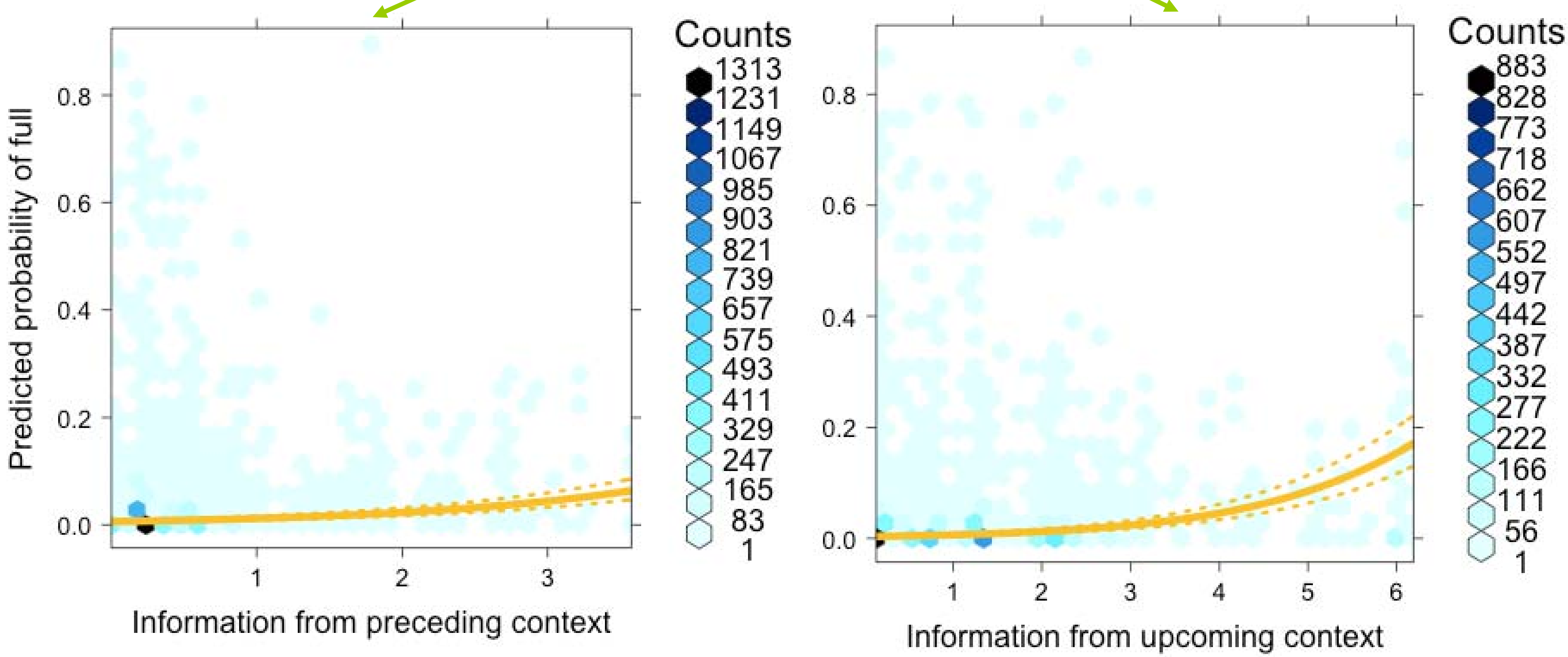
HAVE contraction

and she 's/has been there, too…



NOT contraction

Clinton did n't/not have …



Evidence for UID

Discourse level:

Utterance level:

Phrasal level:

Word level: ?

Morphological level: more information dense → less likely to be contracted

Phonological level: ok [Bell et al., 2003]

Phonetic level: ok [Aylett & Turk, 2004; Bell et al. 2003; von Son & van Santen, 2005]


UNO
UNIVERSIDAD DE ORIENTE

Information Density & Syntactic Reduction

Jaeger* (2006-thesis, submitted, in prep); Levy[+] & Jaeger (2007); Jaeger, Levy, & Ferreira (in prep); Wasow[#], Jaeger, & Orr[^] (in press)

[*]Brain and Cognitive Sciences, University of Rochester
[+]Linguistics, UC San Diego
[*]Linguistics, Stanford University
[^]Yahoo?
[$]Psychology, UC San Diego

UID & syntactic reduction

- UID predicts that **the total of phrasal and onset information density** correlates with *reduction*

