Published in final edited form as:
Cogn Psychol. 2010 August ; 61(1): 23–62. doi:10.1016/j.cogpsych.2010.02.002.

Redundancy and reduction: Speakers manage syntactic information density

T. Florian Jaeger
Department of Brain and Cognitive Sciences, Department of Computer Science, University of Rochester, Meliora Hall, Box 270268, Rochester, NY 14627-0268, United States

Abstract

A principle of efficient language production based on information theoretic considerations is proposed: Uniform Information Density predicts that language production is affected by a preference to distribute information uniformly across the linguistic signal. This prediction is tested against data from syntactic reduction. A single multilevel logit model analysis of naturally distributed data from a corpus of spontaneous speech is used to assess the effect of information density on complementizer *that*-mentioning, while simultaneously evaluating the predictions of several influential alternative accounts: availability, ambiguity avoidance, and dependency processing accounts. Information density emerges as an important predictor of speakers' preferences during production. As information is defined in terms of probabilities, it follows that production is probability-sensitive, in that speakers' preferences are affected by the contextual probability of syntactic structures. The merits of a corpus-based approach to the study of language production are discussed as well.

Keywords

Efficient language production; Rational cognition; Syntactic production; Syntactic reduction; Complementizer *that*-mentioning

1. Introduction

The extent to which language and language use are organized to be efficient has attracted researchers from various disciplines for at least close to a century (Aylett & Turk, 2004; Chomsky, 2005; Fenk-Oczlon, 2001; Genzel & Charniak, 2002; Givón, 1979; Hawkins, 2004; Landau, 1969; Manin, 2006; van Son, Beinum, Koopmans-van, & Pols, 1998; Zipf, 1935, 1949). Probably one of the earliest observations related to efficient language production is the link between word frequency and word form (Schuchardt, 1885; Zipf, 1929, 1935). The observation that frequent words generally have shorter linguistic forms (Zipf, 1935) was an important piece of evidence that led Zipf to propose his famous *Principle of Least Effort*, according to which human behavior is affected by a preference to minimize "the person's average rate of work-expenditure over time" (Zipf, 1949, p. 6). In this context, it is intuitively efficient for more frequent words to have shorter phonological forms. More recent evidence suggests that word length (in phonemes or syllables) is even more strongly correlated with words' average predictability in context than with their frequency (Piantadosi, Tily, & Gibson, 2009; see also Manin, 2006). This inverse relation between contextual probability and linguistic form is expected given information theoretic considerations about efficient communication (Shannon, 1948, for more detail see below): the more probable (expected) a word is in its

context, the less information it carries (the more redundant it is) in that context. The observed link between probability and phonological form can then be restated in terms of information: on average, words that add more (new) information to their context have longer phonological forms. Intriguingly, this link between information, redundancy, and probability on the one hand and linguistic form on the other hand is not limited to the mental lexicon, but seems to extend to lexical production. Several studies over recent years have found that more predictable instances of the same word are on average produced with shorter duration and with less phonological and phonetic detail (Aylett & Turk, 2004, 2006; Bell et al., 2003, Bell, Brenier, Gregory, Girand, & Jurafsky, 2009; Pluymaekers, Ernestus, & Baayen, 2005; van Son et al., 1998; van Son & van Santen, 2005 among others).

In short, the distribution of phonological forms in the mental lexicon as well as evidence from phonetic and phonological reduction during online production suggest that language strikes an efficient balance between the information conveyed by linguistic units and the amount of signal associated with them (cf. Aylett & Turk, 2004). This raises an intriguing possibility. Human language production could be organized to be efficient at *all* levels of linguistic processing in that speakers prefer to trade off redundancy and reduction. Put differently, speakers may be managing the amount of information per amount of linguistic signal (henceforth information density), so as to avoid peaks and troughs in information density. If so, it should be possible to observe effects of this trade-off on speakers' preferences at choice points during utterance planning.

Choice points that would theoretically allow speakers to manage information density are ubiquitous even beyond phonetic and phonological planning. To name just a few: during morphosyntactic production, speakers of many languages can sometimes choose between full or contracted forms (e.g. in auxiliary contraction, as in *he is* vs. *he's*, Frank & Jaeger, 2008); during syntactic production, speakers sometimes have a choice between full and reduced constituents (e.g. in optional *that*-mentioning, as in *This is the friend (that) I told you about*, (Ferreira & Dell, 2000; Race & MacDonald, 2003); optional *to*-mentioning, as in *It helps you (to) focus where your money goes*, Rohdenburg, 2004); speakers often can even elide entire constituents (e.g. optional argument and adjunct omission, as in *I already ate* (*dinner*), Brown & Dell, 1987; Resnik, 1996); and at the earliest stages of production planning, speakers can choose to distribute their intended message over one or more clauses (e.g. *Ok, next move the triangle over there vs. Ok, next take the triangle and move it over there*, Brown & Dell, 1987; Gómez Gallo, Jaeger, & Smyth, 2008; Levelt & Maassen, 1981). Some of these choice points are arguably available during any sentence and similar choice points are available in other languages. If language production is organized to be efficient in that speakers prefer to distribute information uniformly across the linguistic signal, the form with less linguistic signal should be less preferred whenever the reducible unit encodes a lot of information.

Unfortunately, the effect of information density on production beyond the lexical level has remained almost entirely unexplored (but see Genzel & Charniak, 2002; Resnik, 1996; discussed below). This is despite a very rich tradition of research on speakers' preferences during syntactic production (e.g. work on accessibility effects, e.g. Bock & Warren, 1985; Ferreira, 1994; Ferreira & Dell, 2000; Prat-Sala & Branigan, 2000; dependency length minimization, e.g. Elsness, 1984; Hawkins, 1994, 2001, 2004; syntactic priming, e.g. Bock, 1986; Pickering & Ferreira, 2008).

In this article, I explore the hypothesis that language production at all levels of linguistic representation is organized to be communicatively efficient. I present and discuss the hypothesis of *Uniform Information Density* (developed in collaboration with Roger Levy; see Jaeger, 2006a; Levy & Jaeger, 2007). The hypothesis of Uniform Information Density links speakers' preferences at choice points during incremental language production to information

theoretic theorems about efficient communication through a noisy channel with a limited bandwidth (Shannon, 1948). I test the prediction of this hypothesis that syntactic production reflects a preference to distribute information uniformly across the speech signal.

Successful transfer of information through a noisy channel with a limited bandwidth is maximized by transmitting information uniformly close to the channel's capacity (Genzel & Charniak, 2002). Information is defined information theoretically in terms of probabilities. The Shannon information of a word, I(*word*), is the logarithm-transformed inverse of its probability, $I(word)=\log\frac{1}{p(word)}=-\log p(word)$. Since in natural language the probability of a word depends on the context it occurs in, the definition of Shannon information captures that a word's information, too, is context dependent. Intuitively (and simplifying for now), efficient communication balances the risk of transmitting too much information per time (or per signal), which increases the chance of information loss or miscommunication, against the desire to convey as much information as possible with as little signal as possible. If human language use is communication through a noisy channel, linguistic communication would be optimal if (a) *on average* each word adds the same amount of information to what we already know and (b) the rate of information transfer is close to the channel capacity.[1] It seems unlikely that all aspects of language are organized so as to achieve *optimal* communication, given that language is subject to many other constraints (e.g. languages must be learnable). Still, it is possible that language production is *efficient*, in that speakers aim to communicate efficiently within the bounds defined by grammar. If so, speakers should (a) aim for a relatively uniform distribution of information across the signal wherever possible without (b) continuously under- or overutilizing the channel. The hypothesis of Uniform Information Density, which is tested in this paper, focuses on the first prediction (see also Aylett & Turk, 2004; Genzel & Charniak, 2002; Jaeger, 2006a; Levy & Jaeger, 2007).

Uniform Information Density (UID)

> Within the bounds defined by grammar, speakers prefer utterances that distribute information uniformly across the signal (information density). Where speakers have a choice between several variants to encode their message, they prefer the variant with more uniform information density (ceteris paribus).

Two aspects of the definition deserve immediate clarification. For the purpose of this article, 'information density' corresponds roughly to information per time. It is, however, important to keep in mind that the relevant notion of information density of the acoustic signal may also depend on articulatory detail (cf. earlier versions of UID in Jaeger (2006a) and Levy & Jaeger (2007), which did not take this into consideration). Second, the term 'choice' does not imply conscious decision making. It is simply used to refer to the existence of several different ways to encode the intended message into a linguistic utterance.

Given the definition of information, UID assumes that speakers have access to probability distributions over linguistic units (segments, words, syntactic structures, etc.). This distinguishes UID from most existing production accounts, which make different architectural assumptions and do not predict information density to affect speakers' preferences (e.g. availability accounts, Ferreira, 1996; Ferreira & Dell, 2000; Levelt & Maassen, 1981; alignment accounts, Bock & Warren, 1985; Ferreira, 1994; dependency processing accounts

[1] In information theory (Shannon, 1948), the channel capacity defines the maximum amount of information per transmission through a noisy channel that can be transmitted with an arbitrarily small error rate. In other words, communication at and below the channel capacity can in theory (with the right code) turn a noisy channel into an essentially noiseless channel. If a certain error rate is acceptable communications at a rate above the channel capacity are also possible. For simplicity's sake, I assume that human communication strives for an arbitrarily small error rate, but for any given acceptable error rate there is going to be a maximum rate of transmission at which this error rate can be achieved.

Hawkins, 1994, 2004). Among the accounts that share UID's architectural assumption that speakers employ probability distributions during production are connectionist accounts (Dell, Chang, & Griffin, 1999; Chang, Dell, & Bock, 2006) and work on probability-sensitive production (e.g. Aylett & Turk, 2004; Bell et al., 2003, 2009; Gahl & Garnsey, 2004; Resnik, 1996; Stallings, MacDonald, & O'Seaghdha, 1998).

Previous findings from the phonetic and phonological reduction of words in spontaneous speech lend initial support to the hypothesis of Uniform Information Density (see references above). To investigate the effect of information density on production *beyond* the lexical level, I investigate a case of syntactic reduction, optional *that*-mentioning in English complement clauses. When speakers of English produce an utterance with a complement clause, they have the option of mentioning the complementizer, as in (1a), or omitting the complementizer, as in (1b) (example taken from the Switchboard corpus, Godfrey, Holliman, & McDaniel, 1992):

(1) **a.** I know [*that* the expectation for them was, uh, to have sex …].

b. I know [the expectation for them was, uh, to have sex …].

UID predicts that the production system is set up in such a way that information density directly or indirectly affects speakers' preferences during production. That is, as speakers incrementally encode their intended message, their preferences at choice points should be affected by the relative information density of different continuations *compatible with the intended meaning*. Hence, UID does not predict that every word provides the same amount of information, but rather that, where grammar permits, speakers aim to distribute information more uniformly without exceeding the channel's capacity. Fig. 1 serves to illustrate this prediction for *that*-mentioning in complement clauses. The hypothetical distribution of information for the same complement clause with and without the complementizer *that* is shown. Intuitively, mentioning the complementizer distributes the information at the onset of the complement clause over more words (this prediction will be spelled out below). If the information density at the onset of the complement clause is so high that it would otherwise exceed the channel capacity, as in Fig. 1a, speakers are predicted to prefer the full complement clause with *that*, thereby lowering information density. If, however, the information density at the complement clause onset is low, as in Fig. 1b, speakers are predicted to prefer the reduced variant, which avoids unnecessary redundancy.

The goals of this article are twofold. The first goal is to establish UID as a computational account of efficient sentence production. I provide evidence from *that*-mentioning that syntactic production is sensitive to information density and, more generally, that syntactic production is probability-sensitive. I summarize further evidence supporting UID and discuss the relation between UID and existing algorithmic accounts of sentence production, such as availability-based production (e.g. Bock & Warren, 1985; Ferreira & Dell, 2000; Levelt & Maassen, 1981) and ambiguity avoidance accounts (e.g. Bolinger, 1972; Clark & Fox Tree, 2002).

The data in this article are sampled from a corpus of spontaneous speech. The use of such naturally distributed data avoids a serious problem inherent to the use of balanced designs in psycholinguistic experiments that, I argue, has so far been underestimated. There is considerable evidence that listeners and speakers are sensitive to probability distributions (for comprehension, Hale, 2001; Jurafsky, 1996; Kamide, Altmann, & Haywood, 2003; Levy, 2008; MacDonald, 1994; McDonald & Shillcock, 2003; Staub & Clifton, 2006; Trueswell, 1996; for production, Bell et al., 2003, 2009; Gahl & Garnsey, 2004; Stallings et al., 1998, as well as the work presented here) and that they adapt to changes in these distributions (e.g. Saffran, Johnson, Aslin, & Newport, 1999; Wells, Christiansen, Race, Acheson, & MacDonald, 2009). There is even evidence that such adaptation can take place after relatively little exposure

(e.g. Clayards, Tanenhaus, Aslin, & Jacobs, 2008). Consider also that one of the most widely used experimental paradigms in work on production, syntactic priming (Bock, 1986; Pickering & Ferreira, 2008), trades on recent exposure changing speakers' behavior. Hence, it seems paramount to develop methods that facilitate well-controlled investigations of language production without exposing speakers to unusual distributions (such as balanced and hence uniform distributions, as opposed to, for example power law distributions, cf. Zipf, 1935, 1949). The corpus-based approach taken here constitutes such a method. Modern statistical regression models are used to deal with the unbalanced data that inevitably result from natural distributions. Such corpus-based studies are still rare in work on language production and there is skepticism about the use of corpus studies as tests of psycholinguistic hypotheses. The second goal of this article is to illustrate that a corpus-based approach is not only feasible, but a desirable methodological addition to research on the cognitive psychology of language production (see also Baayen, Feldman, & Schreuder, 2006; Bresnan, Cueni, Nikitina, & Baayen, 2007; Jaeger, 2006a, submitted for publication; Roland, Elman, & Ferreira, 2005).

2. Testing Uniform Information Density against syntactic reduction in spontaneous speech

UID predicts that speakers aim to transmit information uniformly close to, but not exceeding, the channel capacity. Mentioning the complementizer *that* at the onset of a complement clause distributes the same amount of information over one more word, thereby lowering information density. Hence, everything else being the same, speakers should be more likely to produce full complement clauses (CCs with *that*) than reduced CCs (without *that*), the higher the information of the CC onset in its context. This prediction, which is not shared by alternative accounts of syntactic production (at least not in their current form), is tested against data from spontaneous speech. For the purpose of the illustration, consider the CC onset to be the first word in CCs without *that* or the first two words in CCs with that. The overall Shannon information of the CC onset then consists of the information contained in the syntactic transition to a complement clause (the information that there is a complement clause) given the preceding context, and the information contained in the first words in the CC given that there is a CC and given the preceding context. In other words, $I(\text{CC onset}|\textit{context}) = I(\text{CC}|\textit{context}) + I(\text{onset}|\textit{context, CC}) = -\log p(\text{CC}|\textit{context}) + -\log p(\text{onset}|\textit{context, CC})$. Since reliable estimates for the entire CC onset's information are considerably harder to obtain than estimates of the first component in the above equation (there necessarily are much fewer observations per type), the current study focuses exclusively on the first component, $I(\text{CC}|\textit{context}) = -\log p(\text{CC}|\textit{context})$. More specifically, the simplest possible estimate of this quantity is used, where the information of the CC onset is only conditioned on the matrix verb. That is, $I(\text{CC}|\textit{context})$ is estimated as $-\log p(\text{CC}|\textit{matrix verb lemma})$. This estimate will be lower the more probable it is that a complement clause follows the matrix verb (e.g. *think* vs. *confirm* in Fig. 1). In the current study, it is hence a verb's subcategorization frequency that is used to estimate the information density at the CC onset (see Appendix A for details).

Most experiments on sentence production test one or two hypotheses at a time, typically using a small set of homogeneous stimuli with lexical and structural properties that are extremely rare in spontaneous language use. Here, I take a different approach. A large number of complement clauses is extracted from a corpus of American English speech (the Switchboard corpus Godfrey et al., 1992) and one single multilevel logit analysis simultaneously tests the predictions of UID while controlling for various alternative accounts of syntactic production, such as availability accounts (Ferreira, 1996; Ferreira & Dell, 2000; Race & MacDonald, 2003), ambiguity avoidance accounts (Bolinger, 1972; Hawkins, 2004; Temperley, 2003), and dependency processing accounts (Elsness, 1984; Hawkins, 2001, 2004), while controlling for syntactic persistence (Bock, 1986; Ferreira, 2003), social effects (Adamson, 1992; Fries, 1940), and effects of grammaticalization (Thompson & Mulac, 1991b; Torres Cacoullos &

Walker, 2009) on *that*-mentioning. With such statistical control, it is possible to benefit from the advantages of corpus-based work on spontaneous language production while minimizing the disadvantages. I will return to the trade-offs of corpus-based research in the general discussion.

Next, I describe the database and the statistical procedure employed in this study. Following that, I describe the controls included in the multilevel logit model analysis.

2.1. Database

The data comes from the Penn Treebank (release 3, Marcus, Santorini, Marcinkiewicz, & Taylor, 1999) subset of the Switchboard corpus of telephone dialogues (Godfrey et al., 1992). The corpus consists of approximately 800,000 words in 642 conversations between two speakers each (approximately gender-balanced) on a variety of topics. The version of the corpus used here is the Paraphrase Stanford-Edinburgh LINK Switchboard Corpus (Bresnan et al., 2002; Calhoun, Nissim, Steedman, & Brenier, 2005). Crucially, to the current study, the corpus was syntactically annotated and manually checked as part of the Penn Treebank project (Marcus et al., 1999). The syntactic annotation makes it possible to extract not only full complement clauses (with *that*) but also reduced complement clauses (without *that*) with high reliability. For an excellent overview of the annotation available for the Paraphrase corpus, see Calhoun (2006). Jean Carletta and colleagues kindly provided us with a version of the corpus in a format compatible with the syntactic search software TGrep2 (Rohde, 2005).

The TGrep2 pattern used to extract all and only CCs from the corpus is given in Appendix B. The pattern returned 7369 complement clauses. Manual inspection of the cases in the database (see Appendix B for more detail) resulted in the removal of 144 cases (2%). This error rate is considerably smaller than in earlier corpus studies using automatically parsed corpora (Roland et al., 2005, see also Roland, Dick, & Elman, 2007). Next, 71 cases (1% of total) were excluded because the matrix verb was incompatible with complementizer omission (see Appendix B). Another 138 cases (1.9% of total) were excluded because the matrix verb lemma did not occur at least 100 times in the corpus. This considerably improves the quality of information density estimates in the database. All remaining cases were automatically annotated for the control variables described in the next section using the TGrep2 Database Tools (Jaeger, 2006b; for more details, see Jaeger, 2006a, chap. 2). An additional 303 cases (4.1% of total) had to be excluded because of missing values for the control variables. The results presented below do not depend on any of these exclusions.

The remaining 6716 CCs come from 346 speakers. Of these, 1173 (17.5%) have a complementizer, while 5543 (82.5%) do not. Similarly low complementizer rates have been reported in previous work (Tagliamonte & Smith, 2005; Torres Cacoullos & Walker, 2009). The low overall rate of *that*-mentioning is primarily due to a few verbs with a low *that*-bias that make up the majority of the cases in the database. Cases with the matrix verbs *think* and *guess* rarely occur with a complementizer. These verbs make up 66% of the database. The remaining 27 verb lemmas, however, have much higher complementizer rates. This is illustrated in Table 1 (see also Table A.1 in Appendix A).

2.2. Statistical procedure: multilevel logit regression

A multilevel logit model, a type of generalized linear mixed model (Breslow & Clayton, 1993; Lindstrom & Bates, 1990; for an overview, see Agresti, 2002, chap. 12), is used to test the partial effect of information density while controlling for other variables known to correlate with *that*-mentioning. The analyzed categorical outcome (dependent variable) is the presence of *that* over its absence. The model contains 25 parameters for controls described in detail below and one parameter to analyze the effect of information density. Together these are the

so-called fixed effects. Additionally, the analysis includes a random speaker intercept, which can be thought of as the individual adjustment to each speaker's rate of *that*-mentioning. The abbreviated model equation is given in (1), where β_1 to β_{25} are the control parameters, $\beta_{InfoDensity}$ is the parameter associated with the information density predictor, and u_i is the random speaker intercept assumed to be normally distributed with variance $\sigma^2_{speaker}$

$$\text{logit}(that)=\text{In}\frac{p(that)}{1-p(that)}=\text{In}\frac{p(that)}{p(\text{no } that)}=\beta_1+\beta_2C_2+\cdots+\beta_{25}C_{25}+\beta_{InfoDensity}*InfoDensity+u_i,\quad u_i\sim N(0,\sigma^2_{speaker}) \tag{1}$$

The model was fit using Laplace Approximation as implemented in the *lmer*( ) function of the *lme4* package (Bates, Maechler, & Dai, 2008) in the statistical software R (R Development Core Team, 2008). Introductions to multilevel logit models in R are available in Gelman and Hill (2006, chap. 14) and, specifically for language researchers, Baayen (2008, chap. 7) and Jaeger (2008). Harrell (2001) provides an excellent overview of regression strategies with reference to R.

