That-Omission Beyond Processing:
Stylistic and Social Effects

T. Florian Jaeger & Laura Staum
Stanford University

NWAV 34

October 21, 2005

The question

- Does morphosyntactic variation show effects of social factors?

Subquestion:

- When morphosyntactic variation shows effects that seem related to formality, can we call these effects stylistic, or would they better be ascribed to register?

Style vs. Register

- How do we (as a community of variationists) use these terms?

Style

Register

Style vs. Register

- **Style** - describes variation along the axis of formality; has been extended to describe variation based on any kind of social meaning (including formality and beyond). (Labov, 1982; Eckert and Rickford, 2001)

- **Register** - describes variation based on text type and genre. Although social meaning can arise from register borrowings, register variation itself is not based on social meaning. (Biber, 1994; Zwicky and Zwicky, 1982)

Style vs. Register

- Things that vary along the axis of formality (casual - formal) are often described as varying stylistically
- Things that are used more often in writing are often interpreted as more formal and may be expected to participate in stylistic variation
- When we observe a modality effect, we can choose to interpret this as stylistic or register-driven (but we should make this choice on a principled basis!).

The variable

Optional *that*-omission in complement clauses (CCs) and relative clauses (RCs):

- I believe [$_{\text{Complement Clause}}$ (*that*) we've pretty much summed everything up].

- I mean everything$_i$ [$_{\text{Relative Clause}}$ (*that*) you spray __$_i$, you know, out in the field].

(RC data does not include wh-relative pronouns, following Tagliamonte et al., 2005.)

Claims about modality/style

- Opposite **modality** findings for CCs/RCs:
  - Complement clauses show **less** *that*-omission in writing than in speech (Biber et al., 1997; Huddleston and Pullum, 2002; Bolinger, 1972)
  - Relative clauses show **more** *that*-omission in writing than in speech (Jaeger & Wasow, 2005)

- BUT, CCs and RCs show no **stylistic** effects in spoken, Labov-style interview data (Cofer, 1972)

Style vs. Register

- If *that* frequency differs between writing and speech, how can we find out whether or not it is socially meaningful (and concomitantly whether this modality difference represents stylistic variation or register-driven variation)?

Claims about social effects

- Null effects of class and ethnicity found in community study of CC and RC *that*-omission in Philadelphia (Cofer, 1972)
- No social effects found for RC *that*-omission in New Zealand (Sigley, 1997, 1998)

BUT

- 'Standard English' letter-writers use more *that* in CCs and RCs than 'Vulgar English' writers (Fries, 1940)
- Apparent socioeconomic class effects found for zero relatives, also in Philadelphia (Adamson, 1992)

Implications

- Style presupposes the social (Labov, 1982; Bell,1984)
- Typical sociolinguistic variables show related patterns for stylistic and social effects (Labov, 1966)

↳ If the observed effects are stylistic, *that*-omission should show social stratification.

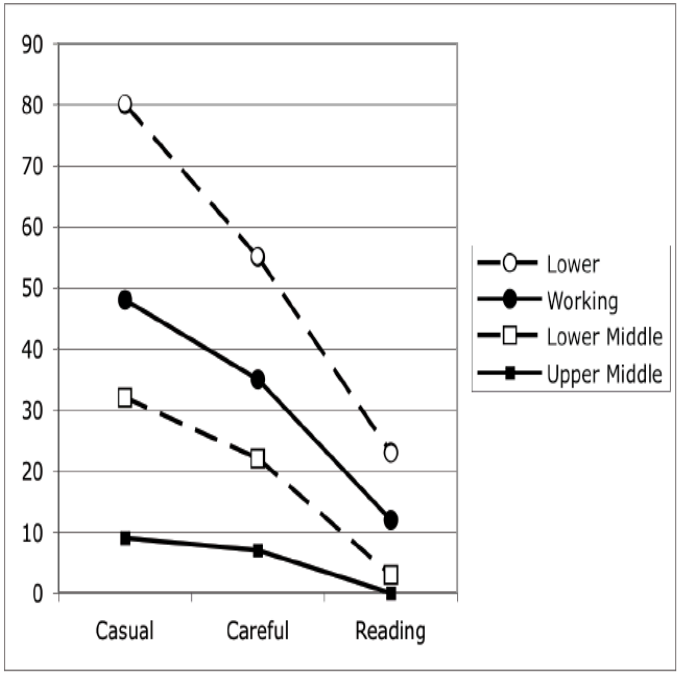


Figure 1. % Reduced –ing. from (Labov, 1966) in three styles and four socioeconomic strata.

(From Eckert's 2005 LSA address)

Two Hypotheses

Keeping the proposed distinction between style and register in mind, we can imagine two competing hypotheses:

Stylistic Omission Hypothesis

That-omission is a socially meaningful variable and thus shows both stylistic and social effects.

Register-driven Omission Hypothesis

That-omission is not socially meaningful, and thus should not show social stratification.


LELAND STANFORD JUNIOR UNIVERSITY
DIE LUFT DER FREIHEIT WEHT
1891

The Study

Goals

- Compare the **Stylistic Omission** and the **Register-based Omission Hypothesis.**
  - Does *that*-omission show social stratification in spoken American English?
  - Also: Is omission in CCs and RCs affected by the same factors?