$$\log \frac{1}{p(CC \mid ctxt)} \qquad \log \frac{1}{p(w_i \mid CC, ctxt)}$$

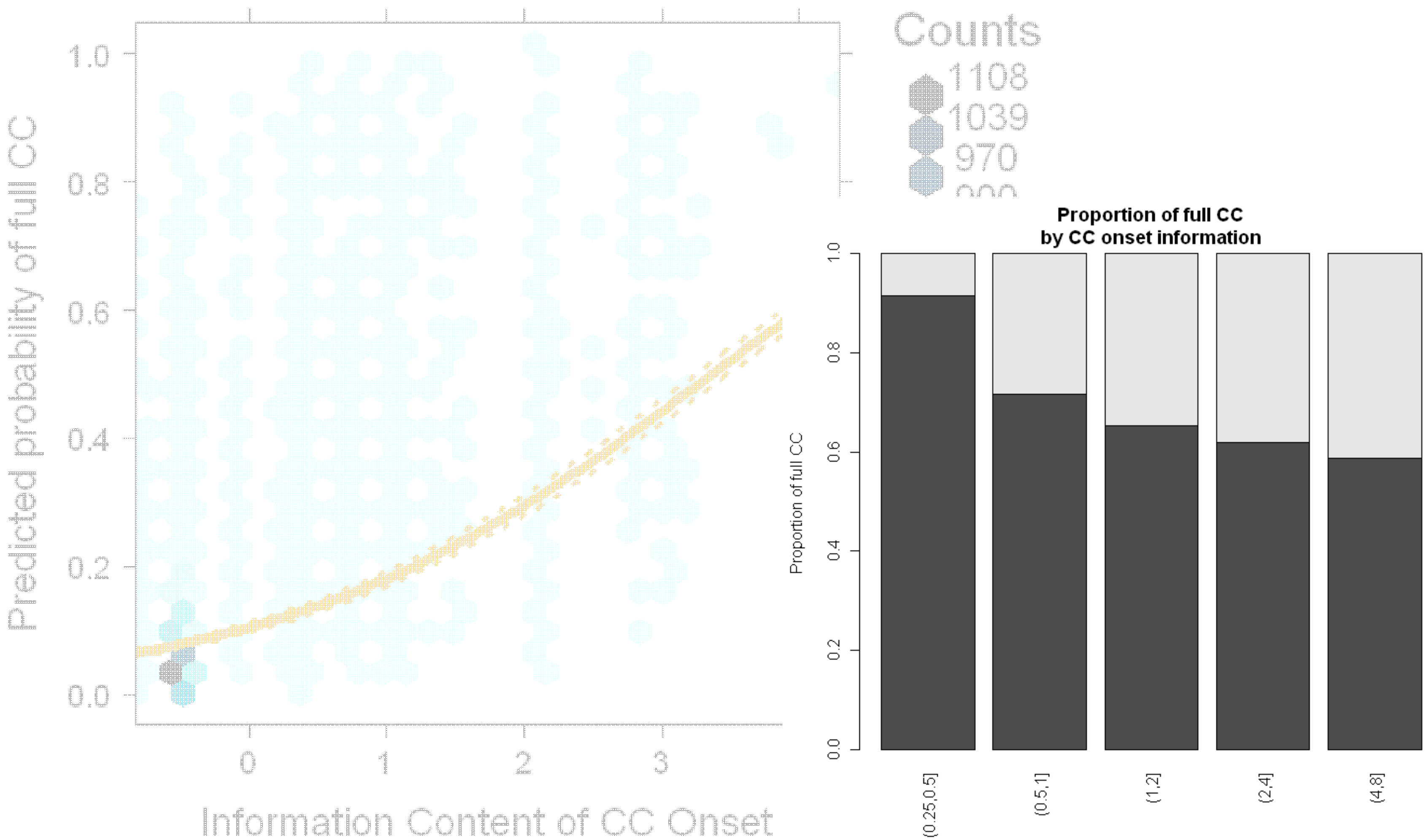
Phrase onset **Lexical content**

I assume { *you will drink with me.* / *that you will drink with me.* }

CC reduction

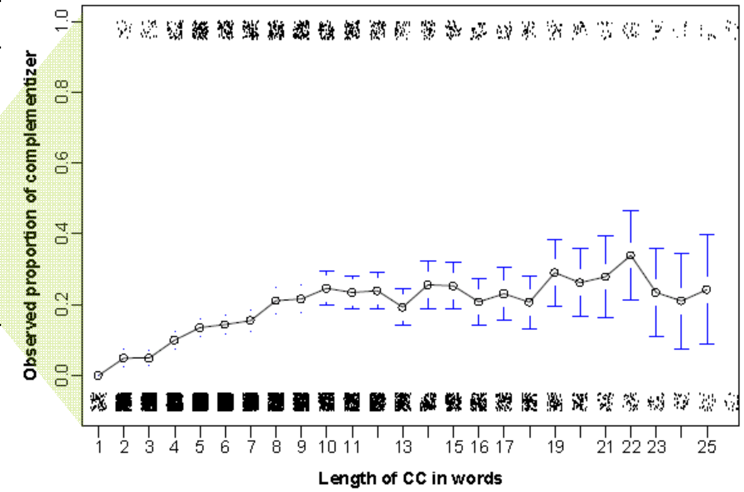
(Jaeger 2006-thesis, 2009-submitted)

I think (that) this is entirely irrelevant.



Predictor	Coef. β	SE(β)	z	p
Intercept	0.119	(0.376)	0.3	> 0.7
Position(Matrix Verb)	0.948	(0.143)	6.6	< 0.0001
(1st restricted comp.)	−27.819	(5.331)	−5.2	< 0.0001
(2nd restricted comp.)	55.185	(10.794)	−5.2	< 0.0001
Length(Matrix Verb-to-CC)	0.172	(0.065)	2.7	< 0.008
Length(CC Onset)	0.180	(0.014)	12.8	< 0.0001
Length(CC Remainder)	0.026	(0.006)	4.3	< 0.0001
Log Speech Rate	−0.700	(0.129)	−5.4	< 0.0001
Sq Log Speech Rate	−0.365	(0.190)	−1.9	< 0.06
Pause	1.100	(0.108)	10.5	< 0.0001
Disfluency	0.395	(0.122)	3.2	< 0.002
CC Subject = *it* vs. *I*	0.037	(0.077)	0.3	> 0.6
= *other pro* vs. *prev. levels*	0.053	(0.033)	1.6	$= 0.11$
= *other NP* vs. *prev. levels*	0.111	(0.023)	4.9	< 0.0001
FQ(CC Subject Head)	−0.019	(0.028)	−0.7	> 0.4
Subject Identity	−0.317	(0.166)	−1.9	< 0.056
Word Form OCP	−0.316	(0.170)	−1.9	< 0.063
FQ(Matrix Verb)	−0.208	(0.030)	−7.0	< 0.0001
Ambiguous CC Onset	−0.116	(0.115)	−1.0	> 0.3
Persistence = *no* vs. *prime w/o that*	0.019	(0.067)	0.3	> 0.7
= *prime w/ that* vs. *prev. levels*	0.058	(0.035)	1.6	$= 0.10$
Matrix Subject = *you*	0.484	(0.152)	3.2	< 0.0015
= *other PRO*	0.616	(0.125)	4.9	< 0.0001
= *other NP*	0.862	(0.128)	6.7	< 0.0001
Male Speaker	−0.157	(0.111)	−1.4	> 0.15
CC Predictability	−0.639	(0.038)	−16.6	< 0.0001

Table 3
Result summary: Coefficient estimates β, standard errors SE(β), associated Wald's z-score ($= \beta/\text{SE}(\beta)$) and significance level p for all predictors in the analysis.



Avoid *that that* sequences
(Walter & Jaeger, in press)

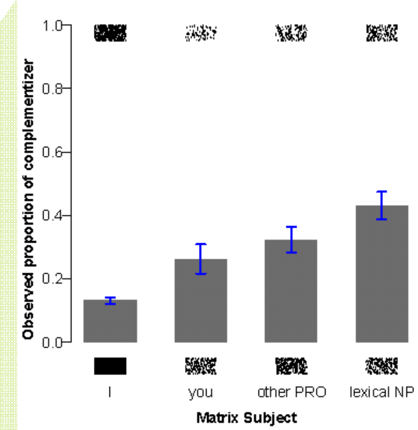


Fig. 5. Effect of matrix subject on *that*-mentioning

Predictor	Coef. β	SE(β)	z	p
Intercept	0.119	(0.376)	0.3	> 0.7
POSITION(MATRIX VERB)	0.948	(0.143)	6.6	< 0.0001
(1st restricted comp.)	−27.819	(5.331)	−5.2	< 0.0001
(2nd restricted comp.)	55.185	(10.794)	−5.2	< 0.0001
LENGTH(MATRIX VERB-TO-CC)	0.172	(0.065)	2.7	< 0.008
LENGTH(CC ONSET)	0.180	(0.014)	12.8	< 0.0001
LENGTH(CC REMAINDER)	0.026	(0.006)	4.3	< 0.0001
LOG SPEECH RATE	−0.700	(0.129)	−5.4	< 0.0001
SQ LOG SPEECH RATE	−0.365	(0.190)	−1.9	< 0.06
PAUSE	1.100	(0.108)	10.5	< 0.0001
DISFLUENCY	0.395	(0.122)	3.2	< 0.002
CC SUBJECT = *it* vs. *I*	0.037	(0.077)	0.3	> 0.6
= *other pro* vs. *prev. levels*	0.053	(0.033)	1.6	$= 0.11$
= *other NP* vs. *prev. levels*	0.111	(0.023)	4.9	< 0.0001
FQ(CC SUBJECT HEAD)	−0.019	(0.028)	−0.7	> 0.4
SUBJECT IDENTITY	−0.317	(0.166)	−1.9	< 0.056
WORD FORM OCP	−0.316	(0.170)	−1.9	< 0.063
FQ(MATRIX VERB)	−0.208	(0.030)	−7.0	< 0.0001
AMBIGUOUS CC ONSET	−0.116	(0.115)	−1.0	> 0.3
PERSISTENCE = *no* vs. *prime w/o that*	0.019	(0.067)	0.3	> 0.7
= *prime w/ that* vs. *prev. levels*	0.058	(0.035)	1.6	$= 0.10$
MATRIX SUBJECT = *you*	0.484	(0.152)	3.2	< 0.0015
= *other PRO*	0.616	(0.125)	4.9	< 0.0001
= *other NP*	0.862	(0.128)	6.7	< 0.0001
MALE SPEAKER	−0.157	(0.111)	−1.4	> 0.15
CC Predictability	−0.639	(0.038)	−16.6	< 0.0001