If guidelines of model evaluation are followed, multilevel models can be used to reliably analyze even highly unbalanced and clustered data like those typically present in corpus studies. For example, in the current data set, a few speakers contribute most of the data while many speakers contribute only a few (mean number of cases per speaker = 19.3, median = 14, mode = 14, range = 1–99, SD = 16.4). This is illustrated in the histogram in Fig. 2. Such clusters, if unaccounted for, can lead to spurious statistical results. The random speaker intercept in (1) addresses this issue.

Additionally, estimated standard errors of fitted parameters become unreliable if the associated predictor is highly collinear with other predictors. The results presented here are not affected by collinearity. If not mentioned otherwise, correlations between fixed effects were very low ($r < 0.2$). Where necessary, this was achieved by means of residualization of correlated predictors (see below).

The approach taken here thus differs from previous large-scale studies of complementizer mentioning (Roland et al., 2005; Torres Cacoullos & Walker, 2009). Roland and colleagues' model contains 186 parameters, many of which are correlated with each other. Torres Cacoullos and Walker (2009), too, do not control for collinearity between predictors in their model. While it is still possible to assess whether a predictor has an effect (e.g. via model comparison, as in Roland et al., 2005), it is difficult to reliably assess effect *directions* for collinear predictors. The goal of the approach taken here is to move closer to the direct tests of theories of sentence production within an integrated model, while simultaneously assessing the independent effects of multiple hypothesized mechanisms. This has the advantage that it results in interpretable parameter effects, thereby not only testing whether a predictor affects *that*-mentioning, but whether it does so in the predicted direction (without requiring post hoc tests, cf. Roland et al., 2005).

2.3. Controls

Syntactic reduction has been attributed to a variety of factors and mechanisms (Adamson, 1992; Bolinger, 1972; Dor, 2005; Elsness, 1984; Ferreira & Dell, 2000; Ferreira, 2003; Finegan & Biber, 2001; Fox & Thompson, 2007; Hawkins, 2001, 2004; Race & MacDonald, 2003; Temperley, 2003; Thompson & Mulac, 1991a; Yaguchi, 2001). To put the hypothesis of Uniform Information Density to as stringent an empirical test as possible, it is necessary to control for other effects known to affect complementizer *that*-mentioning. I briefly summarize the three arguably most influential processing accounts of *that*-mentioning.

Ferreira and Dell (2000) propose an account of syntactic reduction that is exclusively driven by production pressures. The general idea is that speakers insert optional words, such as relativizers or complementizers, when they would not be able to continue production fluently otherwise. That is, speakers are assumed to utter optional words if the material that has to be uttered next is not readily available. This idea of availability-based production is based on the *Principle of Immediate Mention*:

> "Production proceeds more efficiently if syntactic structures are used that permit quickly selected lemmas to be mentioned as soon as possible."
>
> (Ferreira & Dell, 2000, 289 ff.)

The principle of immediate mention predicts that the accessibility of material at the complement clause onset determines whether speakers produce the *that*. Evidence for this prediction comes from production experiments (Ferreira & Dell, 2000; Ferreira & Hudson, 2005) and corpus studies (Elsness, 1984; Roland et al., 2005; Tagliamonte & Smith, 2005; Torres Cacoullos & Walker, 2009; see also Jaeger & Wasow (2006), Race & MacDonald (2003), Tagliamonte, Smith, & Lawrence (2005), Temperley (2003) for similar evidence from relativizer omission).

An alternative account of effects associated with the complement clause onset is that speakers insert *that* to avoid temporary ambiguity (Bolinger, 1972; Hawkins, 2004, see also Temperley (2003) for relativizer omission) as in the following example, where the complement clause subject *you* could lead to temporary ambiguity, if the speaker does not insert *that* before it:

(2) Well, I know (that) you need to go.

The third account focuses on the effect that mentioning *that* has on the time it takes to process all dependencies between elements in the complement clause and elements in the matrix clause (cf. Domain Minimization and Maximize Online Processing, Hawkins, 2001, 2004). On the one hand, mentioning *that* slightly increases the complexity of the complement clause. On the other hand, mentioning *that* can shorten some of the dependencies (e.g. because it clearly marks the beginning of a complement clause). Dependency processing accounts have received support from studies showing that increased distance between the matrix verb (e.g. *know* in (2)) and the complement clause onset correlates with higher preference for complementizer *that* (Elsness, 1984; Hawkins, 2001; Roland et al., 2005; see also Quirk (1957), Fox & Thompson (2007), Race & MacDonald (2003) for relativizer mentioning).

Table 2 summarizes the control predictors included in the analysis to account for availability-based production, ambiguity avoidance, and dependency processing, as well as additional controls that previous studies found to affect *that*-mentioning. Next, I describe these control predictors in detail.

2.3.1. Dependency length and position—Various measures of domain complexity preceding and within the CC have been found to be correlated with *that*-mentioning (e.g. Elsness, 1984; Ferreira & Dell, 2000; Hawkins, 2001, 2004; Roland et al., 2005; Tagliamonte et al., 2005; Thompson & Mulac, 1991a). Four such measures are included in the analysis.

Length(matrix verb-to-CC): Given the results from previous studies, non-adjacent CCs are expected to strongly prefer *that* (Elsness, 1984; Hawkins, 2001; Rohdenburg, 1998). Principles like Hawkins' Domain Minimization and Maximize Online Processing (Hawkins, 2001, 2004) predict an increasing preference for *that* the more material intervenes between the matrix verb and the CC onset. So, the number of words between the matrix verb and the CC onset was included as a continuous predictor. For example, in (3a) there zero intervening words,

whereas in (3b) there are two intervening word. There were 222 cases with intervening material between the verb and the CC (mean number of intervening words for those cases = 2.0).

(3) **a.** My boss thinks [I'm absolutely crazy].

b. I agree with you [that, that a person's heart can be changed].

Length(CC onset): Dependency processing accounts also predict that the length of the CC subject correlates positively with speakers' preference for *that*. Here, the number of fluent words up to and including the CC subject was included as a continuous predictor. For example, the CC onset in (3a) contains one word, whereas the CC onset in (3b) contains four words (*a person's heart*; clitics were counted as separate words). Only fluent words were counted to avoid collinearity with the measures of disfluency introduced below. The complementizer – if present – was also not counted to avoid a trivial correlation with *that*-mentioning.

Length(CC remainder): The number of words in the CC following its subject was also included as a continuous predictor (e.g. three words in both (3a) and (3b)). Dependency processing accounts only predict an effect of this variable if other dependencies hold between material preceding the CC and material following it. This is unlikely to be a frequent event, and even for cases with such dependencies, the effect would be predicted to be very weak.

Position(matrix verb): It is possible that production difficulty differs systematically depending on where speakers are in the process of planning and pronouncing a sentence. Unfortunately, very little is known about relative production difficulty during incremental sentence production (in sharp contrast to sentence comprehension). To capture any potentially non-linear effects of the position of the complement clause in the overall sentence, a restricted cubic spline over the number of words preceding the matrix verb (two words in (3a) and one word in (3b)) was included in the model (using *rcs*( ), package *Design*, Harrell, 2007). Restricted cubic splines provide a convenient way to model non-linear effects of predictors (for a concise summary, see Harrell (2001, pp. 16–24); for an introduction to *rcs( )* for the analysis of language data, see Baayen (2008, chap. 6.2.1)). An effect of matrix verb position might also be predicted by grammaticalization accounts of *that*-mentioning, which consider certain uses of CC-embedding verbs to have become grammaticalized without *that* (Thompson & Mulac, 1991b; discussed below). Three parameters (4 knots) were used for the restrictive cubic spline, allowing for a moderate degree of non-linearity (consistent with the results, as also confirmed by further tests allowing for higher degrees of non-linearity). Following Harrell (2001), knots were placed at the 5th, 35th, 65th, and 95th quantiles, corresponding to 2, 3, 4, or 16 words preceding the complement clause.

2.3.2. Overt production difficulty at CC onset—Speakers are more likely to produce optional *that* when they are experiencing production difficulty (Ferreira & Firato, 2002; Jaeger, 2005). Overt signs of production difficulty provide a window into underlying production difficulty that may go beyond what the controls introduced above capture. Speech rate and pause information were extracted from the time-aligned orthographic transcripts of Switchboard (Deshmukh, Ganapathiraju, Gleeson, Hamaker, & Picone, 1998).

Log speech rate and squared log speech rate: Both the log-transformed and the square of the logtransformed speech rate at the CC onset were included because both have been shown to affect the phonetic reduction of words (Bell et al., 2003). Speech rates in the data set are roughly normally distributed with a heavy right tail (mean = 4.8 syllables/second, SD = 1.4). To reduce collinearity between these two factors, the log-transformed speech rate was centered before it was squared. Since the two measures were still correlated, the LOG SPEECH RATE

was regressed out of SQUARED LOG SPEECH RATE and the residuals were included in the analysis of *that*-mentioning to assess the effect of SQ LOG SPEECH RATE.

Pause: The presence or absence of a suspension of speech at the CC onset was included as a binary control predictor. About 10% (680) of all cases contain a pause at the CC onset. Pause lengths range from 10 ms to almost 13 s (mean = 453 ms, SD = 626).

Disfluency: The number of disfluencies at the CC onset up to and including the CC subject was also included as a control. Disfluencies included restarts (e.g. "*It was a, a, a, guy*"), editing terms (e.g. *you know, I mean*), and filled pauses (e.g. *uh, um*). While these categories arguably differ in many respects, they all are correlates of production difficulty. For example, the CC onset (4) contains three disfluencies (two restarts and one filled pause).

(4) They said [the, that, um, that he was horrible at batting with men on base].

To reduce collinearity between this measure and the length of the CC onset (longer phrases on average have more disfluencies), the disfluency count was regressed against LENGTH(CC ONSET) and the residuals of that regression were included in the main analysis to assess disfluency effects. The number of disfluencies intervening between the matrix verb and CC was not included as control because there were too few cases with disfluencies.

2.3.3. Lexical retrieval at CC onset

CC subject: The accessibility of the CC subject was coded based on its referential expression (Gundel, Hedberg, & Zacharski, 1993). Previous work has found that pronouns, especially the local pronoun *I*, correlate with lower *that* rates than other pronouns and lexical NPs (Elsness, 1984; Ferreira & Dell, 2000; Roland et al., 2005). CC SUBJECT was coded as four ordered levels: *I*(899 cases) > *it*(1294) > - other pronouns, including demonstrative pronouns (2874) > other types of NPs (1649). For example, the CC subject in (3a) is *I*, the CC subject in (3b) was coded as 'other type of NP', and the CC subject in (4) was coded as 'other pronoun'.

The effect of CC subject accessibility on *that*-mentioning has been attributed to an availability-based strategy of speakers to insert *that* when following material is not readily available for pronunciation (Ferreira & Dell, 2000). In the Switchboard corpus, the word *it* is frequently used non-referentially (as in "…*it rains*…") or to refer to a highly salient issue under discussion. Because both of these uses are likely to make *it* easy to retrieve and hence highly available, *it* is expected to pattern with the local pronoun *I*. Three Helmert-contrasts over the four levels of CC SUBJECT were included in the model, comparing each of the three lowest levels against all higher levels (*it* vs. *I; other pronouns* vs. *I* and *it*; etc.).

Subject identity: Both Elsness (1984) and Ferreira and Dell (2000) found that co-referentiality of the matrix and CC subject correlated with a lower *that* rate. While recent experiments (Ferreira & Hudson, 2005) and corpus studies (Torres Cacoullos & Walker, 2009) have not replicated this effect, SUBJECT IDENTITY encoding whether the matrix and CC subjects were string identical was included in the analysis. Since this variable is highly correlated both with the type of matrix subject (see below) and the type of CC subject, the centered binary SUBJECT IDENTITY variable was first regressed against the referential form of the CC subject (CC SUBJECT) and the referential form of the matrix subject (MATRIX SUBJECT). The residuals of this regression were entered into the main analysis. SUBJECT IDENTITY hence tests the effect of subject identity beyond the properties of the matrix and CC subjects.

Frequency(CC subject head): In addition to these form-based predictors, the model included the log-transformed frequency of the CC subject's head lemma (e.g. the frequency of *I* for (3a) or the frequency of *heart* for (3b) above). A lemma-based rather than word form-based estimate

was chosen to reduce data sparsity (by combining the counts from all word forms associated with a lemma). Using word form frequency does not change any of the results reported below. Since most CC subject phrases were pronouns, their head is also the only word in that NP. In 85% of all cases in the database, the subject head is also the only word in the CC subject.

Since more frequent words are produced faster (Jescheniak & Levelt, 1994), availability accounts of *that*-mentioning predict that the less frequent the CC subject's head is, the more likely speakers are to produce *that*. This has indeed been observed in previous studies (Roland et al., 2005).

The frequency of the CC subject is, however, highly correlated with the form of its referential expression (CC SUBJECT, Spearman rank correlation $r = -0.80$). Hence, the effect of CC SUBJECT was regressed out of the log-transformed frequency and the residuals of that regression were entered into the analysis of *that*-mentioning.

Word form similarity: Finally, a binary predictor was included to encode whether the first word in the CC (excluding the complementizer *that*) was the demonstrative determiner or pronoun *that* (as in, e.g. "I believe (that) that paint is what I need."). There is preliminary evidence that speakers are about half as likely to produce complementizer *that* if the next word is demonstrative *that* (Jaeger, in press; Walter & Jaeger, 2008). Walter and Jaeger attribute this effect to a bias against adjacent similar word forms (the Obligatory Contour Principle Leben, 1973), presumably due to interference effect. A bias against adjacent similar or identical linguistic elements has been observed at many levels of linguistic representation (for a recent overview, see Walter, 2007). Similarity-based interference has also been observed in language comprehension (Gordon, Hendrik, & Johnson, 2001; Lewis, 1996) and production (e.g. in agreement errors, Badecker & Lewis, 2007; or phonological priming, Bock, 1987).

About 10% of the CCs in the data (696 cases) start with demonstrative *that* and hence could potentially exhibit an OCP effect. Since by far most of the instances of the demonstrative *that* (96.7%) are the demonstrative pronoun *that* (rather than the demonstrative determiner), which means that WORD FORM SIMILARITY is correlated with SUBJECT FORM == *other pronoun* (Spearman rank correlation $r = 0.37$), it is necessary to dissociate the two effects.

The WORD FORM SIMILARITY variable was regressed against an indicator variable which coded whether the CC subject was a pronoun. The residuals of that linear regression were included in the analysis as the WORD FORM SIMILARITY predictor.

2.3.4. Lexical retrieval immediately preceding CC

Frequency(matrix verb): The matrix verb directly precedes the CC in 93.5% of all cases. Increased processing load associated with the production of less frequent word forms (Jescheniak & Levelt, 1994; Baayen et al., 2006) may spill over from the matrix verb to the CC onset. In that case, the availability- based account of *that*-mentioning predicts higher rates of *that*-mentioning following less frequent matrix verbs. Evidence for this hypothesis comes from previous corpus studies (Elsness, 1984; Garnsey, Pearlmutter, Meyers, & Lotocky, 1997; Roland et al., 2005). The log-transformed frequency of the matrix verb lemma was included in the analysis to account for this effect. The frequency of the matrix verb rather than the word that immediately preceded the CC was chosen to avoid collinearity with the effect for LENGTH(MATRIX VERB-TO-CC).

2.3.5. Ambiguity avoidance at CC onset

Ambiguous CC onset: Without a complementizer some CC onsets can lead to temporary syntactic ambiguity. If the matrix verb is compatible with subcategorization frames other than

complement clauses and if the CC onset is not unambiguously case-marked (e.g. *I* and *we* vs. *you* or *the man*) and if the matrix onset preceding the complement clause does not unambiguously select a verb sense that requires a complement clause, the CC onset without the complementizer could temporarily be interpreted as an argument to the matrix verb. Consider, for example, (5) where *many of them* could also temporarily be interpreted as the direct object to the matrix verb *know*. Such temporary ambiguities can lead to considerably increased processing times (due to 'garden-pathing'; e.g. Garnsey et al., 1997; Trueswell, Tanenhaus, & Kello, 1993).

(5) …and I know [many of them are doing it].

This raises the question as to whether speakers avoid such temporary ambiguities. Indeed, some studies on *that*-mentioning in complement clauses (Bolinger, 1972; Hawkins, 2004) and relative clauses (Bolinger, 1972; Temperley, 2003) have provided evidence that producers mention *that* to avoid temporary ambiguities. However, these studies were generally conducted on small databases of written (often edited) texts and they lacked controls for most of the other factors affecting syntactic reduction. More recent and more controlled studies have failed to find any effect of ambiguity avoidance on *that*-mentioning (Ferreira & Dell, 2000; Roland et al., 2005). Most previous studies did not use spontaneous speech data (but see Roland et al., 2005, 2007) and none of the corpus studies used data that was manually annotated for potential ambiguity.

Here, a CC onset was coded as potentially ambiguous if the onset of the utterance up to and including the CC subject (but without the complementizer, if it was present) could potentially lead to a garden path. First, all cases with case-marked CC subjects or with a matrix verb that never or rarely takes a CC complement (*think*, *suppose*, *wish*, *figure*, *hope*, and *agree*) were automatically marked as unambiguous. This left 2973 cases that were manually annotated. This annotation was conducted by an undergraduate research assistant at the University of Rochester, after initial training by the author on 100 cases. The annotation marked 1012 CCs as potentially containing a temporary ambiguity (only about a third of the cases that would have been judged ambiguous based on only case-marking and verb subcategorization constraints!).

2.3.6. Grammaticalization of epistemic markers

Matrix subject: Several studies have found that *that*-mentioning is also strongly correlated with the form of the *matrix* subject. This effect has variously been attributed to decreased processing load for co-referentiality between the matrix and CC subject (Ferreira & Dell, 2000) or the use of grammaticalized discourse formulae (Thompson & Mulac, 1991b). Here, potential effects of co-referentiality are already modeled by the SUBJECT IDENTITY predictor introduced above.

According to grammaticalization accounts (Thompson & Mulac, 1991b), *that*-mentioning is not an alternation between two meaning-equivalent variants (Thompson & Mulac, 1991b, pp. 313–314). Rather, there are epistemic uses of complement clause embedding verbs, which occur without *that*, and real complement clauses, which are assumed to occur with *that*. For example, in (6) (taken from Thompson & Mulac (1991a, p. 313)) *I think* does not necessarily introduce a complement clause, but rather expresses a "degree of speaker commitment [to the proposition conveyed in the remainder of the clause]" (Thompson & Mulac, 1991a).

(6) I think exercise is really beneficial, for anybody.

The strongest version of the grammaticalization hypothesis, according to which there is no alternation, is incompatible even with the data presented in Thompson and Mulac (1991a). After all epistemic phrases are excluded from their data, there still is considerable variation in

that-mentioning that needs to be accounted for. Previous work (Thompson & Mulac, 1991a; Torres Cacoullos & Walker, 2009) does, however, support weaker grammaticalization accounts. Grammaticalization could be gradient or it could be *one* of several factors contributing to the overall pattern of *that*-mentioning. Indeed, there is evidence that certain highly frequent matrix clause onsets, such as *I think* and *I guess*, are associated with much lower rates of *that*-mentioning than other instances of complement clause embedding verbs (Thompson & Mulac, 1991a; Torres Cacoullos & Walker, 2009). According to Thompson and Mulac (1991b), speakers are most likely to express a degree of commitment when talking about themselves, but do so also when talking about their addressee(s) (Thompson & Mulac, 1991b, p. 243). Hence 1st person matrix subject *I* should correlate with the lowest rates of *that*-mentioning, followed by 2nd person matrix subjects. All other types of matrix subjects should correlate with considerably higher rates of *that*-mentioning. Thompson and Mulac (1991a), 321 present evidence for such a three-way distinction between matrix subjects. Other studies have, however, found that only matrix subject *I* differed significantly from other types of matrix subjects (Torres Cacoullos & Walker, 2009).

Potential effects of the matrix subject form are modeled as a treatment-coded predictor with three levels: *I* (5355 cases) vs. *you* (351) vs. other personal pronoun (515 cases) vs. other NP (495 cases, including zero subjects). Each level was contrasted against *I*, in such a way that the coefficient corresponds to the distance between the two group means.

2.3.7. Additional controls

Syntactic persistence: Optional *that*-mentioning is also affected by syntactic persistence, that is whether the most recent CC was produced with or without a complementizer *that* (Ferreira, 2003; Jaeger & Snider, 2008). SYNTACTIC PERSISTENCE was coded as categorical predictor with three levels: no preceding prime (616 cases), preceding prime with *that* (1239), preceding prime without *that* (4861). This coding does not distinguish between primes by the speaker and primes by their interlocutor. Helmert contrasts were used to model the effect of syntactic persistence, comparing *that*-mentioning for 'no prime' against *that*-mentioning for 'primes without *that*', and the effect of both of these levels against the effect of 'primes with *that*'.