Methods

- Use a large corpus of speech coded for social information

- Use modern statistical modeling that can control for both linguistic factors and speaker effects

The database: Penn Treebank III Switchboard

- About 800,000 words of parsed and POS-tagged telephone dialogues between strangers about pre-selected topics. (Godfrey et al., 1992)

- Sample: CCs and RCs that can exhibit the variation
  - 6,712 CCs (only verbal complements included)
  - 3,465 RCs (no *wh*-relativizers)

- Distribution of social variables in our sample reflects distribution in entire Switchboard corpus (i.e. all kinds of speakers use CCs and RCs)

Modeling problem

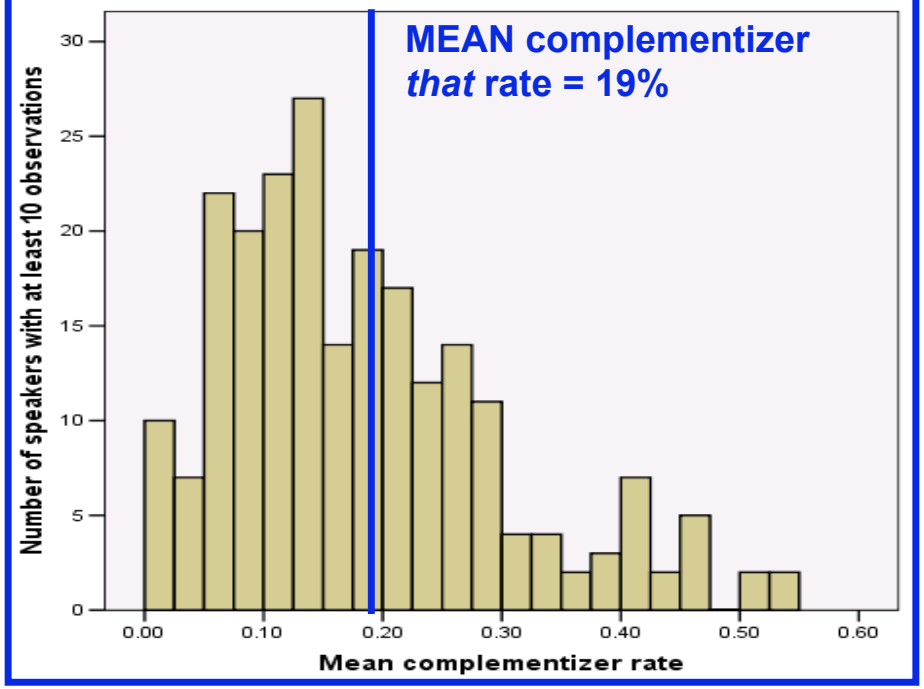
Two common methods of modeling sociolx. variation:

- Case-by-case (all cases of the variant are included)
  - Pro: can include linguistic factors in the model
  - Con: each speaker's information appears multiple times; violates assumption of independence of observations!
- Speaker index (calculate an index of the level of behavior observed for each speaker)
  - Pro: each observation is actually independent
  - Con: can't include linguistic factors in the model

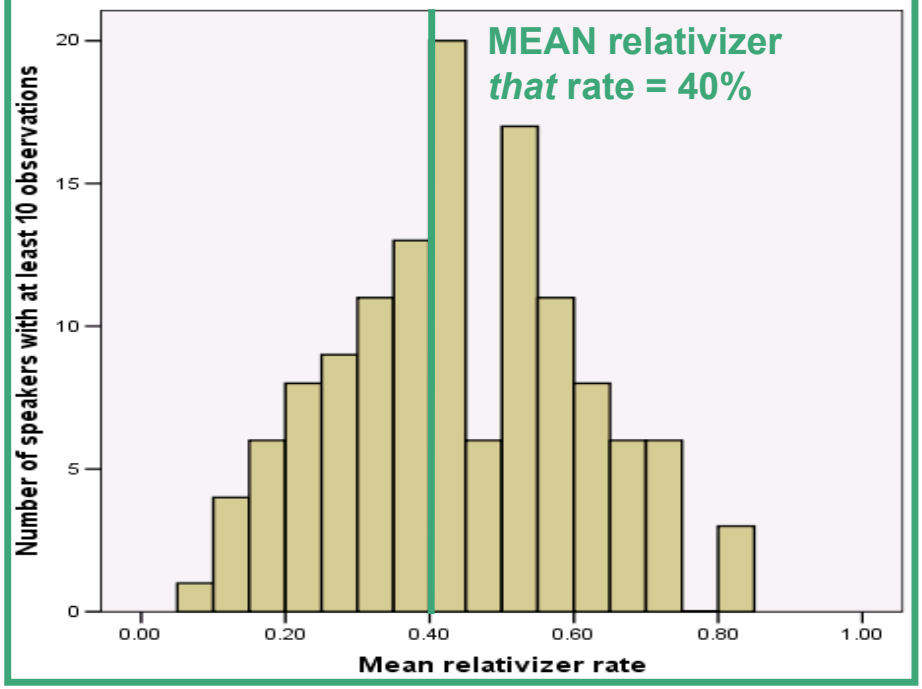
Inter-speaker variation in our samples

- Different speakers have different rates of *that* in CCs and RCs:

- CCs: 350 speakers
  - approx. 19 utterances each (STDEV= 16.3, Range= 1 to 96)



- RCs: 335 speakers
  - approx. 9.5 utterances each (STDEV= 8.0, Range= 1 to 40)



The statistical model

- Logit Generalized Linear Mixed Model
  (R-library *glmmPQL*, cf. Venables & Ripley, 2002)
  - These models provide a way to include both social and processing/linguistic factors in the analysis without incorrectly inflating the social effects (unlike current implementations of VARBRUL).
  - Also deals with individual variation in an adequate way (w/o introducing lots and lots of free parameters).