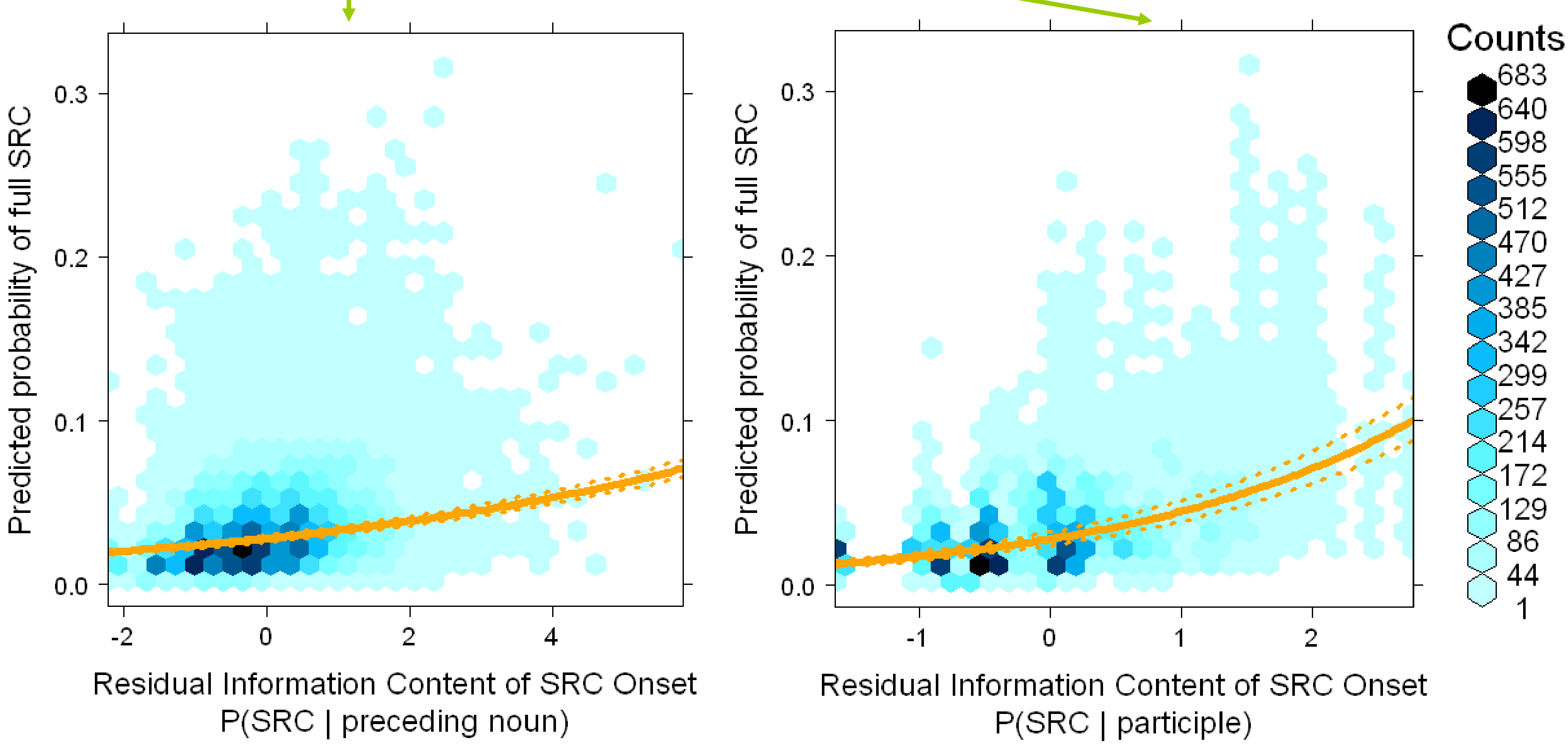
Table 3
Result summary: Coefficient estimates β, standard errors SE(β), associated Wald's z-score ($= \beta/\text{SE}(\beta)$) and significance level p for all predictors in the analysis.

Most important predictor of *that*-mentioning: $\chi^2(1) = 263.0$, $p < 0.0001$ (cf. more than 4 fluency or three domain complexity parameters combined)

SRC reduction

(Jaeger submitted)

A president (who is) adored by most of us would have no need …



- ***that*-mentioning CC reduction** in standard AE speech
  - *I think (that) this is entirely irrelevant.*
- ***that*-mentioning NRC reduction** in standard AE speech
  - *The guy (that) I saw was from …*
- ***that*-mentioning SRC reduction** in British dialects
  - *There's a man (that) lives in my neighborhood.*
- **'*whiz*-deletion' in SRC reduction** in standard BE writing
  - *A president (who is) adored by most of us would have no need …*
- **Infinitival VP reduction in BE** writing
  - *Can you help me (to) pack my suitcase?*

Wasow, Jaeger, & Orr (2006); Jaeger (2006-thesis; 2007-LSA; 2007-AMLaP); Jaeger & Wasow (2007-REL); Jaeger, Levy, Orr, & Wasow (2005/6-AMLaP); Levy & Jaeger (2007-NIPS)

Evidence for UID

Discourse level:

Utterance level: ?

Phrasal level: ok? [Resnik 1996]

Word level: ok [Jaeger, 2006, in progress; Levy & Jaeger, 2007; Wasow et al., in press]

Morphological level: ok [Frank & Jaeger, 2008]

Phonological level: ok [Bell et al., 2003]

Phonetic level: ok [Aylett & Turk, 2004; Bell et al. 2003; von Son & van Santen, 2005]

Information Density & Inter-clausal Planning

Gómez Gallo*, Jaeger[+], & Smyth° (2008-LREC, 2008-CUNY, 2008-CogSci); Gómez Gallo* & Jaeger[+] (in prep)

*CS & LIN, University of Rochester
*BCS & CS, University of Rochester
° LIN, University of Toronto

Inter-clausal planning

- Given an intended message, what determines how speakers distribute the message across clauses?

Mono-clausal

Put a tomato inside Central Park.