Speaker gender: Previous work found that social variables correlate with syntactic reduction (Adamson, 1992; Fries, 1940). Here speaker gender is included in the analysis since there is some evidence that female speakers are more likely to produce *relativizers* in non-subject-extracted relative clauses (Staum & Jaeger, 2005; although we did not find such an effect for complementizer *that*-mentioning). This would be consistent with the observation that female speakers are more likely to choose more formal registers and formal registers are correlated with less reduction (Finegan & Biber, 2001).

2.4. Model evaluation

Before results are reported, it is important to evaluate the fitted model. This involves testing for signs of overfitting, collinearity, and outliers (see Baayen, 2008, Section 6.2.3-4; Jaeger & Kuperman, 2009; Jaeger, submitted for publication). Plotting mean predicted probabilities of *that* vs. observed proportions in Fig. 3 shows an acceptable fit of the model. Unsurprisingly, the fit is best for bins with low observed proportions of *that*, which contain most of the data. Further simulations over the estimated parameters (using *sim( )*, package *arm*, Gelman et al., 2008) showed no signs of overfitting: all estimated coefficients were stable. There were no signs of collinearity in the model (fixed effect correlations rs < 0.2).

The model accounts for a significant amount of information in the variation of *that*-mentioning. While several *pseudo* R^2 measures have also been proposed for logit models as measures of

overall model quality, they lack the intuitive and reliable interpretation of the R^2 available for linear models. One of the most commonly used pseudo R^2 measures is the Nagelkerke R^2. The Nagelkerke R^2 assesses the quality of a model with regard to a baseline model (the model with only the intercept). While this measure is usually employed for ordinary logit models, its definition extends to multilevel logit models. For the current model, Nagelkerke $R^2 = 0.34$ compared against an ordinary logit model with only the intercept as baseline.[2]

2.5. Effect of information density

As predicted by UID, there was a clearly significant effect of information density: the higher the information of the CC onset, the more likely speakers are to produce the word *that* ($\beta = 0.47$; $z = 16.9$; $p < 0.0001$). In other words, speakers are less likely to produce a complementizer *that*, the lower the information density of the CC onset. This effect holds even while other variables that previous studies found to be correlated with *that*-mentioning are controlled for. Fixed effect correlations of information density with other predictors were negligible (all $r < 0.13$), so that collinearity is not a concern for the assessment of the effect. Fig. 4 illustrates the effect on *that*-mentioning.

As a matter of fact, information density is the strongest predictor in the model in terms of its contribution to the model's likelihood $\left(\chi^2_{\Delta(\Lambda)}(1) = 263.0, p < 0.0001\right)$. The χ^2-test serves as a measure of how much the model is improved by the inclusion of the predictor.[3] The partial Nagelkerke R^2 associated with information density is 0.05. This suggests that at least 15% of the model quality (in terms of the Nagelkerke R^2) can be attributed to information density.

This improvement in model quality more than matches that of all four fluency parameters combined $\left(\chi^2_{\Delta(\Lambda)}(4) = 194.2, p < 0.0001\right)$. To put the effect in relation to two theories of sentence production that have received considerable attention in psycholinguistic work on syntactic variation, availability-based sentence production (Bock, 1986; Ferreira & Dell, 2000; Levelt & Maassen, 1981; Prat-Sala & Branigan, 2000, among others) and dependency processing accounts (Hawkins, 1994, 2004): the effect associated with the only parameter fitted for information density outranks the effect of all three parameters associated with dependency length effects in the model (LENGTH(MATRIX VERB-TO-CC), LENGTH(CC ONSET), LENGTH(CC REMAINDER); $\left(\chi^2_{\Delta(\Lambda)}(3) = 167.8, p < 0.0001\right)$. The effect of information density also is much larger than the combined effect of accessibility related parameters in the model (CC SUBJECT, FREQUENCY(CC SUBJECT), SUBJECT IDENTITY; $\chi^2_{\Delta(\Lambda)}(5) = 27.0, p < 0.0002\right)$. That is, information density emerges as the single most important predictor of complementizer *that*-mentioning.

[2]Despite my own skepticism about Nagelkerke R^2s, I have chosen to present them since they may be familiar to some readers. All Nagelkerke R^2s are calculated against an ordinary logit model with only the intercept as baseline. The R code used to calculate the Nagelkerke R^2s presented here is available at http://hlplab.wordpress.com/2009/08/29/nagelkerke-and-coxsnellpseudo-r2-for-mixed-logit-models/.

[3]For large data sets like the current one, differences in deviance (=−2 * log Λ, where Λ is the data likelihood of the model) between nested models that are maximum likelihood fitted on the same data are approximately χ^2-distributed. A χ^2-test can then be used to compare a model against the same model without a given predictor. The degrees of freedom of the χ^2-test correspond to the degrees of freedom associated with the predictor. Non-residualized versions of predictors are used for all model comparisons reported here, since comparisons based on data likelihood are robust against collinearity and using the residualized predictors would potentially underestimate contributions of predictors. It is worth noting that measures of model quality based on the data likelihood (including the Nagelkerke R^2 and other pseudo R^2 measures) have to be interpreted cautiously for *multilevel logit* models. Maximization of the likelihood of multilevel logit models does not have a known analytic solution. Here Laplace approximation was used to fit the model, which has been shown to provide computationally efficient and accurate estimates (Harding & Hausman, 2007). For further discussion, see also https://stat.ethz.ch/pipermail/r-sig-mixed-models/2008q3/001233.html.

It is well-known that embedding verbs differ considerably with regard to their *that*-bias (the associated rate of *that*-mentioning, Garnsey et al., 1997; Roland et al., 2005, 2007). In the current database, *that*-biases range from 1% for *guess* to 75% for *worry* (see Table A.1 in Appendix A). On the one hand, UID differs from availability, ambiguity avoidance, and dependency processing accounts in that it potentially offers an account for these otherwise idiosyncratic differences between verbs. Note for example, that the verb *guess*, which has the lowest *that*-bias, also has one of the highest CC-biases (62.3% of all instances of *guess* in the corpus take a complement clause). In comparison, the verb *worry*, which has the highest *that*-bias, has a more than 20-times lower CC-bias (3.2%). Since the estimate of information density at the CC-onset used in this paper is only based on verbs' subcategorization frequency, these observations are in line with the predictions of UID.

However, embedding verbs differ along dimension besides subcategorization frequency. It is possible that the analysis presented above fails to account for differences between embedding verbs that affect *that*-mentioning. Hence it is conceivable that the observed correlation between information density and *that*-mentioning is due to differences between embedding verbs that are not modeled in the analysis. To test this hypothesis, a random intercept for verb lemmas was added to the model and the analysis was repeated. Indeed, the estimated variance of the random verb lemma intercept is large $\left(S^2_{matrix\ lemma}=1.011\right)$ in line with Roland et al. (2007). The random verb lemma intercept improves the model considerably $\left(\chi^2_{\Delta(\Lambda)}(1)=233.5, p<0.0001\right)$; Nagelkerke R^2 = 0.38, cf. 0.34 for a model without the random verb lemma effect). It is worth noting, that, unlike for the other predictors in the model, the random verb lemma intercept is not motivated by a specific processing account. Also, unlike for the fixed predictor of information density, inclusion of a random effect does not make any assumption as to how verbs align with their *that*-bias. Instead, the random effect simply models arbitrary difference in *that*-bias between verb lemmas. Nevertheless, the information density effect remains significant in the expected direction $\left(\chi^2_{\Delta(\Lambda)}(1)=7.1, p<0.0008\right)$. That is, while future work is necessary to determine exactly how much of the variance in *that*-mentioning associated with lexical differences can be attributed to UID, the observed effect of information density cannot be reduced to arbitrary properties of embedding verbs (see also Appendix C for further evidence that the observed effect is not due to a few verbs, including meta-analyses of production experiments reported in Ferreira and Dell (2000)).

2.6. Results and discussion of control predictors

Next, I discuss the effects of the control parameters grouped by the accounts of syntactic production that they bear on. In an effort to make the results more accessible, I present and discuss findings grouped by the accounts of sentence production on which they bear on. I first discuss availability, ambiguity avoidance, and dependency processing accounts (in that order). Then I discuss to what extent there is evidence that the observed patterns in *that*-mentioning are due to grammaticalization (Thompson & Mulac, 1991b) and syntactic persistence (Bock, 1986).

While the discussion of these controls does not directly bear on the primary effect of interest in this article, the tests provided go beyond the confirmation of previous research in several cases. For example, previous work has not investigated the relation between fluency and syntactic reduction, although, as I argue below, the evidence from fluency bears on and refines the hypothesis of availability- based production. Similarly, previous work has largely ignored the role of phonological similarity based interference during *syntactic* planning (but see Bock, 1987; Walter & Jaeger, 2008; Jaeger, in press). Additionally, the current analysis tests previous hypotheses while controlling for a large number of competing accounts. Indeed, not all previous effects replicate.

Table 3 summarizes the parameter estimates β for all fixed effects in the model, as well as the estimate of their standard error SE(β), the associated Wald's z-score, and the significance level.

The estimated variance of the random speaker intercept is $s^2_{speaker}=0.364$. The partial Nagelkerke R^2 associated with the random speaker effect is 0.01.

2.6.1. Availability-based production—Support for the availability account of syntactic reduction (Ferreira & Dell, 2000) comes from the fluency controls, the controls for accessibility effects of the CC subject, and the other controls for lexical retrieval effects at and immediately preceding the CC onset.

Fluency: Probably the strongest evidence that the availability of material at the CC onset affects *that*-mentioning (Ferreira & Dell, 2000) comes from the clearly significant effect of fluency predictors. Lower SPEECH RATE, presence of a PRECEDING PAUSE, and INITIAL DISFLUENCY correlate with higher complementizer rates. As previous work has found for phonetic reduction (Bell et al., 2003), speech rate has a highly significant effect on *that*-mentioning (log-transformed speech rate: $\beta = -0.70$; $z = -5.5$; $p < 0.0001$; residual squared log-transformed speech rate: $\beta = -0.36$; $z = -1.9$; $p < 0.06$). The effect of CC initial disfluencies, such as filled pauses or restarts, is also highly significant ($\beta = 0.39$; $z = 3.2$; $p < 0.002$): the odds of complementizer *that* are about $1.5(= e^{0.39})$ times higher for CCs with a disfluency than for CCs with fluent onsets. Pauses, too, strongly correlate with higher *that* rates ($\beta = 1.11$; $z = 10.2$; $p < 0.0001$; fixed effect correlation with log speech rate $r = 0.23$). Interestingly, inserting that does not seem to help speakers (or at least not sufficiently) to overcome the production difficulty that seems to cause *that*-insertion (Jaeger, 2005; see also Ferreira & Firato, 2002).

Subject identity, referential form, and frequency of CC subject: As predicted by availability accounts, co-referentiality between the matrix and CC subject correlates with lower *that* rates although the effect does not quite achieve significance ($\beta = -0.32$; $z = 1.9$; $p < 0.052$), thereby providing support, albeit weak support, for Ferreira and Dell (2000) over studies that failed to replicate the co-referentiality effect (Ferreira & Hudson, 2005; Torres Cacoullos & Walker, 2009).

Of the form-based contrasts (CC SUBJECT), lexical and pronoun CC subjects differ significantly ($\beta = 0.11$; $z = 4.9$; $p < 0.0001$; fixed effect correlations with other levels of CC SUBJECT: $0.004 < r\text{s} < 0.28$). As expected, there is no sign that the local pronoun *I* and *it* differ in *that* rates ($p > 0.6$). The contrast between other pronouns vs. *I* and *it* is predicted to be significant by availability accounts, but does not quite reach significance ($\beta = 0.05$; $z = 1.6$; $p = 0.11$). Neither does FREQUENCY(CC SUBJECT HEAD), the residual effect of the CC subject's head frequency beyond CC SUBJECT, reach significance ($p > 0.5$).

As mentioned above, the frequency and form-based effects are highly collinear, and splitting the accessibility effect between the two variables hurts both of them. Without the form-based predictors, CC subject frequency is a highly significant predictor correlating (as expected) negatively with *that* mentioning ($\beta = -0.07$; $z = -4.0$; $p < 0.0001$). Furthermore, CC subject frequency remains a significant predictor ($\beta = -0.18$; $z = -2.5$; $p < 0.012$), if only CCs with pronoun subjects are considered, suggesting that the frequency effect extends beyond the contrast between pronominal and lexical CC subjects. Excluding CC subject frequency does not change the significance of any of the form-based parameters. In sum, there is a clear accessibility effect related to the CC subject, and it is best modeled by the frequency of the subject's head (BIC for model without form-based predictors: 4837; BIC for model without frequency predictor: 4849).[4]

Interference during lexical retrieval: There is also evidence for an interference effect during lexical retrieval due to similarity. If the CC subject is the demonstrative pronoun *that* or if it starts with the demonstrative determiner *that*, speakers are about 1.4 times less likely to mention that complementizer *that* than otherwise, although the effect fails to reach significance ($\beta = -0.31$; $z = -1.8$; $p < 0.08$). The relative weakness of the effect compared to previous work (Jaeger, in press; Walter & Jaeger, 2008) is likely due to the additional controls included in the analysis presented here.

'Spillover': The observed highly significant negative correlation between the lemma frequency of the matrix verb, FREQUENCY(MATRIX VERB), and *that*-mentioning ($\beta = -0.23$; $z = -7.7$; $p < 0.0001$) replicates earlier corpus-based results (Elsness, 1984; Garnsey et al., 1997; Roland et al., 2005). Interestingly, the only previous *experiment* investigating matrix verb frequency found no effect (Ferreira & Dell, 2000, Experiment 3). If confirmed further, the effect is compatible with availability accounts under the assumption that production resources are limited so that high processing load can 'spill over' into the planning of upcoming material.

2.6.2. Ambiguity avoidance—As discussed above, complementizer mentioning has been attributed to ambiguity avoidance (Bolinger, 1972; Hawkins, 2004; see also Temperley, 2003 for relativizer mentioning). However, several other studies have failed to replicate these effects (Ferreira & Dell, 2000; Roland et al., 2005). The results presented here are the first obtained from spontaneous speech and while controlling for a variety of other processing effects. The main analysis did not reveal evidence for ambiguity avoidance. CC onsets that were annotated as potentially ambiguous were not significantly more likely to occur with complementizer *that* ($p > 0.2$). But is possible that the vast majority of the 'potentially ambiguous' cases may lead to negligible or even no garden path effects in context. If speakers only avoid sufficiently severe ambiguity (i.e. cases that actually lead to increased comprehension difficulty), this would explain lack of an ambiguity avoidance effect in the main analysis. Two post hoc tests were conducted to assess this hypothesis.

First, consider that garden path effects have been shown to be weaker, the less probable the 'garden path' parse is (e.g. Garnsey et al., 1997; Trueswell et al., 1993). This observation suggests a modification of previous ambiguity avoidance accounts: maybe speakers only avoid temporary ambiguity if the preceding context biases comprehenders towards the unintended parse (for a similar argument, see Wasow & Arnold, 2003). In other words, speakers should be more likely to insert complementizer *that*, the more likely comprehenders are to choose the wrong parse after observing the first word(s) in the complement clause. This would explain the null effect of AMBIGUOUS CC ONSET since our ambiguity annotation was rather inclusive. That is, cases were annotated as potentially ambiguous, even if there were rather unlikely to lead to a garden path given the subcategorization bias of the verb. Consider the following example, where *this* does not have to start a CC, as in "*Which one would you take?*" - "*I guess this one*".

(7) I guess [this doesn't really have to do with taxes …]

To test whether speakers only avoid potential ambiguity if comprehenders are *likely* to be garden pathed, an interaction term between AMBIGUOUS CC ONSET and the log-transformed probability of a complement clause given the matrix verb was added to the model (except for the sign, this is the same as the information density estimate). This interaction failed to reach significance ($p > 0.5$), providing no evidence that speakers avoid ambiguity.

[4]The BIC (Bayesian Information Criterion, Schwarz, 1978) is a measure of model quality that weighs the model's empirical coverage against its parsimony (BIC = $-2 \ln \Lambda + k \ln n$, where k is the number of parameters in the model, n the number of data points, and Λ is the model's data likelihood). Smaller BICs indicate better models.

But could it be that speakers only avoid potential comprehension difficulty if it is both likely and long-lasting? That is, do speakers only avoid long stretches of potentially ambiguous material (and only if comprehenders are likely to be garden pathed)? To test this hypothesis, all 1012 potentially ambiguous cases in the database were annotated for their disambiguation point. Table 4 summarizes the distribution of potentially ambiguous CC onsets with regard to their disambiguation point.

To test whether speakers only avoid relatively long-lasting ambiguity, the original predictor AMBIGUOUS CC ONSET in the model was replaced by the *length of the potential ambiguity* (in words). Unambiguous cases were coded as having an ambiguity length of 0. Both the (centered) ambiguity length predictor and its interaction with the probability of a complement clause given the matrix verb were entered into the model. The main effect of ambiguity length was not found to be significant ($p > 0.4$), but its interaction with CC predictability is approaching significance in the expected direction ($\beta = -0\ 14$, $z = -1.6$, $p = 0.1$):.For unexpected CCs with otherwise late disambiguation points, speakers are more likely to insert *that*. Including the interaction in the model does lead to a marginal improvement of the model's log-likelihood $\left(\chi^2_{\Delta(\Lambda)}(1) = 2.8, p < 0.1\right)$.

In conclusion, there may be a weak effect on ambiguity avoidance on complementizer *that*-mentioning, but the strong effect found in some earlier studies (e.g. Hawkins, 2004) seems to vanish with the additional controls included in the current study. Given that multiple post hoc tests were performed, the marginal effect found above has to be taken with extreme caution. One reason for this weak or non-existent effect of ambiguity avoidance may be that by far most of the potentially ambiguous CC onsets in the database are almost immediately disambiguated (cf. Table 4). In other words, there are very few ambiguities that would be likely to lead to long-lasting garden path effects (for similar evidence from *that*-mentioning in relative clauses, see Jaeger, 2006a, Section 6.3.1). Syntactic reduction may simply be the wrong place to look for ambiguity avoidance.

Finally, it's crucial to point out that a lack of ambiguity avoidance would not argue against *all* types of audience design. I return to this issue in Section 3, where I discuss how the hypothesis of Uniform Information Density relates to audience design.

2.6.3. Dependency processing—Several effects provide evidence for dependency processing accounts (Elsness, 1984; Hawkins, 2001, 2004). Complementizer mentioning correlates positively with the number of words intervening between the matrix verb and the CC onset ($\beta = 0.17$, $z = 2.5$, $p = 0.01$) and the number of words at the CC onset up to and including the CC subject ($\beta = 0.18$, $z = 12.8$,$p < 0.0001$), as predicted in Hawkins (2001, 2004) and replicating previous work (Elsness, 1984; Hawkins, 2001; Roland et al., 2005; Thompson & Mulac, 1991b; Torres Cacoullos & Walker, 2009). It is also worth noting that the coefficients for these two effects are of similar size, which is consistent with the specific account of complementizer omission outlined in (Hawkins, 2001). These effects hold beyond the availability effects discussed above (collinearity between levels of CC SUBJECT and LENGTHCC ONSET was minimal, $-0.16 < rs < 004$).

Interestingly, the number of words in the CC *following* its CC subject also correlates positively with *that*-mentioning $\beta = 003$, $z = 44$, $p < 0.0001$. As discussed above, Domain Minimization and Maximize Online Processing (Hawkins, 2004) would predict no effect or an effect with a much smaller coefficient than for the other two length measures. The latter is indeed the case. The significant effect of the CC complexity beyond the CC subject may be surprising given experimental evidence that language production seems to be rather radically incremental (e.g. Brown-Schmidt & Konopka, 2008; Griffin, 2003; Wheeldon & Lahiri, 1997). There is, however, considerable evidence that speakers must have at least heuristic weight or complexity

estimates of material that is not yet phonologically encoded (e.g. Bock & Cutting, 1992; Clark & Wasow, 1998; Gómez Gallo et al., 2008; Wasow, 1997). Indeed, the empirical distribution of complementizers by overall CC length, shown in Fig. 5, reveals a strong effect of CC complexity up to the 7th to 9th word, after which no trend is observed. This was confirmed by refitting the model with a restricted cubic spline modeling over CC length. The predicted effect of CC length is displayed in Fig. 6.

It thus would seem that speakers have access to at least a heuristic complexity estimate of the CC onset up to the first 7–9 words and that this affects speakers' decision to produce *that* (see Jaeger (2006a) for similar evidence from relativizer mentioning).

2.6.4. Grammaticalization of epistemic phrases—Previous findings provide clear support for (weak) grammaticalization accounts, according to which the distribution of complementizer *that* is partially driven by certain types of matrix clause onsets, such as *I guess*, that have become grammaticalized as epistemic markers. These epistemic markers are assumed to occur without the complementizer *that* (Thompson & Mulac, 1991b). Three findings in the current study are relevant to the evaluation of such weak grammaticalization accounts.