Predictors in the model

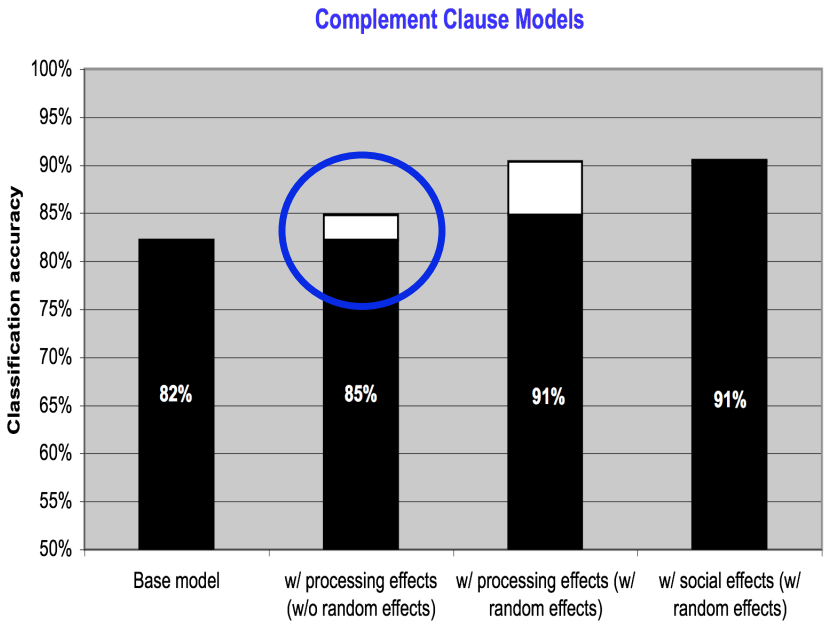
- Processing/linguistic factors (for details see appendix of handout)
- Social factors
  - Gender (2 levels)
  - Education (NO DEGREE; HIGH SCHOOL; COLLEGE; > COLLEGE)
  - Age (mean=37, SD=10.5, range=16-68)
  - Dialect (7 regions + MIXED)

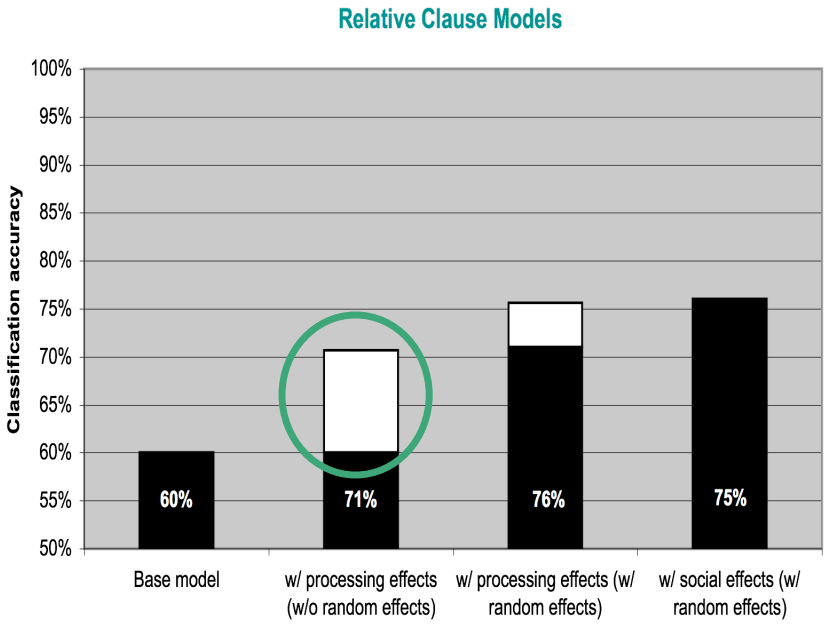
Results: Overview

- CCs
- RCs

Model accuracy

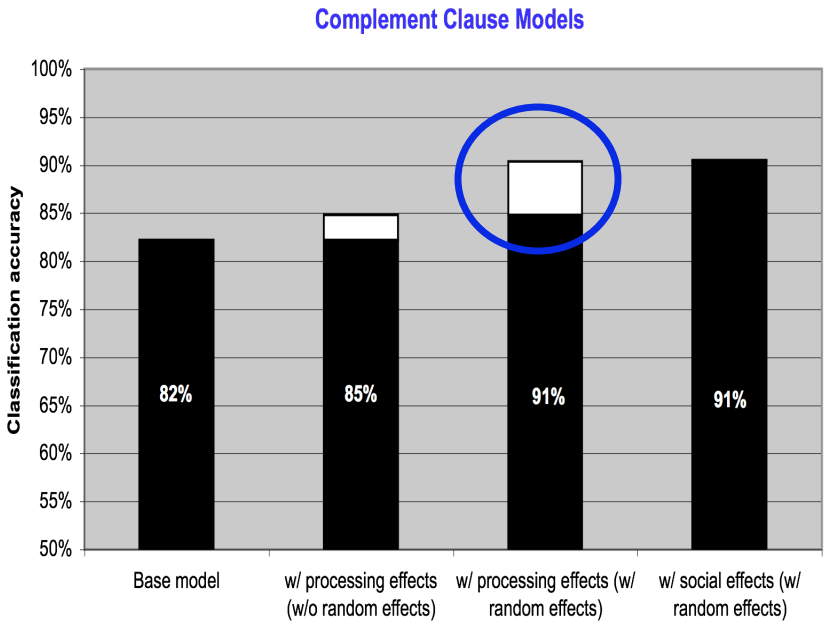
- Processing factors account for a lot of the variation in both complementizer *that* omission and relativizer *that* omission.