Bi-clausal

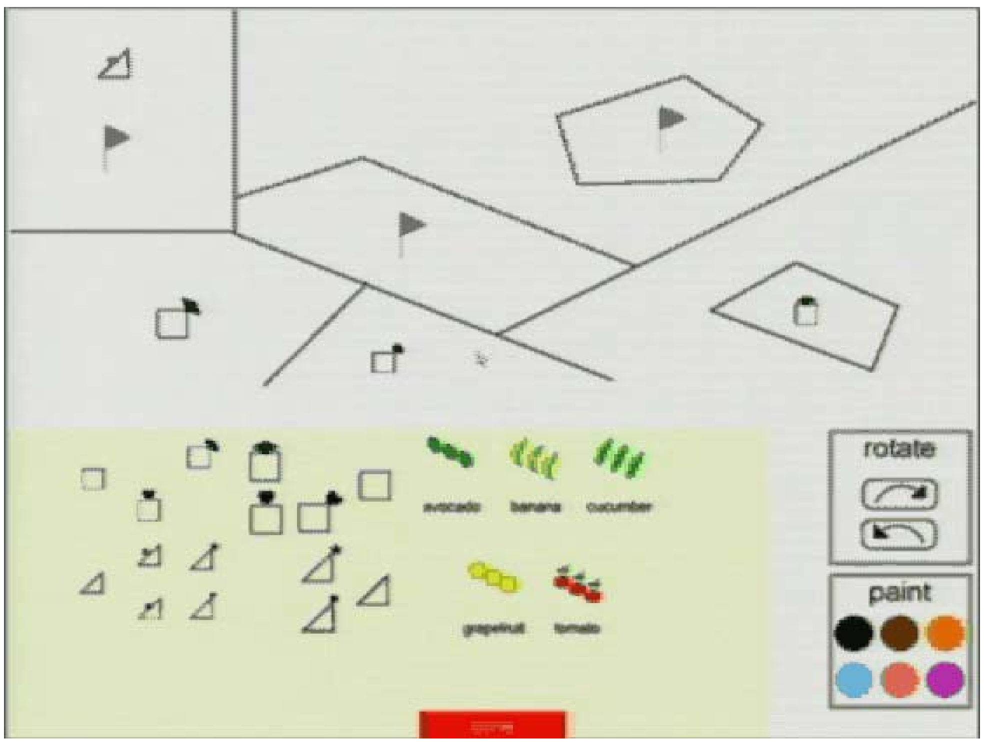
Take a tomato. Put it inside Central Park.

Same message spread over more time/words

→ **UID predictions:** Speakers should distribute information dense messages over several clauses

Data

- Spontaneous elicitation of utterances in task-oriented dialogue (Fruit carts corpus, Gómez Gallo et al., 2007)



"Take a tomato. Put it inside Central Park."

Manipulating information density

Put an apple inside the Central Park.

THEME

- THEME either easy had short label (*tomato, banana,* …) or required complex description (*the triangle with the heart at the corner*)

- LOCATION also differed. But it was hard to determine clause boundaries (… *down a little bit more and to the right*)

Natural production → many variants

- **Mono-clausal MOVE**

  Put an apple inside Central Park.

  The apple is inside Central Park.

  ...

Current study: limited to V-initial

- **Bi-clausal MOVE**

  Take an apple. Put it inside Central Park.

  Take an apple and now put it inside Central Park.

  ...

- **Left-dislocation**

  The apple, put it inside Central Park.

Estimating information of clause

- We approximate information of clause by information content of argument expressions:

Mono-clausal: *Put an apple inside Central Park.*
Bi-clausal: *Take an apple. Put it inside Central Park.*

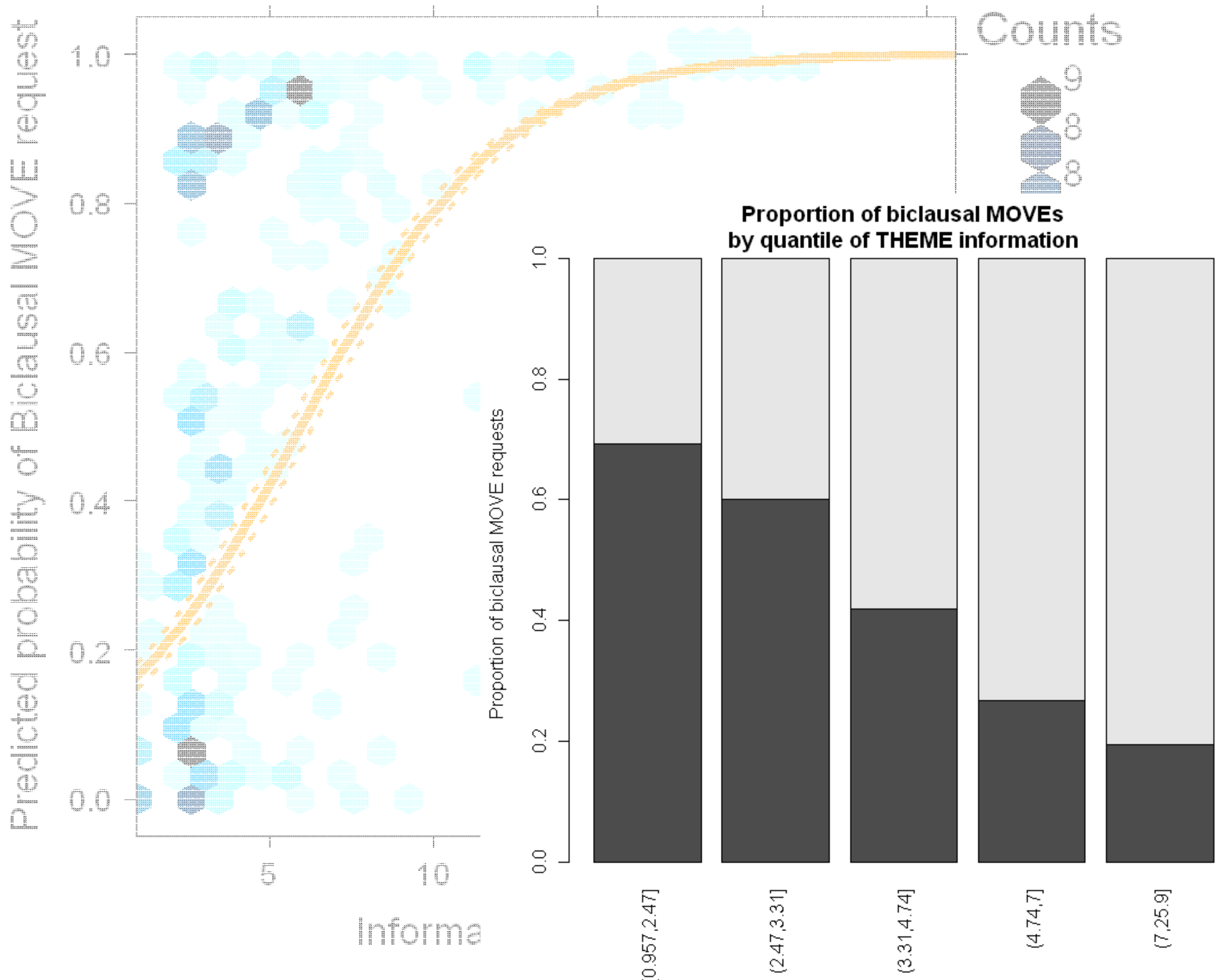
- **Shannon information of theme**:
  - Sum of information content of words
  - Word information content estimated via trigram (w/ back-off & smoothing)

$$\mathrm{I}(w_1 \ldots w_k) = -\sum_{i=1..k} \log(w_i \mid w_{i-2} w_{i-1})$$

Analysis

- Multilevel logit analysis to analyze when speakers' choose **bi-clausal over mono-clausal** utterances depending on information density of theme.

- Simultaneously controlling for:
  - *Length*, fluency, and givenness of theme
  - *Length,* information density, and fluency of location
  - Random effects for individual differences between subjects and elicitation sessions


Predicted probability of Biclausal MOVE request
1.0
0.8
0.6
0.4
0.2
0.0
5
10
Informa
Counts
9
8
8
Proportion of biclausal MOVEs
by quantile of THEME information
Proportion of biclausal MOVE requests
1.0
0.8
0.6
0.4
0.2
0.0
(0.957,2.47]
(2.47,3.31]
(3.31,4.74]
(4.74,7]
(7,25.9]

Time course of clausal planning

- When is the choice for a mono/bi-clausal structure made? Can we be sure this is really a choice about clausal planning?