First, verbs hypothesized by Thompson and Mulac (1991b) to be frequently used as epistemic markers, such as *think* and *guess*, are associated with low rates of *that*-mentioning (see Table A.1 in Appendix A). Lexical differences between verbs account for a considerable amount of variation in *that*-mentioning, as evidenced by the highly significant model improvement associated with the addition of a random verb lemma effect. This is compatible with grammaticalization accounts, but could also be due to other lexical properties not controlled in the analysis.

Second, the distance of the matrix verb from the onset of the sentence had a highly significant nonlinear effect on *that*-mentioning $\left(\chi^2_{\Delta(\Lambda)}(3) = 123.6, p < 0.0001\right)$, as illustrated in Fig. 7. Note that the coefficient-based *significances* of the individual components of the spline given in Table 3 are basically meaningless due to (expected) high fixed effect correlations (rs > 0.9). The total predicted effect of POSITION(MATRIX VERB) depicted in Fig. 7, however, is statistically unbiased (only the standard error estimates of collinear predictors are biased; coefficient estimates remain unbiased).

The effect of the matrix verb's position is mostly driven by a contrast between sentence-initial and other verbs, where the odds of *that* are more than six times lower for sentence-initial CCs compared to CCs that occur more than 10 words into a turn. Most of the effect is due to the contrast between the first four words and all other cases, but there are gradient effects beyond the first four words. It is therefore likely that the positional effect is largely due to epistemic uses of embedding verbs, which often occur sentence-initially (Thompson & Mulac, 1991a; Thompson & Mulac, 1991b). This possibility is explored in ongoing work.

Third, as observed by Thompson and Mulac (1991a) and Torres Cacoullos and Walker (2009), there is an effect of the matrix subject type on *that*-mentioning. Thompson and Mulac (1991a) 321 report that significantly lower *that* rates for cases with 1st or 2nd person matrix subjects (9–10% *that*-mentioning) compared to other matrix subjects (36%). Thompson and Mulac attribute this effect to the fact that these matrix subjects are often used in epistemic phrases. They find that 99% of all cases that are clearly epistemic (such, as "It's just your point of view, *I think*") occur with matrix subject "I" or "you" (Thompson & Mulac, 1991a, p. 321). Hence, following this argument, if a large proportion of the cases with matrix subject "I" or "you" in the database are epistemic, this would explain the low *that* rates for matrix subject "I" (13%) and "you" (26%) compared to other pronouns (32%) and lexical NPs (43%). Without

further stipulations, it would, however, not explain the overall correlation between the complexity of the matrix subject and *that*-mentioning illustrated in Fig. 8.

Post hoc comparison, coding MATRIX SUBJECT as ordered factor with four levels, *I* < *you* < other pronoun < other NP, confirmed the complexity of the matrix subject affects *that*-mentioning beyond the expected contrast between "I" and "you" against all other types of subjects (linear trend: $\beta = 0.66$, $z = 6.8$, $p < 0.0001$; quadratic trend: $\beta = -0.19$, $z = -1.8$, $p < 0.08$; cubic trend: $p > 0.2$). These data suggest that the complexity of the matrix subject affects *that*-mentioning beyond potential effects of grammaticalization. I return to the relation between UID and grammaticalization in Section 3.

2.6.5. Syntactic persistence and implicit learning—Finally, a not-quite marginal effect of primes with a complementizer is found ($\beta = 0.06$; $z = 1.6$, $p < 0.11$). This is due to the small number of CCs without a prime (see above) creating collinearity between the two levels of SYNTACTIC PERSISTENCE (fixed effect correlation $r = -0.48$). Post hoc tests that further removed collinearity found significantly higher *that* rates for primes with *that* ($z = 2.1$, $p < 0.04$). Hence, as in Ferreira (2003), the persistence effect associated with the less frequent structure is stronger, in line with implicit learning accounts of syntactic persistence (Bock & Griffin, 2000).

3. General discussion

The primary goal of this article has been to introduce and test a formalized account of efficient language production, the hypothesis of Uniform Information Density (UID). Based on information theoretic considerations, UID predicts that speakers prefer to distribute information uniformly across their utterances – to the extent that this does not clash with other constraints (e.g., grammatical constraints of English). While there is supporting evidence for UID from phonetic reduction discussed in the introduction, little to nothing was known about similar effects beyond lexical production. If UID affects language production at all levels of linguistic processing, speakers' syntactic choices should reflect a preference for uniform information density. For syntactic reduction, such as the case of *that*-mentioning studied here, speakers should exhibit a higher preference for the full version with *that*, the higher the information density at the complement clause onset. This is indeed observed.

Information density emerges as a strong predictor of *that*-mentioning in a representative sample of English speech, even after controlling for a large number of effects predicted by previous accounts of syntactic production. This result therefore constitutes evidence that syntactic reduction is affected by information density, consistent with the hypothesis that syntactic production is organized to be efficient.

Given that information is an inherently probabilistic notion, the observed sensitivity to information density suggests that syntactic production is probability-sensitive. That is, speakers' preferences during online production are affected by probability distributions (see also Chang et al., 2006; Gahl & Garnsey, 2004; Jaeger, 2006a; Jaeger & Snider, 2008; Resnik, 1996; Stallings et al., 1998). This links the results reported here to findings suggesting that syntactic comprehension is probability-sensitive (Hale, 2001; Hale, 2003; Jurafsky, 1996; Kamide et al., 2003; Levy, 2008; MacDonald, 1994; McDonald & Shillcock, 2003; Staub & Clifton, 2006; Trueswell, 1996). The current results are also compatible with the general architectural assumptions of connectionist accounts (for production, see e.g. Dell et al., 1999; Chang et al., 2006). Most existing accounts of syntactic production, however, neither consider language production to be probability-sensitive nor do they predict the observed effect of information density (Bock & Warren, 1985; Bolinger, 1972; Ferreira & Dell, 2000; Hawkins, 2004; Levelt & Maassen, 1981; Temperley, 2003).

Given its derivation from considerations about optimal communication, it is tempting to see UID in the tradition of rational cognition (Anderson, 1990) or bounded rationality (Simon, 1987). It is worth pointing out though that the rational cognition approach generally aims to test a stronger prediction than what has been tested here so far; rather than test whether human communication deviates from optimal communication, I have only provided evidence that language production exhibits signs of efficiency. It remains to be seen whether a stronger claim can be made given that language is subject to many other constraints (for example, language has to be learnable). Additionally, the current study has only tested the prediction that speakers prefer to distribute information uniformly. The Noisy Channel Theorem (Shannon, 1948), which UID has been derived from, would also lead us to predict that the rate of information transmission is close to the (as of yet unknown) channel capacity. Future work needs to address the link between information theory, in particular the Noisy Channel Theorem, and human communication in more detail.

UID and similar information theoretic approaches (e.g. Aylett & Turk, 2004; Genzel & Charniak, 2002; van Son & Pols, 2003) differ from the majority of work on language production in that these accounts are formulated at the computational level (in the sense of Marr (1982)). As such, they provide a strikingly different perspective on language production. It is therefore especially crucial to show that UID makes novel predictions that distinguish it from previous accounts (contrary to Sakamoto, Jones, & Love, 2008). The prediction about *that*-mentioning which has been tested here serves as an example. I introduce further predictions next, along with preliminary evidence. As with any new proposal, there are a number of open questions about UID, not all of which can be addressed here. I briefly discuss two of them: (1) The relation between UID and grammaticalization beyond the epistemic marker account mentioned above Thompson and Mulac, 1991b; and (2) the notion of the noisy channel evoked in the derivation of UID, in particular its relation to audience design (Brennan & Williams, 1995; Clark & Carlson, 1982; Clark & Murphy, 1982). I close with a discussion of the trade-offs of the corpus-based approach to research on language production.

3.1. Predictions of Uniform Information Density as a principle of efficient language production

As mentioned in the introduction, data from phonetic and phonological reduction patterns align with the predictions of UID. Speakers pronounce less information-dense instances of both syllables and words with shorter duration and less articulatory detail (Aylett & Turk, 2004; Bell et al., 2003, 2009; Gahl & Garnsey, 2004; Pluymaekers et al., 2005; Tily et al., 2009; van Son et al., 1998; van Son & van Santen, 2005). Crucially, these effects are not predicted by availability accounts (Levelt & Maassen, 1981; Ferreira, 1996; Ferreira & Dell, 2000). Availability accounts predict that speakers lengthen a word's duration (and hence possibly its articulatory detail) if the *following* word is hard to retrieve (e.g. because the following word carries a lot of information). Interestingly, such effects of following material have also been observed (e.g. Fox Tree & Clark, 1997), suggesting independent effects of UID and availability-based production. However, previous studies have failed to control for both effects simultaneously, raising the question as to whether either effect is confounded. Future work will need to tease apart these two hypotheses about phonetic reduction.

More generally, UID predicts that reduction of any type is affected by information density. For example, English and languages across the world provide a plethora of reduction phenomena at various levels of linguistic processing (to name a few: optional clitics, optional case-markers, copula drop and reduction, choice of referential expression, argument or adjunct drop, and ellipsis; for further examples, see Jaeger, 2006a, pp. 8151–8152). For all of these types of reduction, UID predicts that speakers show a higher preference for the full form, the more information it carries (as long as both variants are permitted by the grammar of the language

and as long as they convey the same message). To illustrate this prediction, I discuss a few examples spanning different types of reduction alternations. Some of these phenomena have received a lot of attention in the psycholinguistics literature (e.g. syntactic reduction) while very little quantitative data exists for others.

Consider, for example, the phenomenon of optional case-marking in languages like Japanese (Fry, 2003) and Korean (Lee, 2006). These languages have relatively flexible word order compared to English. Case-marking is used to convey the grammatical function of expressions, but in informal speech case-marking is often optional. UID predicts that speakers should be more likely to omit optional case-markers if the case-marker contains less information (i.e. if the marked expression's grammatical function is highly probable, and hence low in information, in its context). While this prediction has not been tested directly, corpus studies on optional case-marking have found, for example, that object expressions with properties that are most typical to grammatical objects (e.g. inanimate, indefinite noun phrases) are less likely to be explicitly marked as object than object expressions with atypical object properties (e.g. animate, definite noun phrases; Fry, 2003; Lee, 2006). UID predicts that this tendency is at least in part due to predictability of the grammatical function given the properties of the noun phrase: The more predictable the grammatical function of an expression, the less likely speakers should be to case-mark the expression (recall that Japanese and Korean have more flexible word order than, for example, English, so that word order alone does not determine grammatical function). Similar reasoning applies to optional object clitics and other types of argument-marking morphology in head-marking languages (e.g. Bulgarian direct object clitics, Avgustinova, 1997; Jaeger & Gerassimova, 2002; resumptive morphology in Yucatec Mayan, Norcliffe, 2009). Availability-based production would not predict such a pattern, though ambiguity avoidance accounts could be taken to make the same prediction as UID.

In this context, it is interesting to consider the predictions of UID about another type of morpho-syntactic choice point. In many languages, including English, speakers can produce auxiliaries in a contracted form (e.g. *he's* vs. *he is*). UID predicts that the more information the contractible element (e.g. a form of BE) contains, the more likely speakers should be to use the full form. Crucially, neither availability nor ambiguity avoidance accounts make this prediction. Preliminary evidence supporting UID comes from several studies on morpho-syntactic contraction reported in Frank and Jaeger (2008): We found that speakers are more likely to choose the full form (e.g. *he is*) over the contracted form (e.g. *he's*) when the form encodes a lot of information. This result was obtained while controlling for potential effects of the availability of material following the contractible element.

The predictions of UID should also be tested against other syntactic reduction phenomena. In line with UID, Wasow, Jaeger, and Orr (in press) found that speakers are less likely to mention relativizer *that* in non-subject-extracted relative clauses (e.g. *I like the way (that) it vibrates*) when the relative clause is predictable given lexical properties of the noun phrase. For example, definite noun phrases are more likely to be modified by a relative clause than indefinite noun phrases, and noun phrases with light head nouns (e.g. *the way*) are more likely to be modified by a relative clause than noun phrases with heavy head nouns (e.g. *the priest*; see also Fox & Thompson, 2007). Estimates of the information density at the relative clause onset based on these cues correlate in the predicted direction with speakers' preference to produce the relativizer (see also, Jaeger, 2006a Study 2, which controls for a variety of effects known to affect relativizer mentioning, including availability-based production).

In Jaeger (submitted for publication), I investigate a related but different type of syntactic reduction, so-called *whiz*-deletion, in written British English. *Whiz*-deletion refers to the optional reduction of, for example, passive subject relatives, as in *The smell (that is) released by a pig or chicken farm is indescribable*, where the relativizer and auxiliary can be omitted

together. In line with UID, writers prefer the full variant with the relativizer and auxiliary when the relative clause onset carries a lot of information (Based on probability estimates derived from the British National corpus). Further studies are necessary to see whether this effect holds for speech and whether data from other languages and other types of syntactic reduction (e.g., *to*-mention after the verb *help* in English, Rohdenburg, 2004) follow the predictions of UID.

It is possible that speakers manage syntactic information density beyond these cases of optional mentioning of function words. Speakers often have a choice between explicitly mentioning arguments (Fillmore, 1968; Resnik, 1996) or adjuncts (Brown & Dell, 1987). UID predicts that the less probable argument and adjunct expression are given preceding context, the more likely speakers are to mention them. This is consistent with evidence that speakers are more likely to mention less typical instruments (Brown & Dell, 1987 e.g. *stabbing with a knife* vs. *an ice pick*). There are, however, alternative explanations for this finding. It is possible that speakers simply are more likely to mention less typical instruments because they are more aware of them.

A slightly stronger piece of evidence comes from a corpus study on 'object drop', as in *Germany lost (the last semi-finals)* (Resnik, 1996). Resnik shows that verbs with a high 'selectional preference', i.e. verbs that contain a lot of information about the distribution of (the semantic classes of) their objects, are on average more likely to omit their objects. An example of a verb with high selectional preference is *eat*. An example of a verb with comparatively low selection preference is *see*. Hearing *Tom ate …* tells us more about the type of objects that we can expect than hearing *Tom saw ….* Specifically, Resnik shows that the relative entropy of a verb *v* with regard to the semantic classes of its objects c_i, $S = \sum_i P(c_i|v)$ 'log $\frac{p(c_i|v)}{p(c_i)}$ is correlated with the verb's rate of object drop (Resnik, 1996, pp. 149-151). This finding is expected given UID, although it constitutes less direct support than the study presented above. Selectional preference is a measure of how much information a verb contains about its objects *on average*, rather than the information of the *actual* object. So, rather than showing that low-information (highly predictable) objects are more likely to be dropped, Resnik's study provides evidence that verbs that generally contain more information about their object also generally have a higher rate of object drop.

Similarly, UID makes predictions about the choice between different referential expressions that refer to the same referent. For example, speakers should be more likely to produce pronouns (e.g. *she*) instead of full noun phrases (e.g. *the girl*) when reference to the expression's referent is probable in that context (for preliminary evidence from referential choice in written language, see Tily & Piantadosi, 2009). Beyond reduction phenomena, UID predicts that speakers should prefer word orders that distribute information more uniformly where grammar permits. In short, UID makes many novel predictions across different levels of linguistic production, across languages, and across different types of alternations. While much work remains to be done, there is preliminary evidence that at least some of these predictions seem to be met.

I close with a brief summary of a puzzling finding that has so far not received a psycholinguistic explanation. Genzel and Charniak (2002) tested whether the distribution of information across discourses follows the hypothesis that linguistic communication is efficient. Recall that Shannon information is defined as the logarithm-transformed inverse of its probability. Entropy is the *expected* (or average) Shannon information, $\Sigma p(word) * I(word) = -\Sigma p(word) * \log p(word)$. If linguistic communication is efficient, the per-word entropy of sentences should stay constant throughout discourses. Unfortunately, a direct test of this constant entropy rate hypothesis is challenging. Adequate estimates of a word's probability and hence a word's information *in its discourse context* require a discourse model of the speakers' state of knowledge given the entire preceding context as well as other knowledge the speaker has access

to (e.g. world knowledge). This makes automatic discourse modeling a notoriously hard problem. Recognizing this technical problem, Genzel and Charniak (2002) proposed an indirect test of the constant entropy rate hypothesis. If the real per-word entropy based on all preceding context, world knowledge, et cetera is constant throughout discourse, estimates of *a priori* per-word entropy that ignore inter-sentential information should increase throughout discourse. Intuitively, words with *a priori* high information (words that are improbable if only their sentential context is considered) should tend to occur later in the discourse than words with low information. The reasoning behind this prediction is that, *on average*, more preceding context makes words more predictable, thereby decreasing their information-content. The indirect prediction of the constant entropy rate hypothesis is supported by evidence from written productions in a variety of languages (Genzel & Charniak, 2002; Genzel & Charniak, 2003; Keller, 2004; Qian & Jaeger, 2009; Qian & Jaeger, submitted for publication). Indirect evidence has also been observed for spoken language (for English, Piantadosi & Gibson, 2008; for Mandarin Chinese, Qian, 2009). While these results are unexpected from the perspective of most accounts of language production, they are highly compatible with the hypothesis of Uniform Information Density. UID predicts that the observed entropy profiles are the combined result of lexical and syntactic choices (for further discussion, see Qian & Jaeger, submitted for publication).

3.2. Uniform Information Density and grammaticalization

There are several ways in which grammaticalization could contribute to the observed patterns of *that*-mentioning. As discussed above, several of the findings provide evidence for the proposal by Thompson and Mulac (1991b) that some matrix clause onsets have become grammaticalized as epistemic markers. There is, however, also evidence that the observed effects of information density cannot be reduced to epistemic markers. If the multilevel logit analysis is repeated on the subset of the data excluding epistemic markers (leaving 3,033 cases), the effect of information density remains highly significant regardless of whether verb lemma was included as random effect ($\beta = 0.41, z = 2.7, p < 0.008$) or not ($\beta = 0.56, z = 113, p < 0.0001$). Here, cases were considered epistemic markers if (a) the matrix subject was *I* or *you*, (b) the matrix verb lemma was either *guess*, *think*, *say*, *know*, or *mean*, (c) the matrix verb was in the present tense, and (d) the matrix clause was not embedded (i.e. the matrix clause was a main clause). A similar analysis excluding all cases with first and second person singular matrix subjects comes to the same conclusion.

Beyond the specific proposal of Thompson and Mulac, there are other grammaticalization accounts compatible with the data. More generally, just like with any other processing effect, the effect of information density could be partially or completely grammaticalized. For example, Bresnan and Hay (2006) provide evidence that accessibility effects on the ditransitive alternation (*He gave her a book* vs. *He gave a book to her*) seem to be at least partially grammaticalized: While the general pattern is observed in both New Zealand and American English (speakers prefer to order expressions referring to animate referents before expressions referring to inanimate referents, cf. Prat-Sala & Branigan, 2000), the strength of the effect differs between the languages. Bresnan and Hay (2006) observe these differences while controlling for many other effects known to affect speakers' preferences in the ditransitive alternation. They also find that the effect of animacy in New Zealand English has changed over time. Bresnan and Hay propose that the animacy effect is grammaticalized as part of gradient grammatical representations (as postulated in, for example, Stochastic Optimality Theory, Boersma & Hayes, 2001).

It is possible that most of the processing effects observed in the current study, including the effect of information density, have become grammaticalized. What grammaticalization accounts leave unanswered though is *why* patterns in alternations become grammaticalized in

one way or another. For *that*-mentioning, Thompson and Mulac (1991b) propose that matrix onsets that are frequently used with complement clauses over time become grammaticalized without the complementizer *that*. This itself is hardly an explanation. Bybee and Hopper (2001) propose that grammaticalization involves reduced forms because of "automatization of neuro-motor sequences which comes about through repetition" (for similar hypotheses that link grammaticalization of reduced forms to frequency of use, see also Bates & MacWhinney, 1982; Bybee, 2002; Givón, 1995; Langacker, 1991; Traugott, 1995). Such 'training' effects may be intuitive for *phonetic* reduction (the case discussed in Bybee & Hopper, 2001), but it is less clear how an automatization account would extend to *syntactic* reduction and, specifically, the complete *omission* of words. More crucially though, the hypothesis that automatization results in reduction (rather than particularly *clear* articulation) seems to imply some notion of efficiency. The information theoretic principles that UID is derived from provide such a motivation: More probable linguistic forms encode less information and speakers aim to distribute information uniformly. In other words, UID may be one reason why epistemic phrases correlate with lower rates of *that*-mentioning.

3.3. The noisy channel

The derivation of UID presented in the introduction to this paper is based on the assumptions that language use involves communication through a noisy channel with limited bandwidth (cf. Shannon, 1948). Uniform Information Density is then provably optimal in terms of the amount of information successfully transferred.