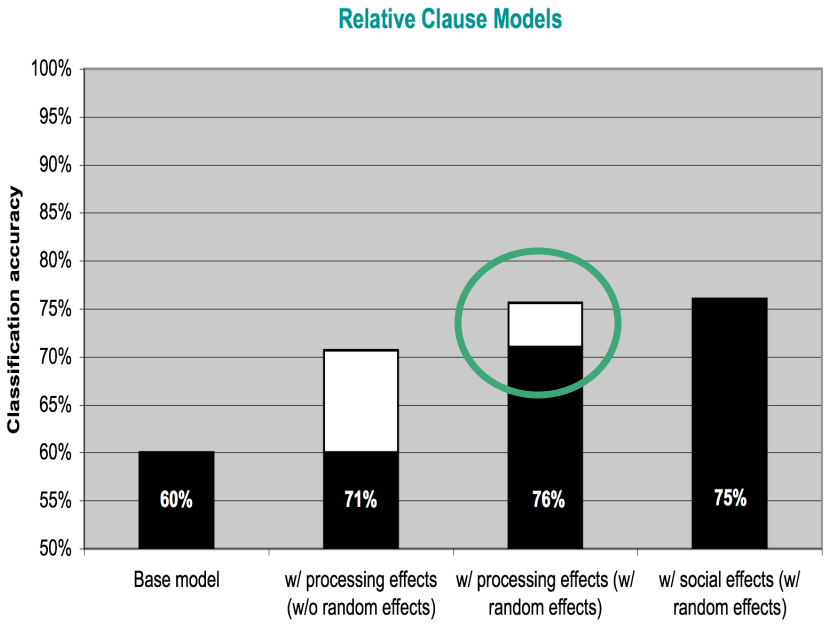




Model accuracy

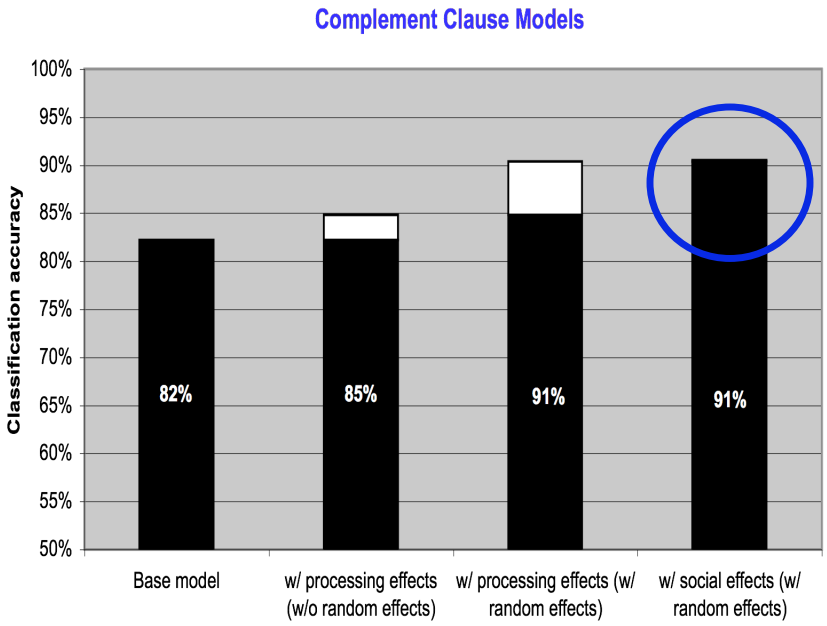
- Accounting for individual speaker effects improves the model significantly.

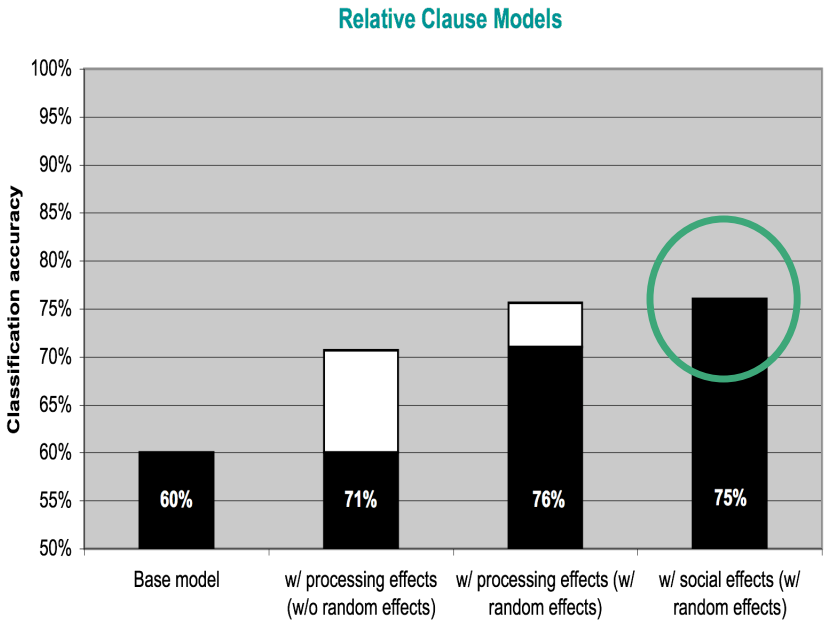




Model accuracy

- Social factors don't matter much.