- The verb already expresses a commitment to the structure and the verb is the first word in the sentence

Verb	**Mono-clausal**	**Bi-clausal**
take	0%	73%
move	28%	0%
put	27%	1%
be	43%	7%

Evidence for UID

Discourse level: ?

Utterance level: ok? [Gomez Gallo et al., 2008a,b; in progress]

Phrasal level: ok? [Resnik 1996; work in progress]

Word level: ok [Jaeger, 2006, in progress; Levy & Jaeger, 2006; Wasow et al., in press]

Morphological level: ok [Frank & Jaeger, 2008]

Phonological level: ok [Bell et al., 2003]

Phonetic level: ok [Aylett & Turk, 2004; Bell et al. 2003; von Son & van Santen, 2005]

Information per word throughout discourse in Mandarin Chinese

Qian & Jaeger (2009-LSA; 2009-CUNY; 2009-CogSci); Qian (2009-BS-thesis)

Brain and Cognitive Sciences, University of Rochester

Constant Entropy Rate (Genzel & Charniak, 2002)

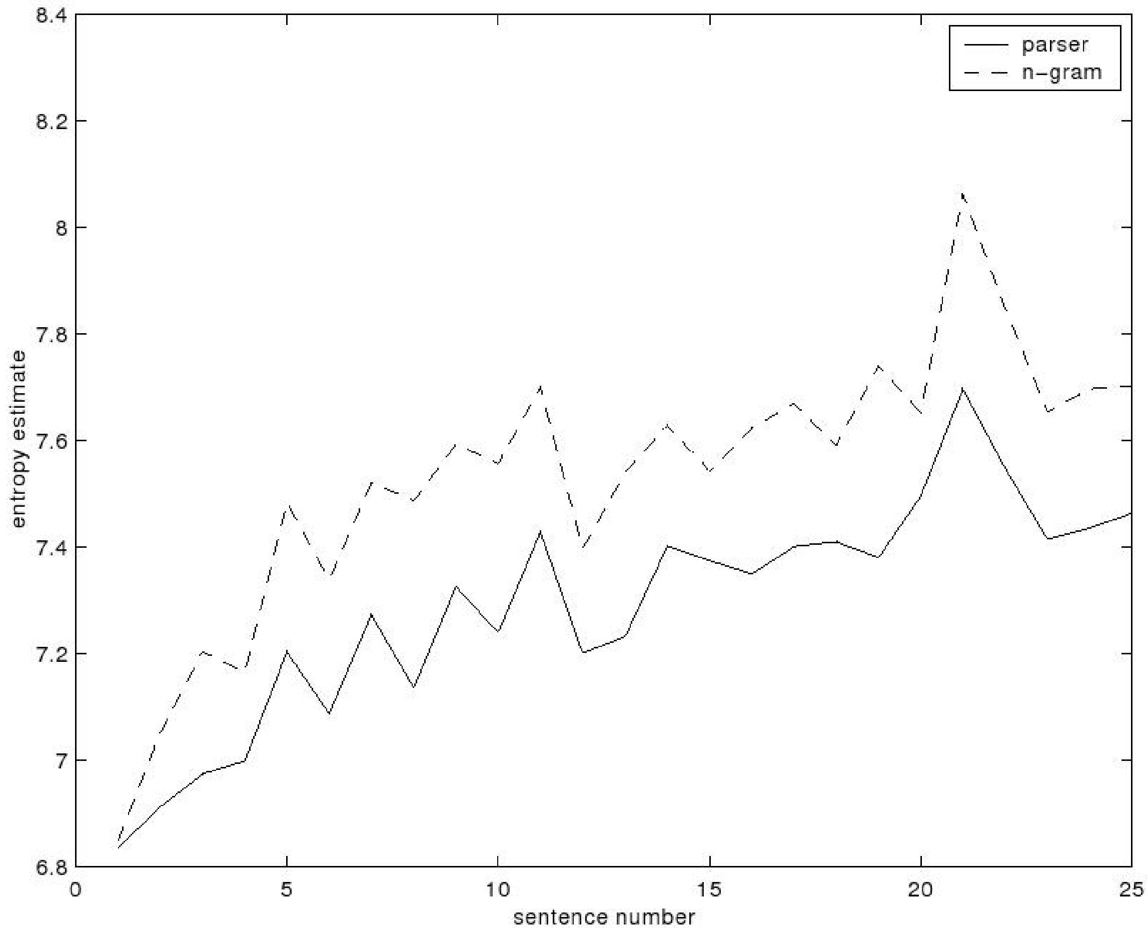
- **Constant Entropy Rate:** average amount of information in sentences constant during a discourse

- **But:**
  - Measuring context is difficult with current NLP techniques (e.g. n-grams, PCFGs).
- **Indirect test:**
  - We could look at *out-of-context* sentence information instead.

- Less context is available at the beginning of a discourse → early sentences should have lower ***out-of-context*** entropy rates.

- Sentences later in a discourse are more predictable w.r.t. the preceding context → speakers can encode more ***out-of-context*** information in later sentences.

Genzel & Charniak (2002)



Also: Genzel & Charniak (2003); Keller (2004); Piantadosi et al. (2008)

Corpus and Data

- **Corpus: Chinese speech**
  - 680,619 characters (P GALE Phase 1 Chinese Broadcast News Parallel Text , part 1 and 2)
  - 46.2 hours of Chinese broadcast news (China Central TV -a broadcaster from Mainland China- and Phoenix TV -a Hong Kong-based satellite TV station)