One intuitive interpretation of the noisy channel assumption is that the 'noisy channel' refers to the channel between interlocutors. It is then tempting to conclude that UID is a form of audience design. In its most general interpretation, audience design (Brennan & Williams, 1995; Clark & Murphy, 1982; Clark & Carlson, 1982) refers to the idea that speakers consider their interlocutors' knowledge and processor state, presumably to improve the chance of successfully achieving their goals (including, but not limited to, the transmission of information). According to this interpretation of UID, speakers would avoid high information stretches in their production, so as to facilitate successful comprehension. Speakers should estimate information density based on their understanding of their interlocutors' expectations. In other words, speakers should estimate information density based on *only* the cues available to their interlocutors - a prediction not tested in the current study.

This may seem to conflict with recent studies, which have generally failed to provide clear evidence for syntactic audience design (e.g. Arnold, Wasow, Asudeh, & Alrenga, 2004; Kraljic & Brennan, 2005; Roland et al., 2005, 2007; though see Hawkins, 2004; Haywood, Pickering, & Branigan, 2005; Temperley, 2003). Note, however, that these studies focused on *ambiguity avoidance*, more specifically, the avoidance of temporary syntactic ambiguity. Speakers are rarely faced with the need and possibility to avoid long-lasting garden paths (i.e. temporary ambiguities that remain unresolved over a long stretch of linguistic input and where comprehenders are *likely* to be strongly biased towards the unintended parse). Recall, for example, Table 4 above, which shows that long-lasting garden paths are extremely rare for complement clauses (for comparable observations about relative clauses, see Jaeger, 2006a, Section 6.3.1). Syntactic ambiguity avoidance might simply not be the best place to look for effects of audience design (for related discussions, see Ferreira, 2008; Wasow & Arnold, 2003). Choice points, on the other hand, where speakers can choose between several variants that differ in information density are ubiquitous in production (see above). Modulating information density at such choice points would therefore be a rational strategy for efficient production.

The evidence provided here is, however, also perfectly compatible with an interpretation of UID that does not make reference to audience design. First, it is possible that speakers use their

information density estimates based on their own perspective as an approximation. Given how closely aligned speaker and comprehenders are in normal conversation, these estimates may be sufficient to achieve efficient communication (most of the cues available to speakers are also available to comprehenders). It is also possible that the information density effects have become partially grammaticalized (see previous section). In either case, UID would not be considered audience design in the strictest sense.

Are there yet other interpretations of UID that do not require reference to a noisy channel between interlocutors? A *priori*, the information theoretic proofs that UID is derived from do not require the noisy channel to be external to the speaker. Consider, for example, a standard production model according to which speakers' intended messages are encoded linguistically via several stages (e.g. Bock & Levelt, 1994; Levelt, 1989). Messages may need to be structured into macro-propositions (Brown & Dell, 1987), functional structures need to be derived, lemmas need to be selected, syntactic and phonological information needs to be retrieved, and so on. Regardless of whether these processes are clearly delineated stages or interacting processes, information needs to be passed from one process to another. Any time information is passed, it is passed through a channel. Given that we are considering a biological system, these channels are arguable noisy. In that sense, producing a sentence involves information transfer through many noisy channels. Regardless of whether the channel is between interlocutors or within speakers, sending information at a rate uniformly close to the channel capacity will -in theory- allow arbitrarily small error rates (Shannon, 1948).

While such an interpretation of UID is in theory possible, it is less intuitive. In particular, it is unclear how the observed effect of information density on *that*-mentioning would be accounted for. Why would producing a complementizer *that* distribute information more uniformly with regard to information transferred from one stage in production to another? What is the encoded information that would be spread our more uniformly by uttering *that*? The most straightforward answer would seem to be that the encoded messages is something akin to a control signal, initiating retrieval of a complement clause structure. In that interpretation, producing the complementizer *that* basically serves as a self-priming cue. The account we have arrived at may be considered an availability account in the broadest senses, though it differs from previous proposals in that it takes availability to be primarily driven by information density rather than accessibility (Ferreira & Dell, 2000; Race & MacDonald, 2003). The production-based interpretation of UID shares with the availability proposal made in Race and MacDonald (2003) that it predicts that producing the complementizer facilitates production. In other words, producing *that* should correlate with higher fluency. This prediction does not, however, seem to be supported (Jaeger, 2005). Another problem of the outlined production-based interpretation of UID is that it fails to account for the findings from morpho-syntactic reduction (Frank & Jaeger, 2008) and object drop (Resnik, 1996) reported above.

I tentatively conclude that the noisy channel referred to in the derivation of UID is situated between interlocutors. Whether speakers consider their interlocutors' perspective when estimating information density is an empirical question that remains for future research.

So far, I have limited myself to discussions of UID that depend on the Noisy Channel Theorem. There are, however, alternative derivations of the theoretical optimality of uniformly distributed information density that do *not* require reference to a noisy channel. For example, in Levy and Jaeger (2007), we showed that uniform information density would also minimize processing difficulty, if processing difficulty is superlinearly related to the information content (= surprisal) of words. While this interpretation seems to be disfavored by recent findings that suggest a linear relation between surprisal and processing difficulty in comprehension (Smith & Levy, 2008), it is still possible that processing resources are consumed superlinearly dependent on information density. Superlinear dependence on information density might be

expected from any system that has access to *limited* resources (such as memory). Future will be necessary to test these alternative explanations for the observed effects of information density.

3.4. Trade-offs of corpus-based research on language production

The findings presented in this article are derived from a corpus-based study rather than a psycholinguistic experiment. While there is a rich tradition of corpus-based work on language production (to name just a few, Bresnan et al., 2007; Clark & Fox Tree, 2002; Fox Tree & Clark, 1997; Hawkins, 2001; Lohse, Hawkins, & Wasow, 2004; Race & MacDonald, 2003; Resnik, 1996; Roland et al., 2005, 2007; Temperley, 2003; Wasow, 1997), a number of objections have been raised against the corpus-based approach. Among the reasonable objections to corpus-based work is the claim that the lack of balance and the heterogeneity of the data make corpus-based findings unreliable.[5] While corpus-based psycholinguistics without doubt faces complex statistical challenges, theoretical and methodological advances over the years have addressed many them.

Modern regression methods facilitate simultaneous assessment of multiple effects (e.g. Bell et al., 2003; Pluymaekers, Ernestusb, & Baayen, 2005; Race & MacDonald, 2003; Roland et al., 2005; Tagliamonte & Smith, 2005; Torres Cacoullos & Walker, 2009). The additional power and the inclusion of random effects of multilevel models have further helped to address some of the common statistical concerns with the corpus-based approach (Bresnan et al., 2007; Jaeger, 2006a).

If employed appropriately and conservatively, multilevel models account for random differences between speakers and provide statistical control approaching that of balanced experiments. In particular, great care was taken in the current study to achieve interpretable predictor coding. Wherever possible inclusion of redundant predictors was avoided (this contrasts with the approach taken in Roland et al., 2005). Where predictors were inherently redundant (e.g. because there were several measure of the same underlying processes, as in the case for fluency, or because two competing accounts made partially overlapping predictions), collinearity was reduced via residualization or via non-redundant variable coding (in contrast to, Roland et al., 2005; Tagliamonte & Smith, 2005; Torres Cacoullos & Walker, 2009; but see Baayen et al., 2006). This was done since collinearity makes it hard to assess effect *direction*, which is usually what psycholinguistic accounts make predictions about (the existence of effect can still be assessed reliably by means of model comparison). Additionally, the inclusion of a random speaker intercept provided a measure of inter-speaker variance in *that*-mentioning. For the current study, speaker differences do *not* seem to account for much variance in *that*-mentioning (partial Nagelkerke $R^2 = 0.01$ out of 0.34) and most - though not all - previous findings replicated.

While admittedly making data analysis more complex, the approach taken here has important advantages that should warrant the additional complexity. Some of these advantages are due to the use of naturally distributed spontaneous speech data, while others are due to the use of multiple regression for analysis. The most immediate advantage of working with data from spontaneous speech is ecological validity (cf. Clark, 1996; Dhami, Hertwig, & Hoffrage, 2004). The database used for the current study is arguably a more representative sample of

[5]Note that it is only for the sake of simplicity that I discuss corpus-based and experiment-based approaches as a dichotomy. Any experiment yields a corpus of data. While experiments typically provide balanced data sets, that is by no means always the case. Unbalanced data can result from high exclusion rates or from designs aimed at eliciting more natural speech. Behavioral paradigms also differ in terms of the typical amount of lexical and structural heterogeneity they elicit. Similarly, balanced designs can be combined with a corpus-based approach (Clark & Wasow, 1998, p. 211). In terms of heterogeneity as well as balance, different behavioral paradigms and corpus-based approaches form a continuum rather than a dichotomy.

English speech than productions elicited in experiments with stimuli drawn from a restricted set of lexical and structurally similar complement clauses. This makes it more likely that the results obtained here extend to all of English. Ecological validity is particularly relevant here since the hypothesis investigated is derived from considerations about communication.

The regression approach allows for the simultaneous theory-driven test of multiple hypotheses. In addition to controlling for the predictions of alternative accounts, the inclusion of multiple controls in the current study proved insightful in at least three ways. First, it is reassuring that many previous results replicate on natural data, lending support for the model overall (e.g. effects of dependency length, Bolinger, 1972; Elsness, 1984; Hawkins, 2001, 2004; accessibility effects, Elsness, 1984; Ferreira & Dell, 2000; Race & MacDonald, 2003; syntactic persistence Ferreira, 2003; grammaticalization effects Thompson & Mulac, 1991b). These results also provide evidence that processing mechanisms are the major driving force behind *that*-mentioning - a standard assumption in the psycholinguistic literature, but by no means outside the field. For example, according to some linguistic analyses *that*-mentioning is not an alternation at all, but rather determined by semantic differences between complement clauses with and complement clauses without *that* (Bolinger, 1972; Dor, 2005; Kaltenböck, 2006; Yaguchi, 2001). Such accounts provide no explanation for the observed effects of fluency, accessibility, and syntactic persistence (for further discussion of semantic accounts, see Jaeger, 2006a, Section 6.2.2). Second, the results provide evidence for several relatively understudied effects (e.g. effects of similarity avoidance, Walter & Jaeger, 2008; Jaeger, in press; fluency and speech rate effects on syntactic reduction, Jaeger, 2005; Roland et al., 2007). Third, the results speak to unresolved debates in psycholinguistics (e.g. ambiguity avoidance, Bolinger, 1972; Hawkins, 2004; Temperley, 2003).

Multiple regression also facilitates effect size comparisons. Understanding not only whether an effect exists but also how much it contributes to the observable behavior is an important part of understanding the complex interactions of multiple mechanisms during language production. For example, there is clear evidence for accessibility effects on *that*-mentioning, but compared to the amount of attention these effects have received in the literature (Ferreira & Dell, 2000; Jaeger & Wasow, 2006; Race & MacDonald, 2003; Roland et al., 2005; Temperley, 2003), their influence on the overall distribution of complementizer *that* is relatively small (see Table 3).

Working with naturally distributed data illustrates that the term ‘effect size’ needs to be used with caution. How much an effect contributes to our understanding of observable behavior does not only depend on its strength, but also on how often the relevant condition is present. Consider, for example, the discussion of ambiguity avoidance in the result section. The results suggest that speakers *may* avoid ambiguity in syntactic production, but they also tell us something else. Complement clause reduction may simply not be the right place to look for ambiguity avoidance. In natural speech, complement clauses rarely ever seem to contain the potential for prolonged temporal ambiguities with high risk of ‘garden-pathing’. In other words, there may not be much of a need to avoid ambiguities in complement clauses. Similar studies on other phenomena that have been taken to argue for or against ambiguity avoidance may reveal that such garden-path causing ambiguities are overall extremely rare in natural speech (cf. Jaeger, 2006a).

The final advantage of naturally distributed data I wish to discuss is arguably the most crucial one. As more and more research finds that speakers and listeners are sensitive to probability distributions (for comprehension, Hale, 2001; Jurafsky, 1996; Levy, 2008; MacDonald, 1994; McDonald & Shillcock, 2003; Trueswell, 1996; among many others; for production, Bell et al., 2003; Bell et al., 2009; Gahl & Garnsey, 2004; Stallings et al., 1998, among others; as well as the current study), it may be necessary to revisit the standard assumption that

experiments with balanced designs are the best way to study language production and comprehension. Consider what it means to be a participant in an experiment with a balanced design. Participants are put into a situation where words and syntactic structures (co-)occur in (uniform) distributions that do not match participants' expectations based on previous experience with naturally distributed data. It is well-known that language users can implicitly learn statistical distributions with little exposure (e.g. Saffran et al., 1999). There is also evidence that language users rapidly adapt to changes in distributions (e.g. Clayards et al., 2008; Wells et al., 2009).

In fact, one of the most widely used experimental paradigms, syntactic priming (Bock, 1986; for a recent overview of the literature, see Pickering & Ferreira, 2008), is based on the observation that recent exposure changes speakers' behavior. Assuming that we are not exclusively interested in language processing during such adaptation, corpus studies (and experiments with unbalanced designs resembling natural distributions) offer a crucial advantage to researchers interested in language production. Given these advantages, it is my hope that the current study helps to further establish corpus-based research as a viable complement to psycholinguistics experiments.

4. Conclusions

Based on data from *that*-mentioning in spontaneous speech, I have presented a test of several sentence production accounts within one single regression analysis. The analysis provided both replicating and novel evidence for availability-based accounts (Levelt & Maassen, 1981; Ferreira, 1996; Ferreira & Dell, 2000), dependency processing accounts (Hawkins, 2001, 2004) and effects of grammaticalization (Thompson & Mulac, 1991b), as well as weak evidence for a revised ambiguity avoidance account. Most crucial to the purpose of this paper, the predictions of the hypothesis of Uniform Information Density were confirmed.

Uniform Information Density states that the production system is organized so that speakers prefer to encode their intended message by distributing information uniformly across their utterances at a rate close to, but not exceeding, the channel capacity. As a consequence of this, speakers should prefer utterances that avoid peaks and troughs in information density (within the bounds defined by grammar). This correctly predicts that speakers should prefer to produce complementizer *that* whenever the complement clause onset would otherwise carry too much information, because inserting the complementizer spreads part of the information that would otherwise be carried by the complement clause onset over more words and hence over more linguistic signal. More generally, speakers should show a preference to produce optional linguistic forms whenever this reduces the information density of upcoming (or possibly even past) material that would otherwise have high information density.

The hypothesis of Uniform Information Density makes predictions about speakers' preferences at choice points at all levels of production. I have outlined the predictions for several such choice points and discussed preliminary evidence in support of the Uniform Information Density hypothesis. If further studies confirm speakers' preference to distribute information uniformly, this would be evidence that speakers are rational (Anderson, 1990) or at least boundedly rational (Simon, 1987).Communication through a noisy channel is most efficient if information is transmitted at a uniform rate close to the channel capacity (Shannon, 1948). The observed preference for Uniform Information Density on spontaneous speech is hence compatible with the hypothesis that language production is organized to transmit information efficiently.

Appendix A

Summary of verb biases

Table A.1 lists all verb lemmas in the database along with their frequency, associated probability of a complement clause (CC-bias), and the number of instances in the final database. Two estimates for verb frequency and hence for CC-bias can be derived (recall that $p(\text{CC} \mid \textit{matrix verb lemma}) = f(\text{CC}, \textit{matrix verb lemma}) = f(\textit{matrix verb lemma})$). The first estimate counts all matches to any inflected form of the verb lemma (word count). The second estimate only considers matches that were marked as verb in the corpus (verb count). The results presented in this paper are based on verb counts since this resulted in a marginally better model, but none of the results changes qualitatively, if word counts are used instead.

Note that (a) some verb lemmas were rarely observed, making their estimates unreliable, (b) some verbs occur rarely with a CC, making their *that*-bias estimate unreliable (the analyses with random effects for verb lemmas presented above addresses this issue) and (c) the table marginalizes over all other variables correlating with *that*-mentioning, but at least some of these variables are correlated with verb lemma. That is, it would be misleading to generalize based on the numbers in Table A.1 since they do not take into account the partial effects of other predictors.

Several points deserve mention. In Table A.1, repetitions of the verb due to disfluencies, as in "*I think, I think I can do this*", were counted as several instances of that verb. Separate analyses confirmed that the results replicate, if repetitions due to disfluency are excluded from verb counts. Second, two verbs in the database can only take a CC in very specific environments: *take* occurs with CCs in combination with it, as in *Um, so, I take it that you like to ski*, or as *take into account*; similarly, *thank* occurs with CCs only in combination with *god*, as in *I thank god every day that I have the resources that we have tapped into*. That is, when these verbs occur with CCs, the CCs are actually rather predictable, which is not captured by estimate of CC-bias used in this paper. There are several other verbs for which the true CC-bias is likely to be underestimated as well: *guess, mean, think*, and *know* occur very frequently as parentheticals, as in *He is, you know, from Mars*. The estimate of CC-bias employed in this paper does not capture that such parenthetical uses are likely to differ prosodically from CC-embedding occurrences of these verbs. The true information carried by a CC-onset in context is likely to be determined by these and other cues.

Table A.1

Alphabetically sorted verb lemma, along with verb lemma frequency (Frequency), the number of instances in the final database, subcategorization bias for a complement clause (CC-bias), and its *that*-bias. Both counts based on word forms and counts restricted to terminals part-of-speech tagged as verbs are given.

Verb lemma	Frequency in corpus		CCs	CC-bias		*that*-bias
	Word count	Verb count	In database	Word-based	Verb-based	
Agree	263	263	27	0.10	0.10	0.74
Believe	288	287	117	0.42	0.42	0.38
Bet	120	116	51	0.43	0.45	0.02
Consider	88	87	6	0.07	0.07	0.33
Decide	196	195	52	0.29	0.29	0.40
Expect	89	82	5	0.06	0.06	0.40
Feel	686	675	108	0.16	0.16	0.59

Verb lemma	Frequency in corpus		CCs	CC-bias		*that*-bias
	Word count	Verb count	In database	Word-based	Verb-based	
Figure	158	146	42	0.29	0.32	0.14
Find	688	688	119	0.18	0.18	0.68
Guess	1590	1590	917	0.61	0.61	0.01
Hear	602	601	72	0.13	0.13	0.43
Hope	179	167	103	0.58	0.62	0.20
Imagine	124	123	36	0.31	0.31	0.25
Know	12,386	12,377	559	0.05	0.05	0.32
Mean	2280	2251	43	0.02	0.02	0.28
Notice	104	100	26	0.27	0.28	0.27
Read	619	573	8	0.02	0.02	0.50
Realize	127	127	54	0.43	0.43	0.48
Remember	317	317	39	0.14	0.14	0.10
Say	2076	2000	517	0.26	0.27	0.27
See	2401	2153	55	0.02	0.03	0.65
Show	326	130	8	0.02	0.06	0.62
Suppose	71	62	35	0.54	0.61	0.06
Take	834	833	7	0.01	0.01	0.29
Teach	126	106	2	0.02	0.02	0.50
Tell	545	544	113	0.21	0.21	0.39
Thank	71	71	3	0.04	0.04	0.33
Think	5669	5622	3465	0.64	0.64	0.11
Understand	250	238	22	0.09	0.10	0.64
Wish	118	118	101	0.87	0.87	0.14
Worry	118	97	4	0.03	0.04	0.75

Appendix B

Data extraction

TGrep2 (Rohde,2005) and the Tgrep2 Database Tools (Jaeger, 2006b) were used to extract all and only structures of interest from the Paraphrase Stanford-Edinburgh LINK Switchboard Corpus (Bresnan et al., 2002; Godfrey et al., 1992; Marcus et al., 1999; for an overview of the available annotations, see Calhoun, 2006). The TGrep2 search pattern for CCs is given in (9). Fig. B.1 illustrates most of the structural constraints expressed by the TGrep2 pattern in form of a tree diagram. The first line of the search pattern defines a subordinate clause SBAR (for *S*) that follows and is the sister of a verb (all verbs in the Treebank are marked by a part-of-speech tag that starts with VB). This restricts the matches to complement clauses to verbs (rather than nouns or adjectives). The first line of the pattern also correctly excludes parenthetical uses of some verbs, such as *I mean and you know*, as in (8a) and (8b) respectively, since such parentheticals are marked differently in the Penn Treebank (Meteer et al., 1995).

(8) **a.** How do you feel about them, I mean, since you've kind of been close to that.

b. Situations like that you don't realize, you know, until you start thinking about it …

The second line states that the clause must directly dominate either a complementizer *that*, which is marked by the part-of-speech tag IN, or a -NONE- node which marks the absence of a complementizer. This excludes all kinds of other complement clauses, such as interrogative complement clauses and infinitival complement clauses. The third line of the search pattern excludes CCs with subject gaps since they cannot occur with a complementizer (Huddleston & Pullum, 2002, p. 953; for an empirical confirmation, see Jaeger, 2006a, Appendix B.2). The fourth line excludes extraposed CCs (in the corpus used here, extraposed CCs are marked with an N node governing an alphanumeric index to the position that they are extraposed from). The last line excludes some *whether*-clauses, free relatives, etc. that otherwise would be included due to irregularities in the Treebank annotation. Note that the pattern includes both CCs adjacent to the matrix verb and CCs that are preceded by other phrases that follow the matrix verb.