Results: Social effects

- CCs
- RCs

CCs: Social factors

Factor	Significance
Speaker's gender	n.s.
Speaker's education	
• HIGH SCHOOL more *that* than NO DEGREE	p = 0.07
• COLLEGE more *that* than HIGH SCHOOL	n.s.
• > COLLEGE more *that* than COLLEGE	n.s.
Speaker's primary dialect	n.s.
Speaker's age	n.s.

⇨ Education effect is based solely on the *NO DEGREE* category

Summary of results

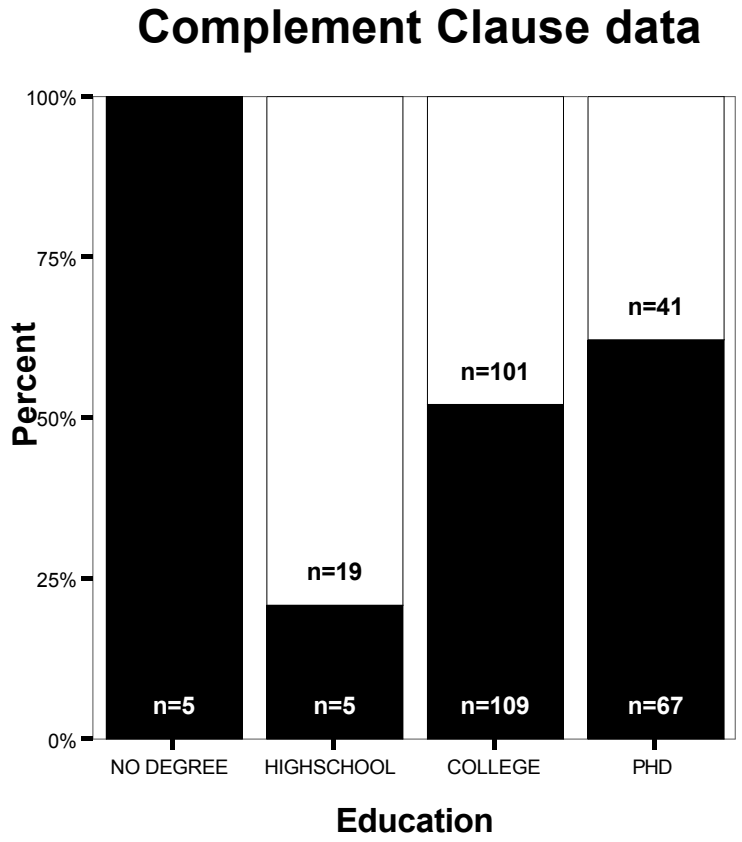
- There are some non-significant dialectal contrasts in the model, **but** ...
  - ... they are weak (by contrast, most processing factors are associated with p-values << 0.0001)
  - ... they don't form a clear interpretable pattern.

- So what about the effect of education?
  - Goes in the direction expected by the Stylistic Omission Hypothesis
  - But why is only the *NO DEGREE* level relevant?
  - There is evidence that this level (unlike the others) is unreliable:

NO DEGREE

- Extremely small category in Switchboard (only 13 speakers)
- In our CC sample, there were 5 speakers in the *NO DEGREE* category.

- Additionally, *all* of them are men, which makes it hard to distinguish the effect from a gender effect.
- So, is SWBD education coding unreliable?

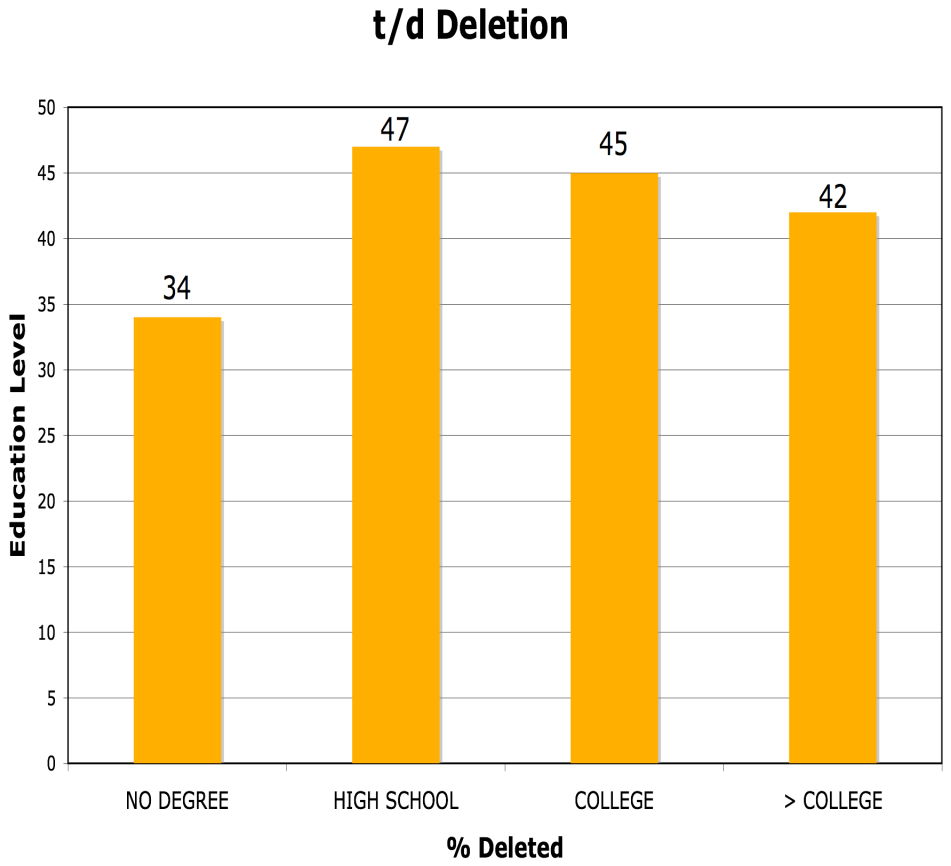


Other sociolx. work with SWBD

- In SWBD, t/d-deletion is distributed as expected (except for NO DEGREE) (Strassel, 2001)

⇨ SWBD education-coding is fine-grained and accurate enough to see real effects.