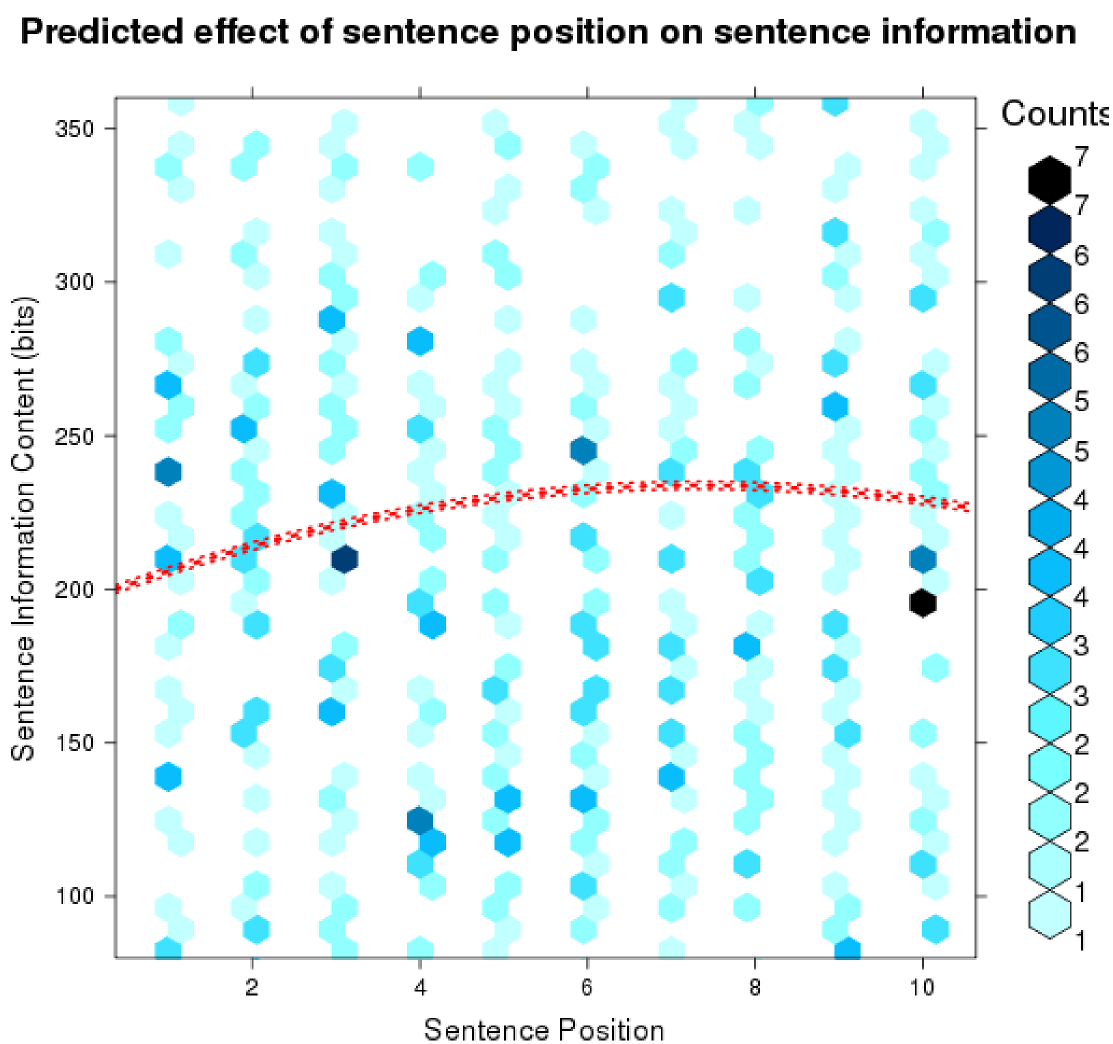
- **Data:** We selected a total of 674 segments (each segment is a complete news story or news report), all of which have at least more than 10 sentences
- Only the first 10 sentences of each segment are considered here → 6,740 sentences

The model

- The information content of each word was estimated via a trigram model trained on these 674 segments (training 607; testing 67).

- Using a linear mixed model, sentence position is regressed against word information content, while controlling for
  - sentence length (linear and non-linear)
  - out-of-vocabulary words
  - text effects as random effects

Per-sentence Information throughout discourses of Chinese speech



Summary

- Even though the specific mechanism responsible for this effect are yet to be understood, the distribution of information across discourses follows the prediction of UID
  - English writing (Genzel & Charniak, 2002)
  - Russian and Spanish writing (Genzel & Charniak, 2003)
  - English speech (Piantadosi et al., 2008)
  - Chinese speech and writing (Qian & Jaeger, submitted, in progress)

Evidence for UID

Discourse level:	ok [Genzel & Charniak, 2002, 2003; Keller, 2004; Piantadosi, 2008; Ting & Jaeger, 2008]
Utterance level:	ok? [Gomez Gallo et al., 2008a,b; in progress]
Phrasal level:	ok? [Resnik 1996; work in progress]
Word level:	ok [Jaeger, 2006, in progress; Levy & Jaeger, 2006; Wasow et al., in press]
Morphological level:	ok [Frank & Jaeger, 2008]
Phonological level:	ok [Bell et al., 2003]
Phonetic level:	ok [Aylett & Turk, 2004; Bell et al. 2003; von Son & van Santen, 2005]

High Information Environments (Producing Dispreferred Structure)

Wagner Cook*, Jaeger+, & Tanenhaus+ (2008-CUNY, submitted)

*Psychology, University of Iowa
+ Brain and Cognitive Sciences, University of Rochester

Production of dispreferred structure

- What happens when speakers choose a dispreferred structure (i.e. an unexpected and hence high information content structure)?

→**UID predictions:** Speakers should have 'repair strategies' to distribute information more uniformly when entering an unexpected parse/interpretation
  - Disfluencies (Shriberg 1996)
  - Gestures (Gerwing & Bavelas, 2005; Rauscher, Krauss & Chen, 1996; Goldin-Meadow et al., 2001, Wagner et al., 2003)

Method

- Controlled elicitation of spontaneous descriptions using 3D animated videos

PO *Simon is handing [the backpack] to [the girl].*
DO *Simon is handing [the girl] [the backpack].*

- Use distribution of structures in *those descriptions* to assess which structure (given the properties of the message) is preferred (a.k.a. probable → low in information content)

Method

Design

- Use contrasts (→ modification → length of arguments) and verb bias to create items with widely differing associated structural preferences.

Verb bias

Less PP ← → More PP

Contrast

Less PP ↕ More PP

	Show	Give	Offer	Throw	Hand	Take
theme						
both, neither						
recipient						

Information density of structures

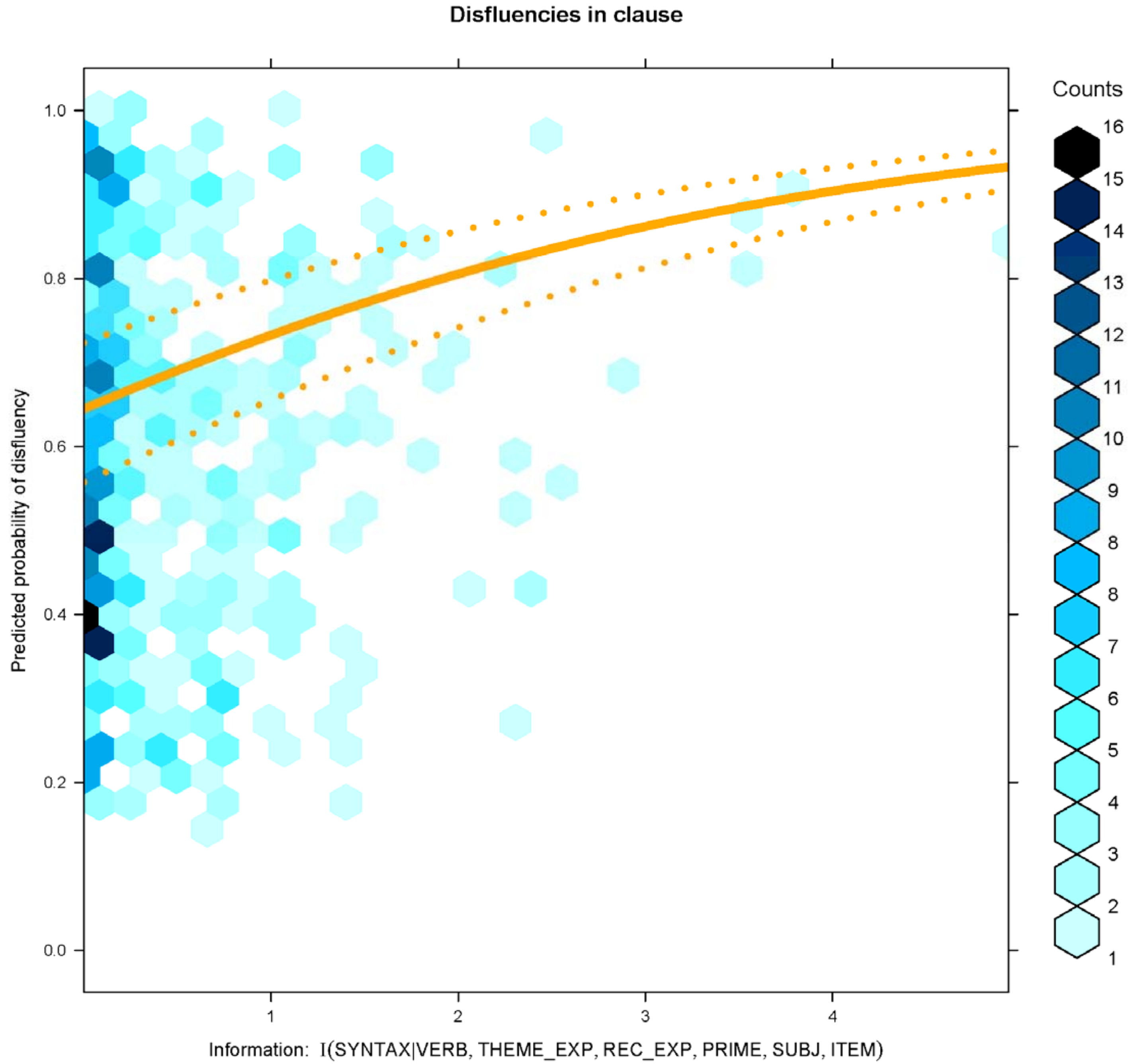
- Multilevel logit model to predict the **probability of a DO vs. PO structure**, based on
  - Recipient NP: pronominality & modification (length)
  - Theme NP: pronominality & modification (length)
  - Syntactic prime
  - Verb bias
  - Speaker and video (as random effects)