(9) TGrep2 search pattern for CCs

```
/Λsbar/$,, /Λvb/
        [< (in< “that”) |< “-none-”]
        <(/Λs/ < (/-sbj/ !< “-none-”))
        !< (/Λn/ < ! /Λ ’ 0,1[a-za-Z]+.*/)
        !</Λwh/
```

Using the pattern in (9), 7369 complement clauses were extracted from the corpus. Manual inspection of all cases with unusual matrix verbs (e.g. *let, be, end up*, etc.) revealed 144 erroneously included cases. These include complement clauses to adjectives, as in ‘… [*it*] *was funny* [*that they kept pushing the three eighty-six* [*price*]’, as well as adjunct clauses, as in ‘… *it ends up* [*the best is to take a high deductible* …]’. Those cases were removed from the database, resulting in 2.0% data loss. As mentioned above, it has been claimed that not all matrix verbs are compatible with complementizer omission. For example, CCs to non-bridge verbs like manner of speech verbs such *as sing, whisper*, etc. supposedly cannot occur without a complementizer. If so, those verbs should not be included in the database. The average complementizer rate was calculated for each verb lemma in the database. There were 50 verbs that are associated with a complementizer rate of 0% or 100% in the database. All of these verbs occur fewer than seven times in the database. This may mean that the true complementizer rates associated with these verbs are different from 0 or 100%. In any case, these cases were excluded from the analysis, resulting in the removal of 71 cases (1.0%). Further exclusions due to missing information about control variables or due to low-count observations are described in the methods section. Overall, 8.5% of the extracted data were excluded prior to statistical analysis.

Appendix C

Additional analyses

This appendix provides a summary of additional analyses conducted to address potential concerns that the observed effects of information density could be driven by outliers. I present two post hoc analyses and three meta-analyses of data from production experiments (Ferreira & Dell, 2000).

First, I replicated the multilevel logit analysis using an alternative statistical approach, bootstrapping with random cluster replacement (e.g. Feng, McLerran, & Grizzle, 1996). Instead of modeling properties of clusters in the data (such as subjects and verb lemmas) explicitly, the boostrapping approach fits an ordinary logistic regression model repeatedly over data that was randomly sampled with replacement from the original data. Crucially, the

sampling takes place over entire clusters of data (e.g. all instances including a particular matrix verb) rather than over individual cases. Since the sampling takes place with replacement, a given data set could theoretically consist of only repeated instances of the same cluster (e.g. only cases with the same matrix verb). The final model is then derived by summarizing the models that were fit to the different samples. Hence, bootstrapping with random cluster replacement provides an alternative method to correct for potentially deflated standard errors (and hence anti-conservative *p*-values) that are due to violations of the assumption of independence. Using the Design package (Harrell, 2007) in R, an ordinary logistic regression model with the same predictors as the multilevel logit analysis (but no random effects) was bootstrapped 20,000 times over verb lemma clusters. The result replicates the multilevel model analysis: Information density is still a significant predictor ($\beta = 0.42$, $z = 2.3$, $p < 0.03$; 1 out of 20,000 samples failed to result in a fit). Hence the results reported above do not depend on the particular analyses chosen.

Second, the multilevel logit analysis was repeated after excluding all cases with one of the five most frequent matrix verbs, *guess*, *think*, *say*, *know*, or *mean*. Unfortunately, this left only 1215 cases, less than 20% of the original data. This means that the limited sample size (the less frequent outcome) is dangerously small given the number of parameters in the analysis. Even for balanced data, it is usually suggested to have at least 10–15 times as many data points as parameters in the model to avoid overfitting (Babyak, 2004; Peduzzi, Concato, Kemper, Holford, & Feinstein, 1996). Reducing the data set, of course, also drastically reduces the statistical power of the analyses. Indeed, the effect of information density lost significance ($p = 0.15$), although the effect is still in the expected direction. Note that the loss of significance is not due to the exclusion of epistemic cases. Recall also the post hoc analyses reported in Section 3 on grammaticalization. After removal of epistemic cases, the effect of information density remained significant. This argues that the effect cannot be reduced to the grammaticalization of a few types of highly frequent matrix clause onsets as epistemic markers (Thompson & Mulac, 1991b).

Finally, I conducted a meta-analysis of Experiments 1, 2, and 4 on *that*-mentioning presented in Ferreira and Dell (2000). All three experiments were spoken recall experiments. Subjects were presented stimuli that they had to recall later after a prompt was displayed. All three experiments manipulated the accessibility of the complement clause subject to test the predictions of availability-based production. The experiments used different items but the same 48 complement clause embedding verbs. I conducted three separate multilevel logit analyses on the three data sets (kindly provided by V. Ferreira). The analyses tested the effect of information density while controlling for the original design factors see Ferreira and Dell, 2000 and random effects for both subjects and items (intercepts and slopes). Information density estimates were based on verb bias information from Garnsey et al. (1997). The predicted information density effect always had the expected direction and reached significance for two of the three experiments (Experiment 1: $p < 0.03$; Experiment 2: $p < 0.003$; Experiment 4: $p = 0.21$).

Acknowledgments

I am grateful for many inspiring discussions that have influenced this paper, foremost of all with T. Wasow, R. Levy, D. Jurafsky, V. Ferreira, M. Tanenhaus, H. Clark, and A. Frank. I also wish to thank S.W. Cook, M. Gillespie, M. Tanenhaus, D. Jurafsky, Gary Dell, A. Fine, E. Hirshorn, C. Kurumada, and C. Hansen-Karr for feedback on earlier versions of this manuscript, and A. Wu for annotation work. This work was supported by RAships at the Linguistics Department, Stanford University (sponsored by T. Wasow and D. Jurafsky), a post-doctoral fellowship at the Department of Psychology, UCSD (V. Ferreira's NICHD Grant R01 HD051030), and NSF Grant BCS-0845059 to the author.

References

Adamson HD. Social and processing constraints on relative clauses. American Speech 1992;67(2)

Agresti, A. An introduction to categorical data analysis. 2. New York, NY: John Wiley and Sons Inc; 2002.

Anderson, JR. The adaptive character of thought. Hillsdale, NJ: Lawrence Erlbaum; 1990.

Arnold JE, Wasow T, Asudeh A, Alrenga P. Avoiding attachment ambiguities: The role of constituent ordering. Journal of Memory and Language 2004;55(1):55–70.

Avgustinova, T. Word order and clitics in Bulgarian. German Research Center for Artificial Intelligence; 1997.

Aylett MP, Turk A. The smooth signal redundancy hypothesis: A functional explanation for relationships between redundancy, prosodic prominence, and duration in spontaneous speech. Language and Speech 2004;47(1):31–56. [PubMed: 15298329]

Aylett MP, Turk A. Language redundancy predicts syllabic duration and the spectral characteristics of vocalic syllable nuclei. The Journal of the Acoustical Society of America 2006;119:30–48.

Baayen, RH. Analyzing linguistic data: A practical introduction to statistics using R. Cambridge, UK: Cambridge University Press; 2008.

Baayen RH, Feldman L, Schreuder R. Morphological influences on the recognition of monosyllabic monomorphemic words. Journal of Memory and Language 2006;55(2):290–313.

Babyak M. What you see may not be what you get: A brief, nontechnical introduction to overfitting in regression-type models. Psychosomatic Medicine 2004;66(3):411–421. [PubMed: 15184705]

Badecker, W.; Lewis, R. A new theory and computational model of working memory in sentence production: Agreement errors as failures of cue-based retrieval; Proceedings of the 20th annual CUNY conference on human sentence processing; 2007.

Bates D, Maechler M, Dai B. lme4: Linear mixed-effects models using s4 classes [R package version 0.999375-28]. 2008

Bates, E.; MacWhinney, B. Functionalist approaches to grammar. In: Wanner, E.; Gleitman, LR., editors. Language acquisition: The state of the art. Cambridge, UK: Cambridge University Press; 1982. p. 173-218.

Bell A, Brenier J, Gregory M, Girand C, Jurafsky D. Predictability effects on durations of content and function words in conversational English. Journal of Memory and Language 2009;60(1):92–111.

Bell A, Jurafsky D, Fosler-Lussier E, Girand C, Gregory M, Gildea D. Effects of disfluencies, predictability, and utterance position on word form variation in English conversation. Journal of the Acoustical Society of America 2003;113(2):1001–1024. [PubMed: 12597194]

Bock JK. Syntactic persistence in language production. Cognitive Psychology 1986;18:355–387.

Bock JK. An effect of the accessibility of word forms on sentence structures. Journal of Memory and Language 1987;26(2):119–137.

Bock JK, Cutting J. Regulating mental energy: Performance units in language production. Journal of Memory and Language 1992;31(1):99–127.

Bock JK, Griffin ZM. The persistence of structural priming: Transient activation or implicit learning? Journal of Experimental Psychology 2000;129(2):177–192. [PubMed: 10868333]

Bock, JK.; Levelt, W. Grammatical encoding. In: Gernsbacher, M., editor. Handbook of psycholinguistics. San Diego, CA: Academic Press; 1994. p. 945-985.

Bock JK, Warren RK. Conceptual accessibility and syntactic structure in sentence formulation. Cognition 1985;21(1):47–67. [PubMed: 4075761]

Boersma P, Hayes B. Empirical tests of the gradual learning algorithm. Linguistic Inquiry 2001;32(1): 45–86.

Bolinger, D. That's that. Vol. 155. The Hague and Paris: Mouton [Studia Memoria Nicolai van Wijk Dedicata]; 1972.

Brennan SE, Williams M. The feeling of another's knowing: Prosody and filled pauses as cues to listeners about the metacognitive states of speakers. Journal of Memory and Language 1995;34:383–393.

Breslow NE, Clayton DG. Approximate inference in generalized linear mixed models. Journal of the American Statistical Association 1993;88(421):9–24.

Bresnan J, Carletta J, Crouch R, Nissim M, Steedman M, Wasow T, et al. Paraphrase analysis for improved generation. LINK project: HRCR Edinburgh-CLSI Stanford. 2002

Bresnan, J.; Cueni, A.; Nikitina, T.; Baayen, RH. Predicting the dative alternation. In: Bouma, G.; KrSmer, I.; Zwarts, J., editors. Cognitive foundations of interpretation. Amsterdam: Koninklijke Nederlandse Akademie van Wetenschapen; 2007. p. 69-94.

Bresnan, J.; Hay, J. Gradient grammar: An effect of animacy on the syntax of give in varieties of English. Ms., Stanford University; 2006.

Brown P, Dell GS. Adapting production to comprehension: The explicit mention of instruments. Cognitive Psychology 1987;19(4):441–472.

Brown-Schmidt S, Konopka AE. Little houses and casas peque nas: Message formulation and syntactic form in unscripted speech with speakers of English and Spanish. Cognition 2008;109:274–280. [PubMed: 18842259]

Bybee J. Word frequency and context of use in the lexical diffusion of phonetically conditioned sound change. Language Variation and Change 2002;14(3):261–290.

Bybee, J.; Hopper, P. Introduction. In: Bybee, JL.; Hopper, PJ., editors. Frequency and the emergence of linguistic structure. Amsterdam: John Benjamins; 2001. p. 229-254.

Calhoun, S. Unpublished doctoral dissertation. University of Edinburgh; 2006. Information structure and the prosodic structure of English: A probabilistic relationship.

Calhoun, S.; Nissim, M.; Steedman, M.; Brenier, J. A framework for annotating information structure in discourse. Proceedings of Frontiers in corpus annotation II: Pie in the sky, ACL2005 conference workshop; Ann Arbor, MI. 2005.

Chang F, Dell GS, Bock JK. Becoming syntactic. Psychological Review 2006;113(2):234–272. [PubMed: 16637761]

Chomsky N. Three factors in language design. Linguistic Inquiry 2005;36(1):1–22.

Clark, HH. Using language. Cambridge, UK: Cambridge University Press; 1996.

Clark HH, Carlson TB. Hearers and speech acts. Language 1982;58:332–373.

Clark HH, Fox Tree JE. Using “uh” and “um” in spontaneous speech. Cognition 2002;84:73–111. [PubMed: 12062148]

Clark, HH.; Murphy, GL. Language and comprehension. Amsterdam: North-Holland; 1982. Audience design in meaning and reference; p. 287-299.

Clark HH, Wasow T. Repeating words in spontaneous speech. Cognitive Psychology 1998;37:201–242. [PubMed: 9892548]

Clayards M, Tanenhaus M, Aslin R, Jacobs R. Perception of speech reflects optimal use of probabilistic speech cues. Cognition 2008;108(3):804–809. [PubMed: 18582855]

Dell GS, Chang F, Griffin ZM. Connectionist models of language production: Lexical access and grammatical encoding. Cognitive Science: A Multidisciplinary Journal 1999;23(4):517–542.

Deshmukh N, Ganapathiraju A, Gleeson A, Hamaker J, Picone J. Resegmentation of SWITCHBOARD. Fifth international conference on spoken language processing. 1998

Dhami M, Hertwig R, Hoffrage U. The role of representative design in an ecological approach to cognition. Psychological Bulletin 2004;130:959–988. [PubMed: 15535744]

Dor D. Toward a semantic account of that-deletion in English. Linguistics 2005;43(2):345–382.

Elsness J. That or zero? A look at the choice of object clause connective in a corpus of American English. English Studies 1984;65:519–533.

Feng Z, McLerran D, Grizzle J. A comparison of statistical methods for clustered data analysis with Gaussian error. Statistics in Medicine 1996;15:1793–1806. [PubMed: 8870161]

Fenk-Oczlon G. Familiarity, information flow, and linguistic form. Frequency and the emergence of linguistic structure 2001:431–448.

Ferreira F. Choice of passive voice is affected by verb type and animacy. Journal of Memory and Language 1994;33:715.

Ferreira VS. Is it better to give than to donate? Syntactic flexibility in language production. Journal of Memory and Language 1996;35(5):724–755.

Ferreira VS. The persistence of optional complementizer mention: Why saying a “that” is not saying “that” at all. Journal of Memory and Language 2003;48:379–398.

Ferreira VS. Ambiguity, accessibility, and a division of labor for communicative success. Learning and Motivation 2008;49:209–246. [PubMed: 19710948]

Ferreira VS, Dell GS. The effect of ambiguity and lexical availability on syntactic and lexical production. Cognitive Psychology 2000;40:296–340. [PubMed: 10888342]

Ferreira VS, Firato CE. Proactive interference effects on sentence production. Psychonomic Bulletin & Review 2002;9:795–800. [PubMed: 12613685]

Ferreira, VS.; Hudson, M. Proceedings of architectures and mechanisms of language processing. Belgium: Ghent; 2005. An emotion (that) I second: Effects of formulation difficulty and ambiguity on sentence production.

Fillmore, C. The case of case. In: Bach, E.; Harms, R., editors. Universals in linguistic theory. New York, NY: Holt, Rinehart, and Winston; 1968.

Finegan, E.; Biber, D. Style and sociolinguistic variation. Cambridge, UK: Cambridge University Press; 2001. Register variation and social dialect variation: The register axiom; p. 235-267.

Fox B, Thompson SA. Relative clauses in English conversation: Relativizers, frequency and the notion of construction. Studies in Language 2007;3:293–326.

Fox Tree JE, Clark HH. Pronouncing “the” as “thee” to signal problems in speaking. Cognition 1997;62:151–167. [PubMed: 9141905]

Frank, A.; Jaeger, TF. Speaking rationally: Uniform information density as an optimal strategy for language production. The 30th annual meeting of the Cognitive Science Society (CogSci08); Washington, DC. 2008. p. 939-944.

Fries, CC. American English grammar. New York, NY: D Appleton-Century Company; 1940.

Fry, J. Ellipsis and wa-marking in Japanese conversation. Routledge; 2003.

Gahl S, Garnsey SM. Knowledge of grammar, knowledge of usage: Syntactic probabilities affect pronunciation variation. Language 2004;80(4):748–775.

Garnsey SM, Pearlmutter NJ, Meyers E, Lotocky MA. The contributions of verb bias and plausibility to the comprehension of temporarily ambiguous sentences. Journal of Memory and Language 1997;37:58–93.

Gelman, A.; Hill, J. Data analysis using regression and multilevel/hierarchical models. Cambridge, UK: Cambridge University Press; 2006.

Gelman A, Su Y-S, Yajima M, Hill J, Pittau MG, Kerman J, et al. Arm: Data analysis using regression and multilevel/ hierarchical models. R package version 1.1-5. 2008

Genzel, D.; Charniak, E. Proceedings of the association of computational linguistics. Philadelphia, PA: 2002. Entropy rate constancy in text; p. 199-206.

Genzel D, Charniak E. Variation of entropy and parse trees of sentences as a function of the sentence number. Proceedings of empirical methods in natural language processing 2003:65–72.

Givón, T. On understanding grammar. New York, NY: Academic Press; 1979.

Givón, T. Functionalism and grammar. Amsterdam: John Benjamins; 1995.

Godfrey J, Holliman E, McDaniel J. SWITCHBOARD: Telephone speech corpus for research and development. Proceedings of ICASSP-92 1992;1:517–520.

Gómez Gallo, C.; Jaeger, TF.; Smyth, R. Incremental syntactic planning across clauses. The 30th annual meeting of the Cognitive Science Society (CogSci08); Washington, DC. 2008. p. 1294-1299.

Gordon PC, Hendrik R, Johnson M. Memory interference during language processing. Journal of Experimental Psychology: Learning, Memory and Cognition 2001;27:1411–1423.

Griffin ZM. A reversed word length effect in coordinating the preparation and articulation of words in speaking. Psychonomic Bulletin & Review 2003;10(3):603–609. [PubMed: 14620353]

Gundel JK, Hedberg N, Zacharski R. Cognitive status and the form of referring expressions in discourse. Language 1993;69(2):274–307.

Hale J. A probabilistic early parser as a psycholinguistic model. Proceedings of the North American association of computational linguistics. 2001

Hale J. The information conveyed by words in sentences. Journal of Psycholinguistic Research 2003;32 (2):101–123. [PubMed: 12690827]

Harding MC, Hausman J. Using a laplace approximation to estimate the random coefficients logit model by non-linear least squares. International Economic Review 2007;48(4):1311–1328.

Harrell, FEJ. Regression modeling strategies. Oxford: Springer-Verlag; 2001.

Harrell FEJ. Design: Design package. R package version 2.1-1. 2007

Hawkins, JA. A performance theory of order and constituency. Vol. 73. Cambridge, UK: Cambridge University Press; 1994.

Hawkins JA. Why are categories adjacent? Journal of Linguistics 2001;37:1–34.

Hawkins, JA. Efficiency and complexity in grammars. Oxford: Oxford University Press; 2004.

Haywood SL, Pickering MJ, Branigan HP. Do speakers avoid ambiguities during dialogue? Psychological Science 2005;16(5):362–366. [PubMed: 15869694]

Huddleston, R.; Pullum, GK. The Cambridge grammar of English. Cambridge, UK: Cambridge University Press; 2002.

Jaeger, TF. Proceedings of the disfluency in spontaneous speech workshop. France: Aix-en-Provence; 2005. Optional that indicates production difficulty: Evidence from disfluencies; p. 103-109.

Jaeger, TF. Ph.D. thesis. Stanford University; Stanford, CA: 2006a. Redundancy and syntactic reduction in spontaneous speech.

Jaeger, TF. TGrep2 database tools: A toolkit for the construction of databases using TGrep2 outputs. 2006b. http://www.hlp.rochester.edu/

Jaeger TF. Categorical data analysis: Away from ANOVAs (transformation or not) and towards logit mixed models. Journal of Memory and Language 2008;59:434–446. [PubMed: 19884961]

Jaeger, TF. Phonological optimization and syntactic variation: The case of optional “that”. Proceedings of the 32nd meeting of BLS; Berkeley, CA. in press

Jaeger TF. submitted for publication. Corpus-based research on language production: Information density affects the syntactic reduction of subject relatives.

Jaeger, TF.; Gerassimova, VA. Proceedings of the lfg02 conference. Stanford: CSLI Publications; 2002. Bulgarian word order and the role of the direct object clitic.

Jaeger TF, Kuperman V. Standards in fitting, evaluating, and interpreting regression models. 2009 UC Davis [Presentation given at the Workshop on Ordinary and Multilevel Modeling].

Jaeger, TF.; Snider, NE. Implicit learning and syntactic persistence: Surprisal and cumulativity. Proceedings of the 29th annual Cognitive Science Society (CogSci09); Washington, DC. 2008. p. 1061-1066.

Jaeger, TF.; Wasow, T. Processing as a source of accessibility effects on variation. In: Cover, RT.; Kim, Y., editors. Proceedings of the 31st annual meeting of the Berkeley Linguistic Society; Ann Arbor, MN: Sheridan Books; 2006. p. 169-180.

Jescheniak JD, Levelt WJM. Word frequency effects in speech production: Retrieval of syntactic information and of phonological form. Journal of Experimental Psychology Learning, Memory and Cognition 1994;20:824–843.

Jurafsky D. A probabilistic model of lexical and syntactic access and disambiguation. Cognitive Science 1996;20:137–194.