⇨ **NO DEGREE category is probably unreliable.**



DASL project results from TIMIT and SWBD for t/d deletion:
http://www.ldc.upenn.edu/Projects/DASL/

Intermediate conclusions - CCs

- Given that the *only* social effect we observed comes from an unreliable category (education = NO DEGREE), we conclude that there is no evidence for the **Stylistic Omission Hypothesis** (repeated below):

Stylistic Omission Hypothesis

That-omission is a socially meaningful variable and thus shows both stylistic and social effects.

RCs: Social factors

Factor	Significance
Speaker's gender	p = 0.01
Speaker's education	n.s.
Speaker's primary dialect	n.s.
Speaker's age	n.s.

- Men use less *that* than women do in relative clauses.

Intermediate conclusions - RCs

- The picture here looks different from the CC findings, because gender has a significant effect on the variation for RCs.
  - N.B. This gender effect is several orders of magnitude smaller than the effects we see for processing factors.

- However, we still don't see the kind of social stratification in the education or age factors that we would need to support the **Stylistic Omission Hypothesis**.

Discussion

- Overall the results show insufficient/little evidence of traditional social meaning for *that*-omission.
- Consistent with Tagliamonte et al., 2005; Sigley, 1997, 1998; and Cofer, 1972.
- Appears to conflict with Fries, 1940 and Adamson, 1992.

Discussion

- Fries, 1940
  - based his designations of 'Standard English' and 'Vulgar English' writers on other linguistic features in the texts
  - This is fully consistent with the Register-driven Omission Hypothesis
- Adamson, 1992
  - Studied production of zero relatives
  - His non-zero category includes both *that* and *wh*-relative pronouns - we expect this to show social stratification (Tagliamonte et al., 2005)

Conclusions

- Expected social patterns for stylistically conditioned variation do not appear for *that*-omission.

- Modeling variation requires controlling properly for both processing/linguistic factors and speaker effects, and modern statistical models provide a way to do so.

  (Tagliamonte et al., 2005; Weiner and Labov, 1983)

  - The models we derived (w/ speaker effect modeling) perform significantly better than standard logistic regression models.

- If we refer to the modality effects we see for *that*-omission as stylistic effects, we lose the relationship between stylistic and social effects, suggesting that in this case modality effects should fall under the rubric of register variation.

Possible Further Work

- Gender effect on RCs (what's going on?)
- Complementizer variation in other languages (Danish, Swedish)
- Complementizer Reduction vs. deletion project
  - Reduction in general shows social effects, *that*-reduction shows gender effects (Bell et al., 2003)
  - If women both reduce more and delete more RC *that*, why doesn't this extend to CC *that*?

Acknowledgments

Special thanks to:

Elizabeth Coppock

Penelope Eckert

Rafe Kinsey

Roger Levy

Christopher Manning

Robert Sigley

Tom Wasow

Arnold Zwicky

Thank you!

Please feel free to contact us:

lstaum@stanford.edu (Laura Staum)
tiflo@stanford.edu (Florian Jaeger)

Linguistics Department
Margaret Jacks Hall
Stanford, CA 94305


LELAND STANFORD JUNIOR UNIVERSITY
DIE LUFT DER FREIHEIT WEHT
1891

Appendix

Construction of databases and Exclusion

- CCs:
  - CC is Complement of verb (rather than adjective or noun)
  - CC immediately adjacent to verb
  - CC is not coordinated with other CC
  - Complementizer is either *that* or *zero* [6,912]

- RCs:
  - Extracted element is not pied-piped
  - Extracted element is not subject of RC [4,406]
  - Relativizer is either *that* or *zero* (no *wh*-relativizers) [3,701]
  - Hand-labeling determined
    - Case is relative clause (judgment) [3,619]
    - Case can undergo variation (conservative judgment) [3,465]

Social characteristics of speakers in the samples

	CC	RC	ALL
NEW ENGLAND	5%	5%	4%
NYC	5%	5%	6%
NORTHERN	16%	15%	14%
SOUTHERN	11%	10%	11%
NORTH MIDL.	13%	13%	15%
SOUTH MIDL.	31%	31%	29%
WESTERN	14%	15%	16%
MIXED	5%	5%	5%
unknown	<1%	<1%	<1%