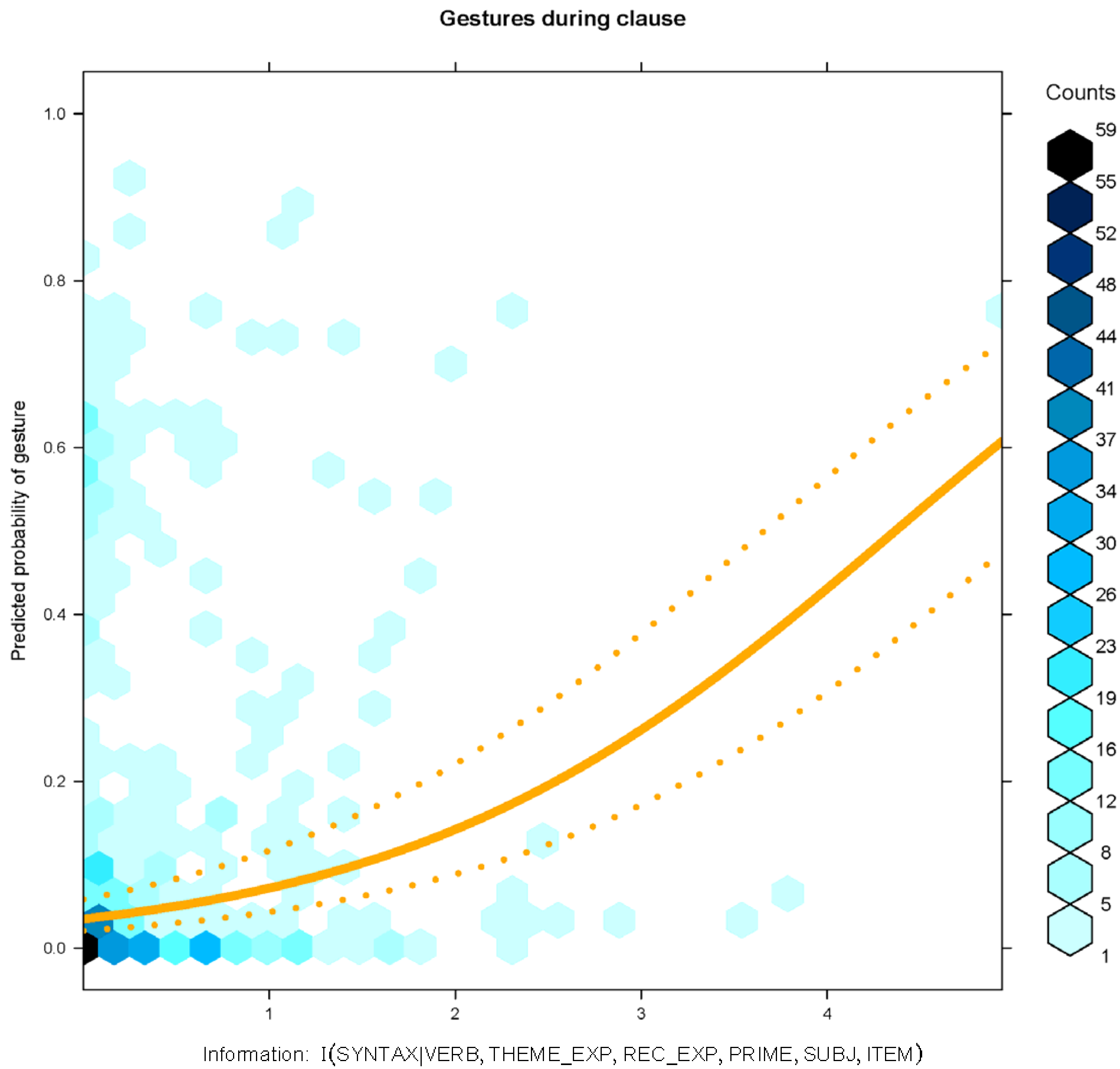
- $I(\textit{structure}) = -\log p_{\text{model}}(\textit{structure})$

Analyses

- Multilevel logit analysis to analyze when speakers' produce **disfluent** utterances and when speakers produce **gestures** depending on information density of structure.

- Simultaneously controlling for:
  - *Length* of sentence
  - Random effects for individual differences between subjects and items


Disfluencies in clause
Predicted probability of disfluency
0.0
0.2
0.4
0.6
0.8
1.0
1
2
3
4
Information: I(SYNTAX|VERB, THEME_EXP, REC_EXP, PRIME, SUBJ, ITEM)
Counts
16
15
14
13
12
11
10
9
8
8
7
6
5
4
3
2
1


Gestures during clause
Predicted probability of gesture
0.0
0.2
0.4
0.6
0.8
1.0
1
2
3
4
Information: I(SYNTAX|VERB, THEME_EXP, REC_EXP, PRIME, SUBJ, ITEM)
Counts
59
55
52
48
44
41
37
34
30
26
23
19
16
12
8
5
1

Summary

- Speakers uttering a dispreferred structure produce …
  - … more disfluencies (see also Tily et al., 2007-CUNY)
  - … more gestures

→Speakers use several channels (gesture and speech) and stretch out the high information via disfluencies.

→Evidence for UID even in those cases when speakers apparently make a suboptimal choice.