Kaltenböck G. ‘…That is the question’: Complementizer omission in extraposed that-clauses. English Language and Linguistics 2006;10:371–396.

Kamide Y, Altmann G, Haywood S. The time-course of prediction in incremental sentence processing: Evidence from anticipatory eye movements. Journal of Memory and Language 2003;49(1):133–156.

Keller, F. The entropy rate principle as a predictor of processing effort: An evaluation against eye-tracking data. Proceedings of the conference on empirical methods in natural language processing; 2004. p. 317-324.

Kraljic T, Brennan SE. Prosodic disambiguation of syntactic structure: For the speaker or for the hearer? Cognitive Psychology 2005;50:194–231. [PubMed: 15680144]

Landau M. Redundancy, rationality, and the problem of duplication and overlap. Public Administration Review 1969:346–358.

Langacker, R. Foundations of cognitive grammar. Vol. II. Stanford: Stanford University Press; 1991.

Leben, W. Unpublished doctoral dissertation. Cambridge, MA: MIT; 1973. Suprasegmental phonology.

Lee H. Parallel optimization in case systems: Evidence from case ellipsis in korean. Journal of East Asian Linguistics 2006;15(1):69–96.

Levelt, WJM. Speaking: From intention to articulation. Cambridge, MA: MIT Press; 1989.

Levelt, WJM.; Maassen, B. Lexical search and order of mention in sentence production. In: Klein, W.; Levelt, WJM., editors. Crossing the boundaries in linguistics. Dordrecht, The Netherlands: D. Reidel; 1981. p. 221-252.

Levy R. Expectation-based syntactic comprehension. Cognition 2008;106(3):1126–1177. [PubMed: 17662975]

Levy, R.; Jaeger, TF. Speakers optimize information density through syntactic reduction. In: Schlökopf, B.; Platt, J.; Hoffman, T., editors. Advances in neural information processing systems (NIPS). Vol. 19. Cambridge, MA: MIT Press; 2007. p. 849-856.

Lewis RL. Interference in short term memory: The magical number two (or three) in sentence processing. Journal of Psycholinguistic Research 1996;25:93–117. [PubMed: 8789368]

Lindstrom MJ, Bates DM. Nonlinear mixed effects models for repeated measures data. Biometrics 1990;46:673–687. [PubMed: 2242409]

Lohse B, Hawkins JA, Wasow T. Domain minimization in English verb–particle constructions. Language 2004;80(2):238–261.

MacDonald MC. Probabilistic constraints and syntactic ambiguity resolution. Language and Cognitive Processes 1994;9:157–201.

Manin D. Experiments on predictability of word in context and information rate in natural language. Journal of Information Processes 2006;6(3):229–236.

Marcus MP, Santorini B, Marcinkiewicz MA, Taylor A. Treebank-3. 1999

Marr, D. Vision: A computational approach. San Francisco: Freeman and Co; 1982.

McDonald S, Shillcock R. Eye movements reveal the on-line computation of lexical probabilities during reading. Psychological Science 2003;14(6):648–652. [PubMed: 14629701]

Meteer M. colleagues. Dysfluency annotation stylebook for the Switchboard corpus. 1995 revised by Ann Taylor.

Norcliffe, E. Ph.D. thesis. Stanford University; 2009. Head marking in usage and grammar: A study of variation and change in Yucatec Maya.

Peduzzi P, Concato J, Kemper E, Holford T, Feinstein A. A simulation study of the number of events per variable in logistic regression analysis. Journal of Clinical Epidemiology 1996;49(12):1373–1379. [PubMed: 8970487]

Piantadosi, ST.; Gibson, E. Uniform information density in discourse: A cross-corpus analysis of syntactic and lexical predictability. Proceedings of 21st annual CUNY conference on sentence processing; Chapel Hill, NC. 2008.

Piantadosi, ST.; Tily, H.; Gibson, E. The communicative lexicon hypothesis. The 31st annual meeting of the Cognitive Science Society (CogSci09); 2009. p. 2582-2587.

Pickering MJ, Ferreira VS. Structural priming: A critical review. Psychological Bulletin 2008;134(3): 427. [PubMed: 18444704]

Pluymaekers M, Ernestus M, Baayen RH. Articulatory planning is continuous and sensitive to informational redundancy. Phonetica 2005;62:146–159. [PubMed: 16391500]

Pluymaekers M, Ernestusb M, Baayen RH. Lexical frequency and acoustic reduction in spoken Dutch. Journal of the Acoustical Society of America 2005;118:2561–2569. [PubMed: 16266176]

Prat-Sala M, Branigan HP. Discourse constraints on syntactic processing in language production: A cross-linguistic study in English and Spanish. Journal of Memory and Language 2000;42(2):168–182.

Qian, T. B.S. thesis. University of Rochester; 2009. Efficiency of language production in native and non-native speakers.

Qian, T.; Jaeger, TF. Constant entropy rate in Mandarin Chinese. The 31st annual meeting of the Cognitive Science Society (CogSci09); 2009. p. 851-856.

Qian T, Jaeger TF. submitted for publication. Entropy profiles in language: A cross-linguistics investigation. Entropy.

Quirk R. Relative clauses in educated spoken English. English Studies 1957:97–109.

R Development Core Team. R: A language and environment for statistical computing. Vienna; Austria: 2008.

Race, DS.; MacDonald, MC. The use of "that" in the production and comprehension of object relative clauses. Proceedings of the 26th annual meeting of the Cognitive Science Society; 2003. p. 946-951.

Resnik P. Selectional constraints: An information-theoretic model and its computational realization. Cognition 1996;61:127–159. [PubMed: 8990970]

Rohde, D. Tgrep2 manual. 2005. http://tedlab.mit.edu/~dr/Tgrep2/tgrep2.pdf

Rohdenburg, G. Clausal complementation and cognitive complexity in English. In: Neumann, FW.; Schulting, S., editors. Anglistentag, erfurt. Trier: Wissenschaftlicher Verlag; 1998. p. 101-112.

Rohdenburg, G. The role of functional constraints in the evolution of the English complementation system. University of Paderborn; 2004.

Roland D, Dick F, Elman J. Frequency of basic English grammatical structures: A corpus analysis. Journal of Memory and Language 2007;57(3):348–379. [PubMed: 19668599]

Roland D, Elman JL, Ferreira VS. Why is that? Structural prediction and ambiguity resolution in a very large corpus of English sentences. Cognition 2005:1–28.

Saffran J, Johnson E, Aslin R, Newport E. Statistical learning of tone sequences by human infants and adults. Cognition 1999;70(1):27–52. [PubMed: 10193055]

Sakamoto Y, Jones M, Love B. Putting the psychology back into psychological models: Mechanistic vs. rational approaches. Memory and Cognition 2008;36:1057–1065.

Schuchardt, H. über die Lautgesetze: Gegen die Junggrammatiker. Berlin: Robert Oppenheim; 1885. Excerpted with English translation. In T. Vennemann, & T. H Wilbur Athenaum, Frankfurt (Eds.) (1972). *Schuchardt, the Neogrammarians, and the transformational theory of phonological change*

Schwarz G. Estimating the dimension of a model. Annals of Statistics 1978;6:461–464.

Shannon C. A mathematical theory of communications. Bell Systems Technical Journal 1948;27(4):623–656.

Simon, HA. Bounded rationality. In: Eatwell, J.; Milgate, M.; Newman, P., editors. 77K New Palgrave Dictionary of Economics. Vol. 1. London: Macmillan; 1987. p. 266-268.

Smith, N.; Levy, R. Optimal processing times in reading: A formal model and empirical investigation. The 30th annual meeting of the Cognitive Science Society (CogSci08); Washington, DC. 2008.

Stallings LM, MacDonald MC, O'Seaghdha PG. Phrasal ordering constraints in sentence production: Phrase length and verb disposition in Heavy-NP Shift. Journal of Memory and Language 1998;39:392–417.

Staub A, Clifton CJ. Syntactic prediction in language comprehension: Evidence from Either.. Or. Journal of Experimental Psychology: Learning, Memory, and Cognition 2006;32(2):425–436.

Staum, LW.; Jaeger, TF. "that"-omission beyond processing: Stylistic and social effects. New York, NY: 2005. Presented at NWAV

Tagliamonte S, Smith J. No momentary fancy! The zero in English dialects. English Language and Linguistics 2005;9(2):289–309.

Tagliamonte S, Smith J, Lawrence H. No taming the vernacular! Insights from the relatives in northern Britain. Language Variation and Change 2005;17:75–112.

Temperley D. Ambiguity avoidance in English relative clauses. Language 2003;79(3):464–484.

Thompson, SA.; Mulac, A. A quantitative perspective on the grammaticization of epistemic parentheticals in English. In: Traugott, E.; Heine, B., editors. Grammaticalization. Vol. ii. Amsterdam: John Benjamins; 1991a. p. 313-339.

Thompson SA, Mulac A. The discourse conditions for the use of complementizer that in conversational English. Journal of Pragmatics 1991b;15:237–251.

Tily H, Gahl S, Arnon I, Kothari A, Snider N, Bresnan J. Pronunciation reflects syntactic probabilities: Evidence from spontaneous speech. Language and Cognition 2009;1

Tily, H.; Piantadosi, ST. Proceedings of the workshop on the production of referring expressions: Bridging the gap between computational and empirical approaches to reference. Amsterdam: The Netherlands; 2009. Refer efficiently: Use less informative expressions for more predictable meanings.

Torres Cacoullos R, Walker JA. On the persistence of grammar in discourse formulas: A variationist study of "that". Linguistics 2009;47(1):1–43.

Traugott, EC. Subjectification in grammaticalisation. In: Stein, D.; Wright, S., editors. Subjectivity and subjectivisation: Linguistic perspectives. Vol. 1. Cambridge, UK: Cambridge University Press; 1995. p. 31-54.

Trueswell JC. The role of lexical frequency in syntactic ambiguity resolution. Journal of Memory and Language 1996;35:566–585.

Trueswell JC, Tanenhaus MK, Kello C. Verb-specific constraints in sentence processing: Separating effects of lexical preference from garden-paths. Journal of Experimental Psychology Learning, Memory, and Cognition 1993;19:528–553.

van Son, RJJH.; Beinum, FJ.; Koopmans-van; Pols, LCW. Efficiency as an organizing principle of natural speech. Fifth international conference on spoken language processing; Sydney. 1998.

van Son RJJH, Pols LCW. How efficient is speech? Proceedings of the Institute of Phonetic Sciences 2003;25:171–184.

van Son RJJH, van Santen JPH. Duration and spectral balance of intervocalic consonants: A case for efficient communication. Speech Communication 2005;47(1):100–123.

Walter, MA. Unpublished doctoral dissertation. Massachusetts Institute of Technology; 2007. Repetition avoidance in human language.

Walter, MA.; Jaeger, TF. Constraints on optional that: A strong word form OCP effect. Proceedings of the main session of the 41st meeting of the Chicago linguistic society; Chicago, IL. Chicago Linguistic Society; 2008. p. 505-519.

Wasow T. Remarks on grammatical weight. Language Variation and Change 1997;9(1):81–105.

Wasow, T.; Arnold, J. Post-verbal constituent ordering in English. Rohdenburg, G.; Mondorf, B., editors. Berlin: Walter de Gruyter; 2003. p. 119-154.

Wasow, T.; Jaeger, TF.; Orr, D. Lexical variation in relativizer frequency. In: Wiese, H.; Simon, H., editors. Proceedings of the workshop on expecting the unexpected: Exceptions in grammar at the 27th annual meeting of the German Linguistic Association; Berlin and New York. Mouton de Gruyter; in press

Wells J, Christiansen M, Race D, Acheson D, MacDonald M. Experience and sentence processing: Statistical learning and relative clause comprehension. Cognitive Psychology 2009;58(2):250–271. [PubMed: 18922516]

Wheeldon L, Lahiri A. Prosodic units in speech production. Journal of Memory and Language 1997;37 (3):356–381.

Yaguchi M. The function of the non-deictic “that” in English. Journal of Pragmatics 2001;33(7):1125–1155.

Zipf GK. Relative frequency as a determinant of phonetic change. Harvard Studies in Classical Philology 1929;40:1–95.

Zipf GK. The psychobiology of language. 1935

Zipf, GK. Human behavior and the principle of least effort: An introduction to human ecology. Addison-Wesley; 1949.

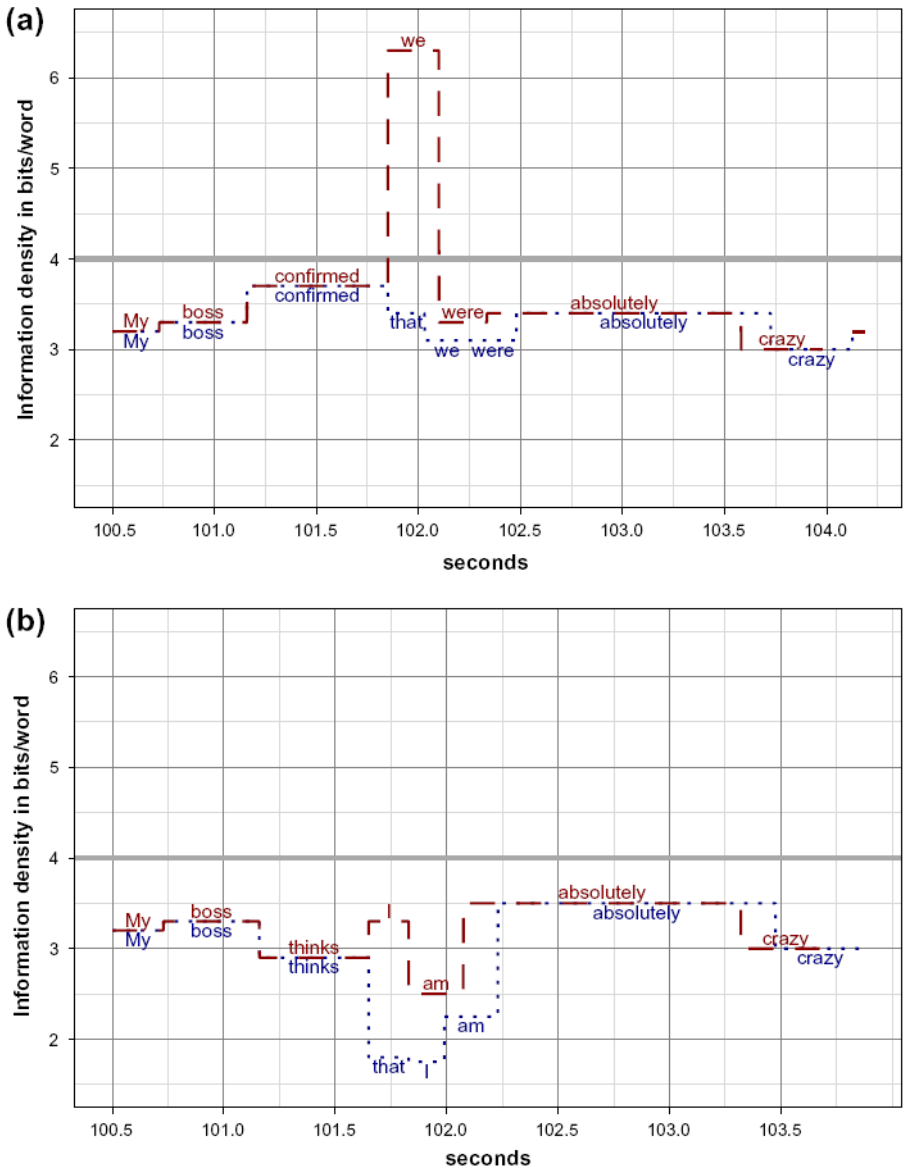

Fig. 1.

Illustration of the development of information density over time (here simply information per word) for two alternative ways to encode the same message with a complement clause. Dashed lines indicate the variants without the complementizer *that*. Dotted lines indicate the variants with the complementizer that. A purely hypothetical channel capacity is indicated by the solid horizontal line (the value of 4 bits/word is arbitrarily chosen). (a) An example with high a priori information density at the complement clause onset. (b) An example with low information density.

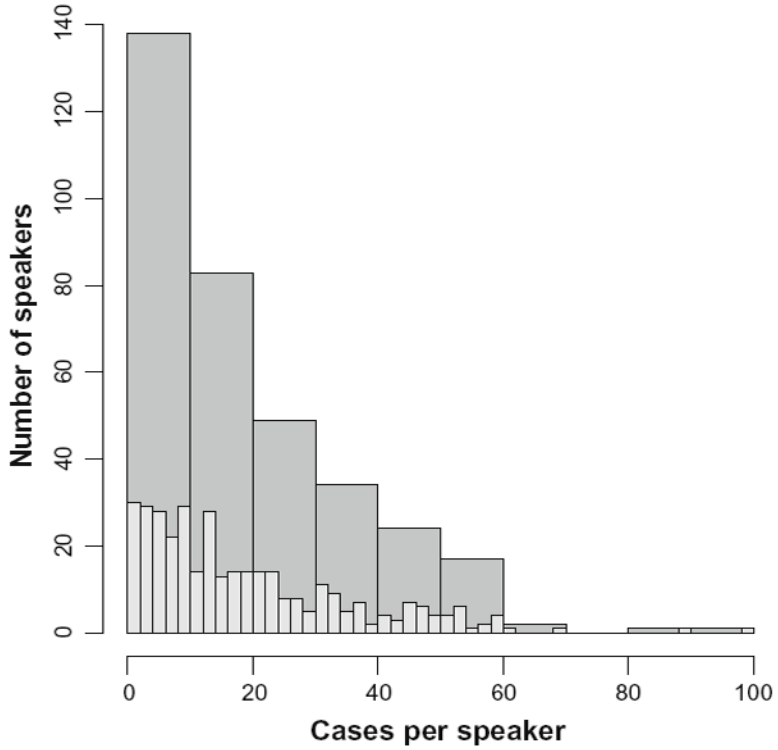

Fig. 2.
Histogram of CCs per speaker in the database. Wider bars show bins of width 10, thinner bars show bins of width 2.

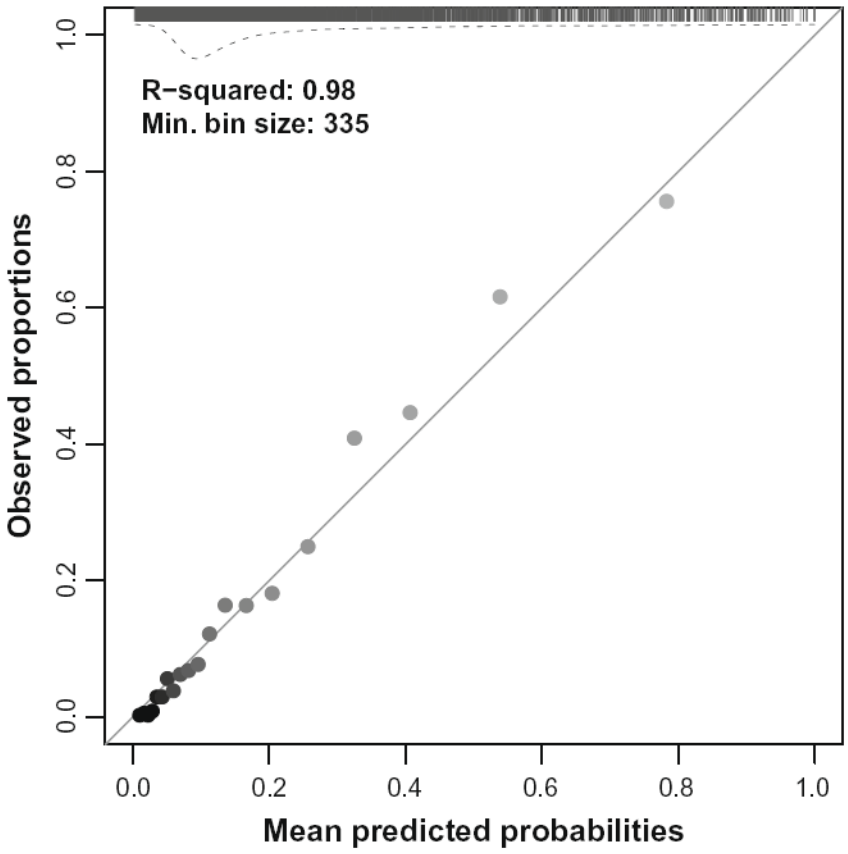

Fig. 3.

Mean predicted probabilities vs. observed proportions of *that*. The data are divided into 20 quantiles, each containing at least 335 data points. The data rug and the kernel density plot at the top of the plot visualizes the distribution of the predicted values. The R^2 of the predicted probabilities vs. observed proportions is given (this is not to be misunderstood as a measure of model quality).