	CC	RC	ALL
NO DEGREE	1%	1%	3%
HIGH SCHOOL	7%	7%	7%
COLLEGE	60%	60%	57%
PHD	31%	32%	33%
unknown	<1%	<1%	<1%

	CC	RC	ALL
MALE	54%	54%	55%
FEMALE	46%	46%	45%

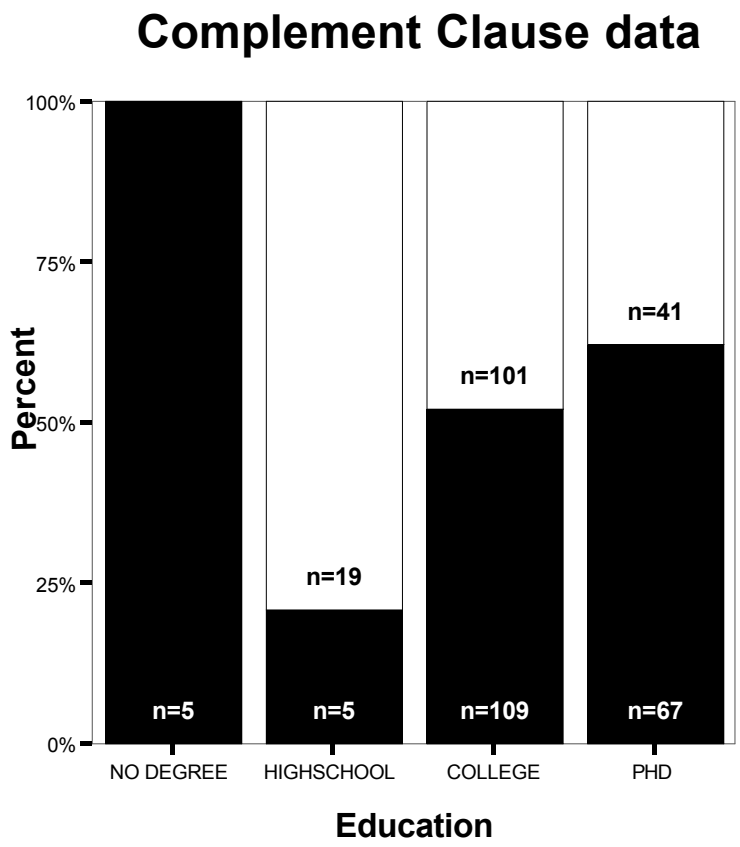
Data sparsity?

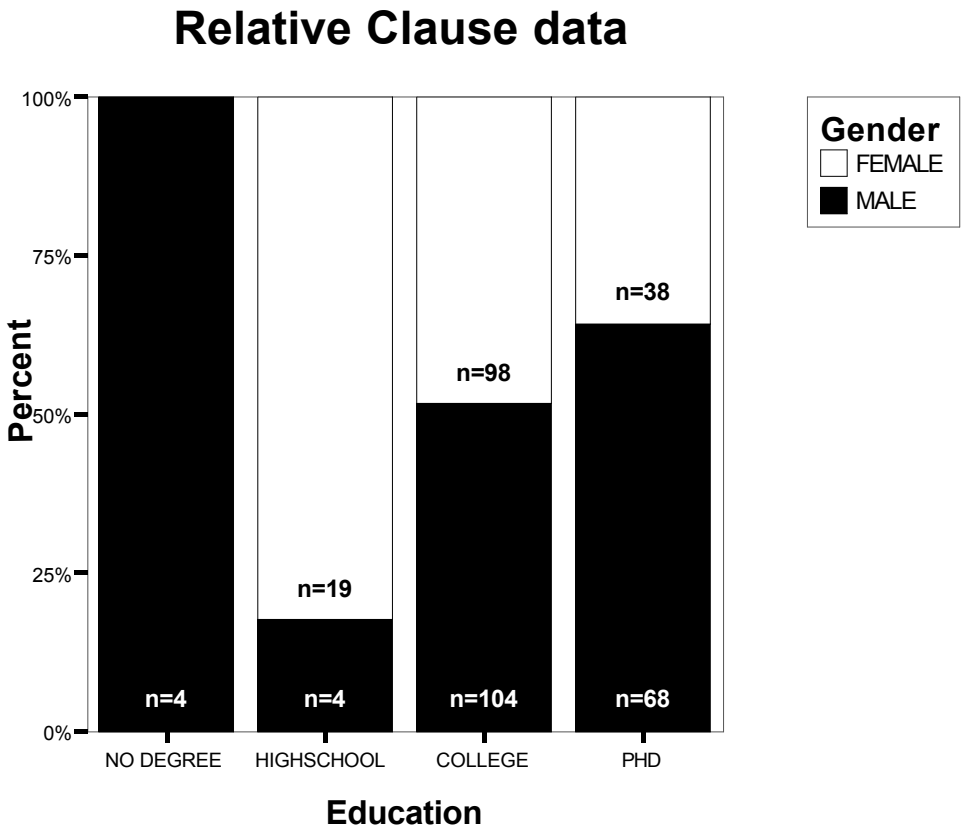
- Did we not get an effect because of data sparsity?
- Generally: *no*, since there should be enough data to fit up to approximately
  - 170 free parameters for the RC data
  - 120 free parameters for the CC data
  - We used far fewer free parameters
- But, yes, *some* social variables are distributed unevenly (e.g. there almost no data on education = '*no high school*')