UID and Mechanistic accounts

UID and availability accounts

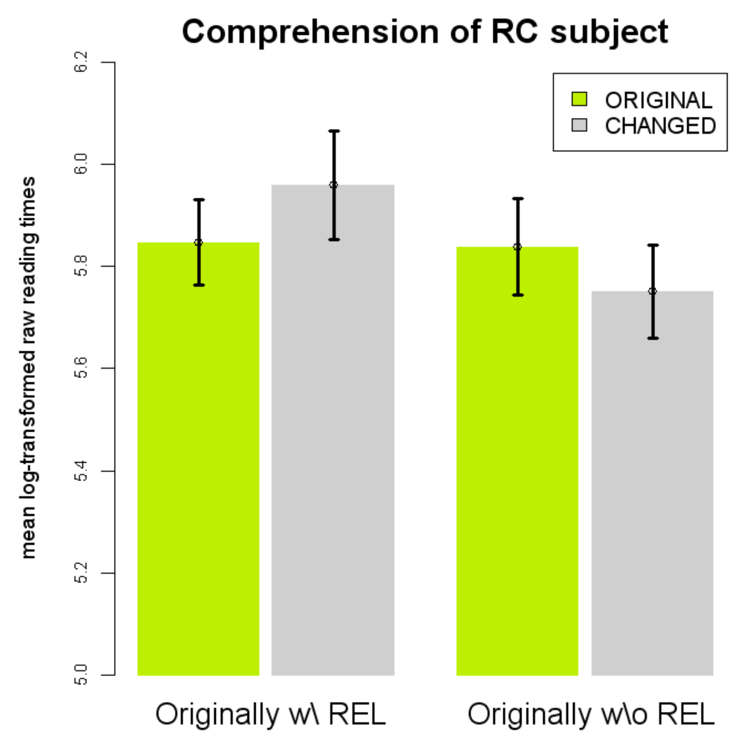
- Found evidence for both, but UID effects are much stronger in all studies we conducted.
- Some, but not all, of UID results compatible with various specific availability accounts, but additional assumptions necessary.
- In any case, availability needs to incorporate predictability/information content (rather than traditional accessibility: animacy, givenness, concreteness…)

UID and Audience design

- If UID holds at the channel between speaker and audience, it *can*, but does not have to, be seen as an instantiation of audience design
- … where information density (distribution of redundancy/uncertainty over time/linguistic units) is what speakers aim to optimize (and not ambiguity avoidance)
- This also leads to questions regarding whether speakers estimate information density from their audience’s perspective

cnt'd

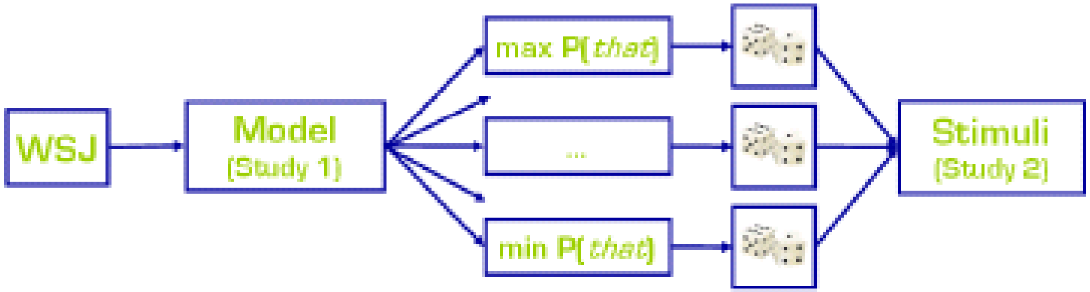
- UID also provably optimal in terms of minimizing processing difficulty (if difficulty depend super-linearly on surprisal; Levy & Jaeger, 2007)
  - To be minimized:

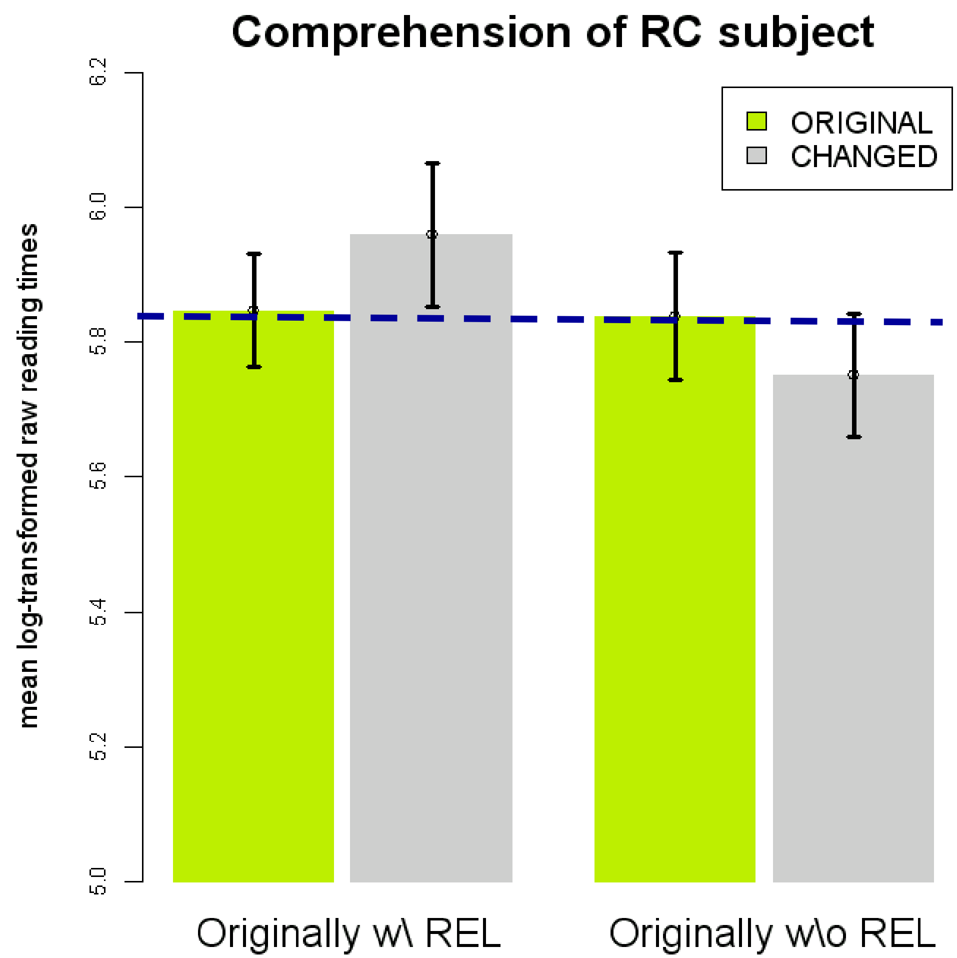


$$diff(w_i) \propto [-\log P(w_i|w_{1 \ldots i-1})]^k$$

Information Density and Syntactic Reduction: Production and Comprehension

T. Florian Jaeger (fjaeger@bcs.rochester.edu)
Department of Brain and Cognitive Science, University of Rochester, Meliora Hall
Rochester, NY 14627 USA





What mechanism is UID?

- Internal monitoring loop? (cf. Levelt, 1989)
  - Prediction: UID defects when internal monitoring defect?

- Conventionalized preference? (cf. learning, skill maintenance accounts)
  - Prediction: no non-linguistically conditioned probabilities (information content) should affect speakers' choice. ← seems unlikely.

- Training?

UID and training accounts

- These may indeed be closely related, but there aren't (m)any well-developed training accounts for the type of phenomena we've been looking at yet.
- Also, not clear how results would account for CER results (Qian & Jaeger, 2009).
- UID provides a computational derivation of why highly trained sequences may be reduced.

Summary

- Evidence that language production is efficient: given a choice, speakers prefer variants that allow them to distribute information uniformly (UID)
- Information density drives speakers' decisions at possibly all levels of linguistic production
- Our ability to use language includes 'access' to probability distribution (information density is defined through probabilities)