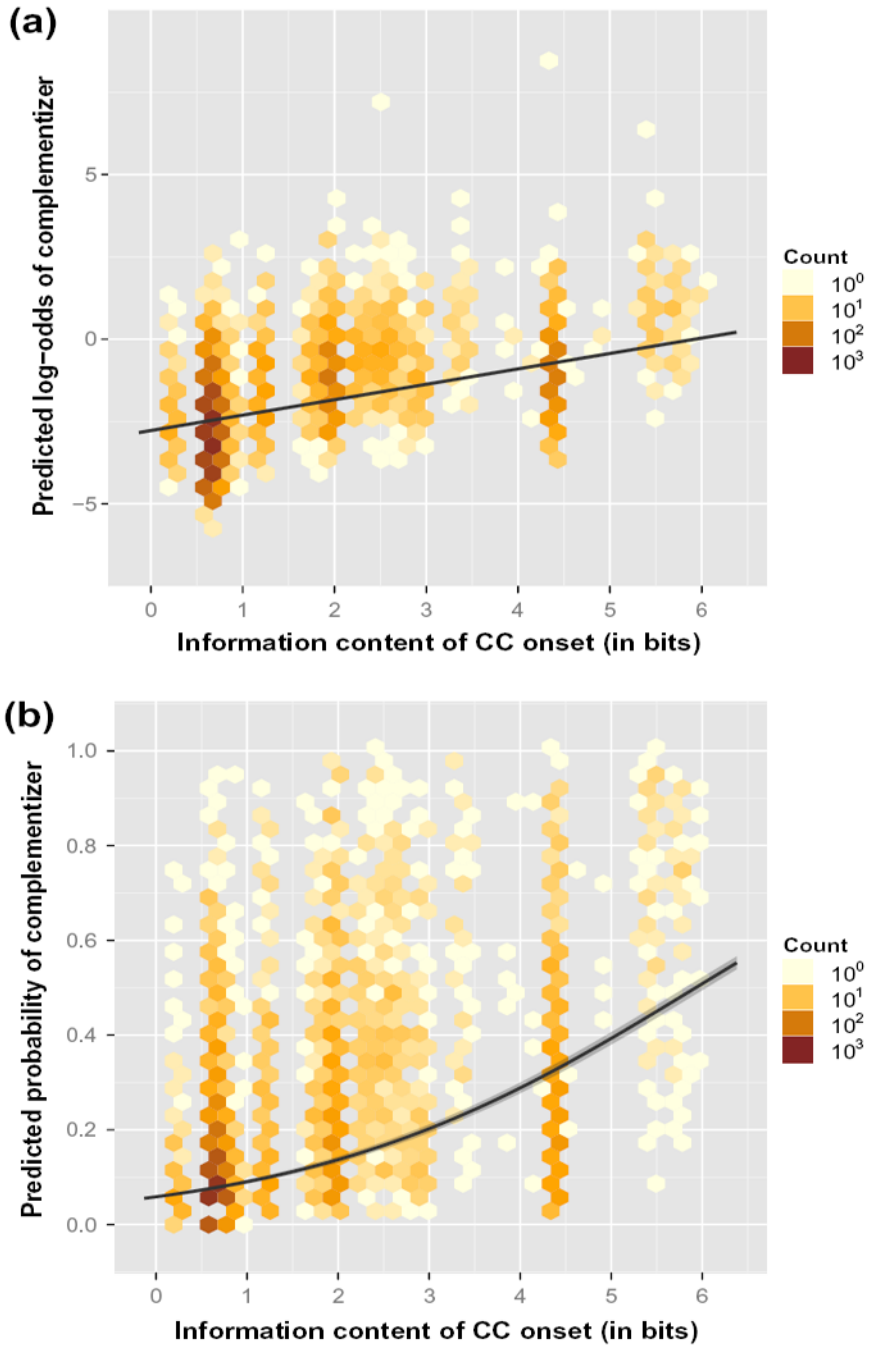

Fig. 4.

Effect of information density at the complement clause onset on *that*-mentioning along with 95% CIs (shaded area, which is hard to see because the CIs are very narrow around the predicted mean effect). (a) The effect on the log-odds of complementizer *that* (the space in which the analysis was conducted). (b) The effect transformed back into probability space. Hexagons indicate the distribution of information density against predicted log-odds (a) and probabilities (b) of *that*, considering *all* predictors in the model. Fill color indicates the number of cases in the database that fall within the hexagon.

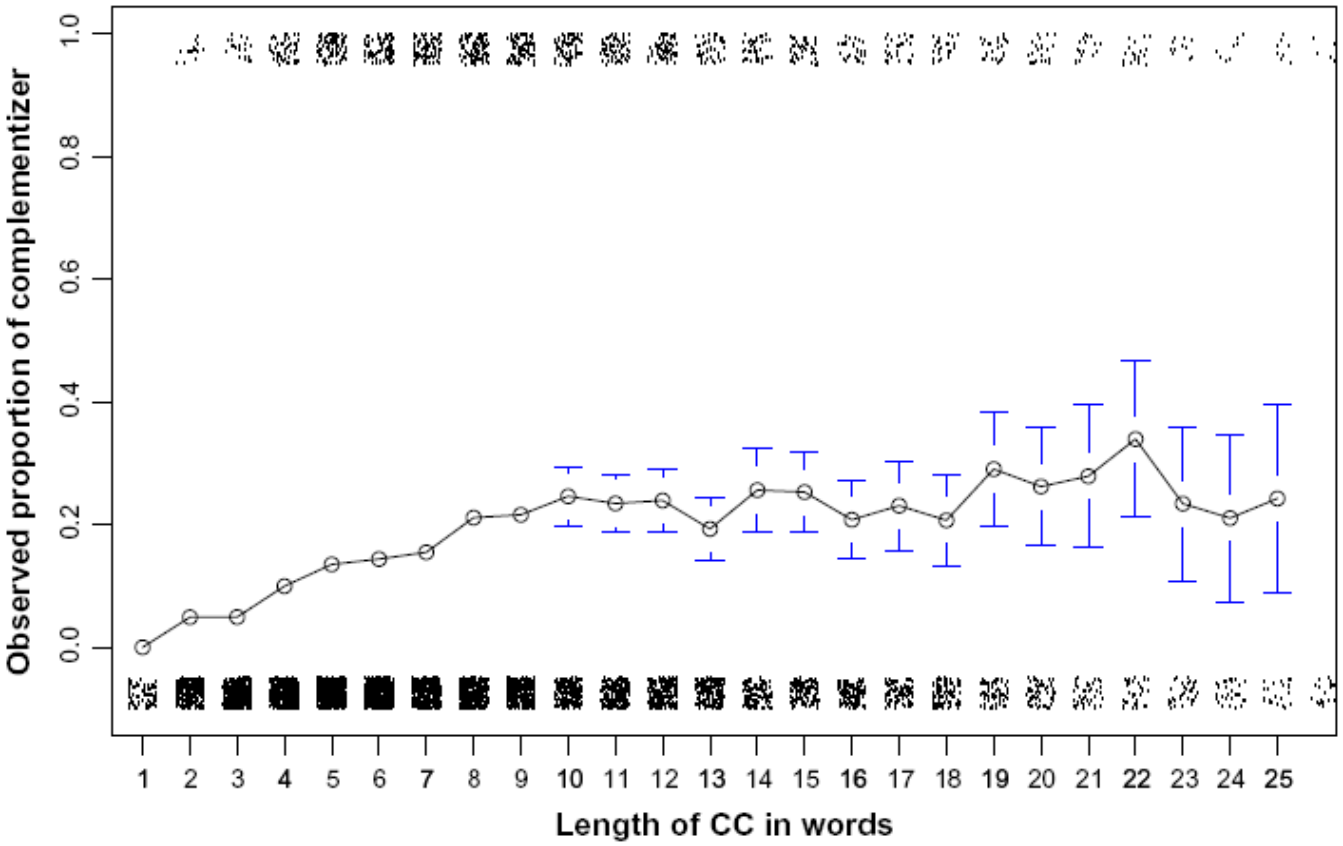

Fig. 5.
Observed proportions of *that* by CC length in words (limited to CCs up to 25 words); jittered points are bottom and top of each cell represent individual cases; error bars indicate 95% confidence intervals. Note that CCs of length 1 are observed (though infrequent, 0.01%) because speakers were interrupted or for other reasons did not complete all CCs.

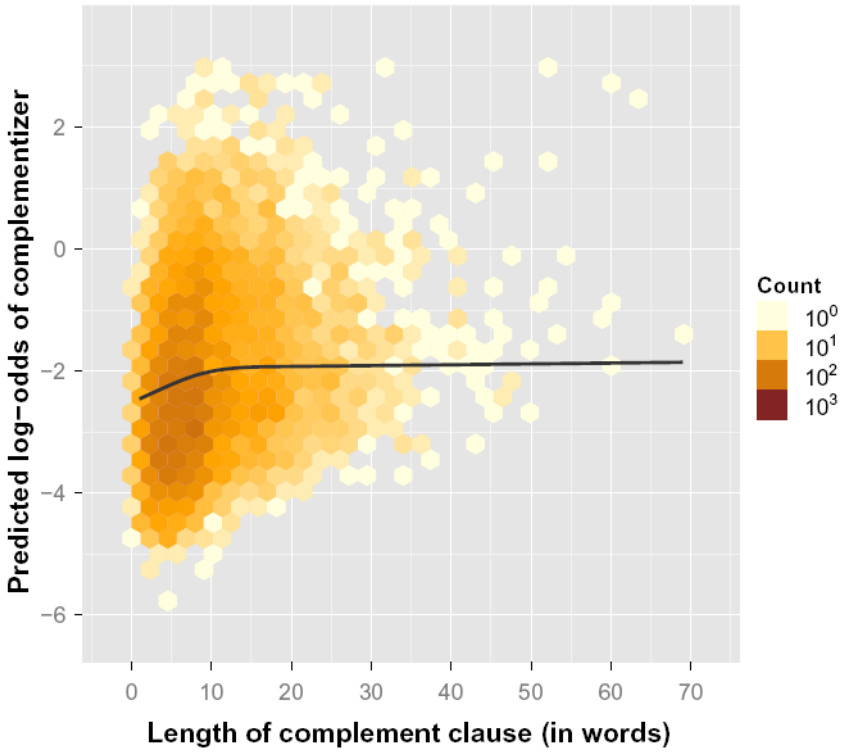

Fig. 6.

Non-linear effect of the complement clause length on log-odds of *that*-mentioning. Hexagons indicate the distribution of the predictor against predicted log-odds of *that*, considering *all* predictors in the model. Fill color indicates the number of cases in the database that fall within the hexagon.

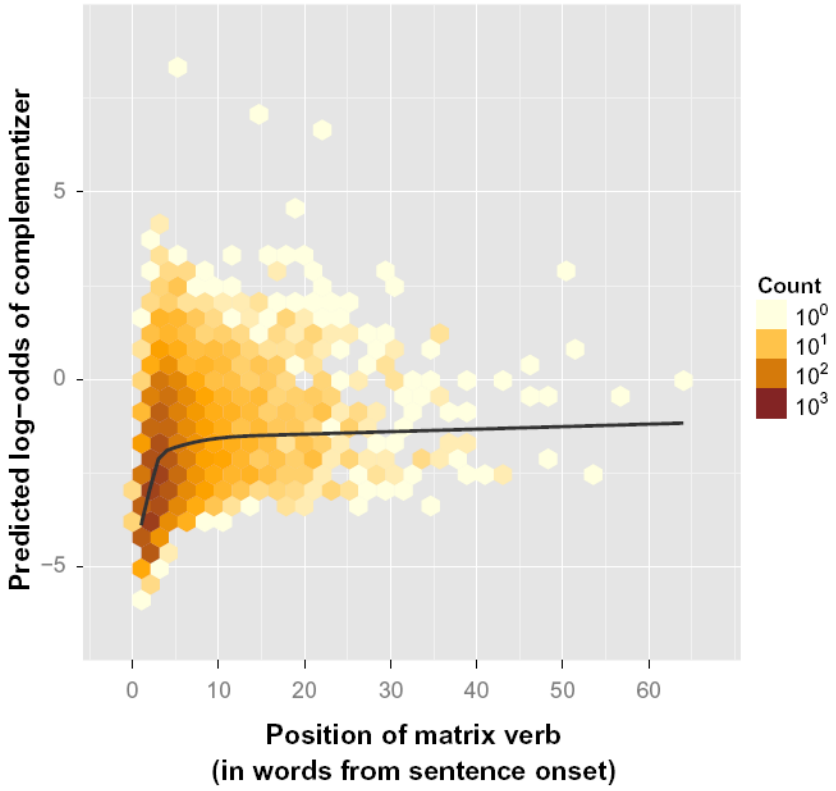

Fig. 7.

Non-linear effect of the position of the matrix verb on log-odds of *that*-mentioning. Hexagons indicate the distribution of the predictor against predicted log-odds of *that*, considering *all* predictors in the model. Fill color indicates the number of cases in the database that fall within the hexagon.

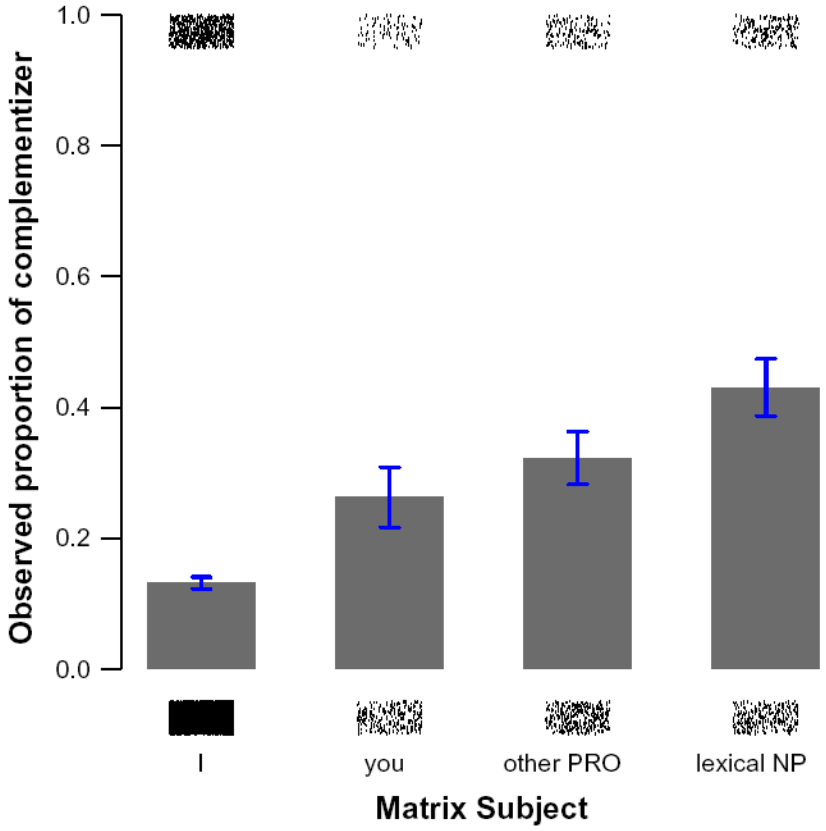

Fig. 8.
Effect of matrix subject on *that*-mentioning; jittered points are bottom and top of each cell represent individual cases; error bars indicate 95% confidence intervals.

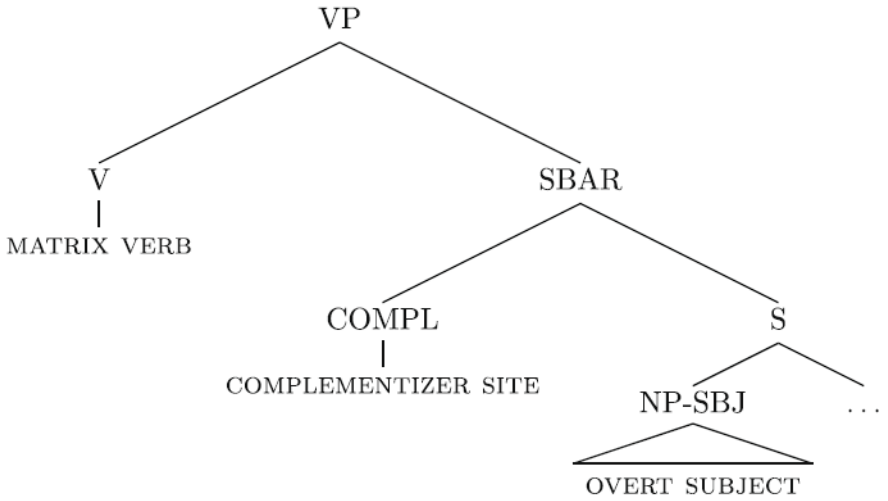

Fig. B.1.
Tree representation of TGrep2 search pattern for CCs.

Table 1

The four most frequent verbs in the database and observed proportions of *that*.

Verb lemma	Percent of database (%)	*that*-bias in database (%)
think	52	11
guess	14	1
know	8	32
say	8	27
Remaining 27 verbs	17	47

Table 2

Predictors in the analysis. The name and description of each input variable are listed. The last column describes the predictor type ('cat', categorical; 'cont', continuous) along with the number of parameters associated with it

Predictor	Description	Type (βs) (1)
INTERCEPT		
Dependency length and position of CC		
POSITION(MATRIX VERB)	CC position in the sentence	cont(3)
LENGTH(MATRIX VERB-TO-CC)	Distance of CC from matrix verb	cont(1)
LENGTH(CC ONSET)	Length of CC onset	cont(1)
LENGTH(CC REMAINDER)	Length of remainder of CC	cont(1)
Overt production difficulty at CC onset		
SPEECH RATE	Log and squared log speech rate	cont(2)
PAUSE	Pause immediately preceding CC	cat(1)
DISFLUENCY	Normalized disfluency rate at CC onset	cont(1)
Lexical retrieval at CC onset		
CC SUBJECT	Type of CC subject	cat(3)
SUBJECT IDENTITY	Matrix and CC subject are identical	cat(1)
FREQUENCY(CC SUBJECT HEAD)	Log frequency CC subject head lemma	cont(1)
WORD FORM SIMILARITY	Potential for double *that* sequence	cat(1)
Lexical retrieval before CC onset		
FREQUENCY(MATRIX VERB)	Log frequency of verb lemma	cont(1)
Ambiguity avoidance at CC onset		
AMBIGUOUS CC ONSET	CC onset ambiguous without *that*	cat(1)
Grammaticalization		
MATRIX SUBJECT	Type of matrix subject	cat(3)
Additional controls		
SYNT. PERSISTENCE	Prime (if any) w/ or w/o *that*	cat(2)
MALE SPEAKER	Speaker is male	cat(1)
Total number of control parameters in model plus intercept		25

Table 3

Result summary: coefficient estimates β, standard errors SE(β), associated Wald's z-score (= β/SE(β)) and significance level p for all predictors in the analysis

Predictor	Coef. β	SE(β)	z	p
Intercept	0.12	(0.38)	0.3	>0.7
POSITION(MATRIX VERB)	0.95	(0.14)	6.6	<0.0001
(1st restricted comp.)	−27.94	(5.33)	−5.2	<0.0001
(2nd restricted comp.)	55.43	(10.80)	−5.1	<0.0001
LENGTH(MATRIX VERB-TO-CC)	0.17	(0.065)	2.5	=0.01
LENGTH(CC ONSET)	0.18	(0.014)	12.8	<0.0001
LENGTH(CC REMAINDER)	0.03	(0.006)	4.4	<0.0001
LOG SPEECH RATE	−0.70	(0.13)	−5.5	<0.0001
SQ LOG SPEECH RATE	−0.36	(0.19)	−1.9	<0.06
PAUSE	1.11	(0.11)	10.2	<0.0001
DISFLUENCY	0.39	(0.12)	3.2	<0.002
CC SUBJECT =*it vs. I*	0.04	(0.08)	0.5	>0.6
=*other pro vs. prev. levels*	0.05	(0.03)	1.6	<0.11
=*other NP vs. prev. levels*	0.11	(0.02)	4.9	<0.0001
FREQUENCY(CC SUBJECT HEAD)	−0.02	(0.03)	−0.7	>0.5
SUBJECT IDENTITY	−0.32	(0.17)	−1.9	<0.052
WORD FORM SIMILARITY	−0.31	(0.17)	−1.8	<0.08
FREQUENCY(MATRIX VERB)	−0.23	(0.03)	−7.7	<0.0001
AMBIGUOUS CC ONSET	−0.12	(0.12)	−1.0	>0.2
MATRIX SUBJECT =*you*	0.48	(0.15)	3.1	<0.002
=*other PRO*	0.60	(0.13)	4.8	<0.0001
=*other NP*	0.85	(0.13)	6.7	<0.0001
PERSISTENCE =*no vs. prime w/o that*	0.02	(0.07)	0.3	>0.7
=*prime w/ that vs. prev. levels*	0.06	(0.04)	1.6	<0.11
MALE SPEAKER	−0.15	(0.11)	−1.3	>0.19
Information density	**0.47**	**(0.03)**	**16.9**	**<0.0001**

Table 4

Distribution of disambiguation points for potentially ambiguous CC onsets.

Disambiguation point – word:	**1**	**2**	**3**	**4**	**5–9**	**>9**
Number of instances	773	86	61	48	35	9