The statistical model

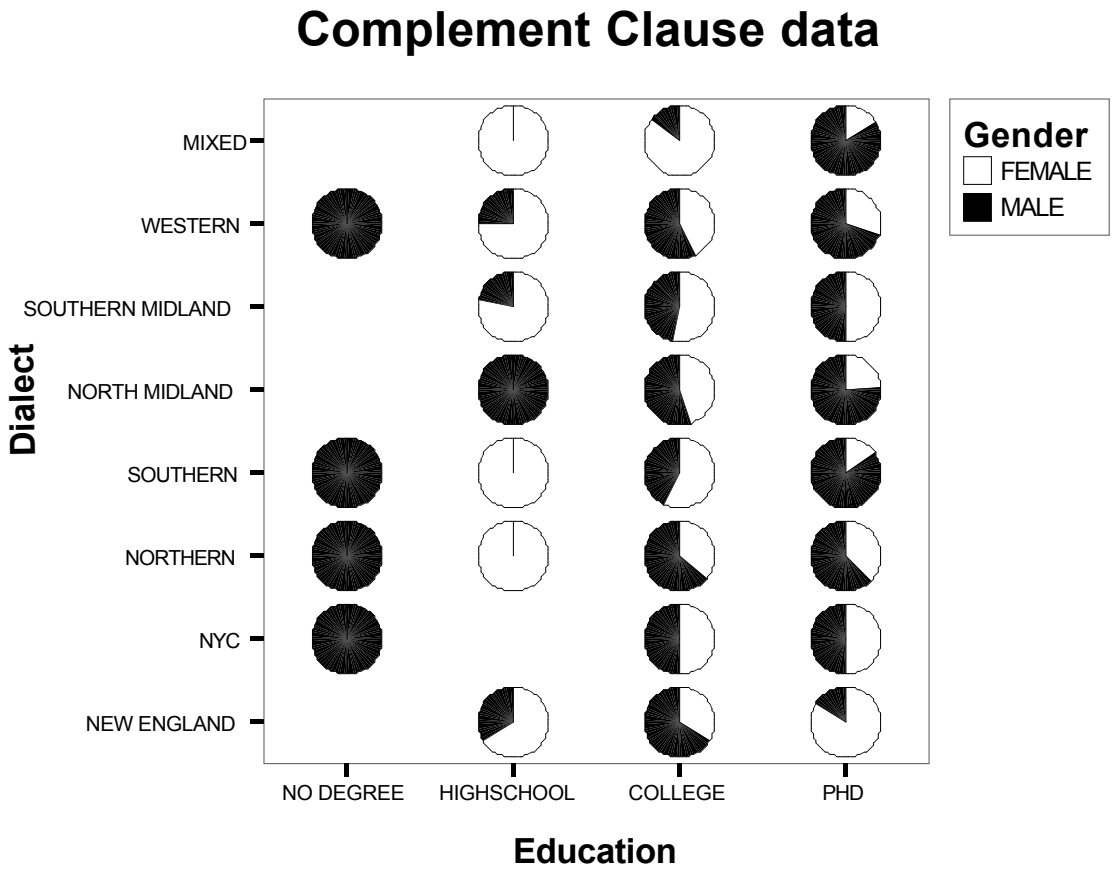
- Logit Generalized Linear Mixed Model (R-library *glmmPQL*, cf. Venables & Ripley, 2002).
- We used normally distributed random intercepts to model speaker effects (to avoid violations of the assumption of the independence of observations) = *better way to model speaker effects*.
- Additionally the models contain:
  - Processing/linguistic factors (as within-speaker factors)
  - Social factors (as between-speaker factors)

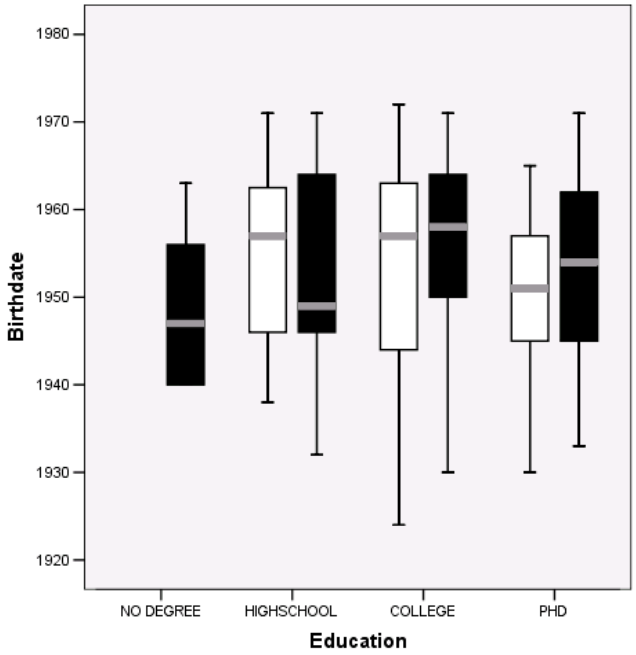
Dependencies among social variables:
Education & Gender





Dependencies among social variables: more





CCs: processing/linguistic factors

Factor	Effect
Matrix clause: subject	+complexity → +P(*that*)
Matrix clause: negation?	yes → +P(*that*)
Matrix clause: embedded?	yes → +P(*that*)
CC: Predictability of CC	+predictability → -P(*that*)
CC: Complexity of subject	+complexity → +P(*that*)
CC: Length of CC	+length → +P(*that*)
CC: disfluency present?	yes → +P(*that*)

Ferreira & Dell, 2000; Roland, Elman, & Ferreira (2005); Ferreira (2003); Ferreira et al. (2005)

RCs: processing/linguistic factors

Factor	Effect
Matrix clause: negation?	yes → +P(*that*)
Matrix clause: embedded?	yes → +P(*that*)
Matrix clause: verb type?	constructional variation
Modified NP: RC-favoring type of determiner?	yes → -P(*that*)
Modified NP: uniqueness requiring adjective	yes → -P(*that*)
Modified NP: type of head noun (semantic weight)	+weight → +P(*that*)
Modified NP: GF in matrix clause	constructional variation
RC: GF of extracted head (ADV, OBJ)	OBJ → +P(*that*)
RC: RC adjacent to head noun?	yes → -P(*that*)
RC: Complexity of RC subject	+complexity → +P(*that*)
RC: Length of RC	+length → +P(*that*)

Fox & Thompson (in press); Jaeger, Levy, Wasow & Orr (2005); Jaeger & Wasow (2005a,b); Jaeger, Orr, & Wasow (2005); Race & MacDonald (2003); Quirk (1957)