New ways of analyzing dialect grammars: complementizer omission in traditional British English dialects

Benedikt Szmrecsanyi (KU Leuven) and Daniela Kolbe-Hanna (University of Trier)

Abstract

The paper is concerned with complementizer retention and omission in English complement clauses (as in *I don't think that I do* versus *I don't think* ___ *I do*). We draw on a richly annotated dataset and several state-of-the-art multivariate analysis methods (mixed-effects logistic regression analysis, random forest modeling, and conditional inference trees) to explore the determinants of this variation. We show that modern approaches to analyzing grammatical variation which cross sub-disciplinary boundaries (variationist sociolinguistics, dialectology, and research on knowledge/processing/cognition) can yield new insights into well-known grammatical variation phenomena.

Acknowledgments

We are grateful for extremely helpful feedback by two reviewers and the editors. Our thanks also go to Benedikt Heller for help with formatting the manuscript. The first-named author acknowledges an Odysseus grant by the Research Foundation Flanders (FWO, grant no. G.0C59.13N). The usual disclaimers apply

1. Introduction

In this paper, which is a refined and extended version of the analysis offered in Kolbe-Hanna and Szmrecsanyi (in press), we explore grammatical variation in British English dialects, based on the *Freiburg Corpus of English Dialects* (FRED), a naturalistic corpus sampling dialect speech all over Great Britain. More specifically, we investigate complementizer retention and omission in contexts where speakers have the choice between retaining the explicit complementizer *that*, as in (1a), or omitting it, as in (1b); both examples are produced by the same speaker, in the same interview. The latter variant is also referred to in the literature as the use of the zero variant, or as complementizer deletion.

(1) a. <u IntRR> Do you remember any accidents on the farm?

<u ICS_JE> No, I don't *think*$_{\text{matrix verb}}$ [*that* I do]$_{\text{complement clause}}$.
(FRED CON_007)[1]
complementizer retention

b. <u IntRR> Something different. Do you remember any special diseases or

[1] All linguistic examples in this paper are drawn from the *Freiburg Corpus of English Dialects* (FRED) (see Section 3) and are referenced by FRED text identifiers.

disasters?

<u ICS_JE> No, I don't *think*$_{\text{matrix verb}}$ [___ I do] $_{\text{complement clause}}$.
(FRED CON_007)
complementizer omission

Our study uses a variationist research design in the spirit of e.g. Tagliamonte and Smith (2005), restricting attention to the frequent matrix verbs *think, say,* and *know*. We first identified all variable contexts in the data, as is customary in variationist linguistics. Next, because many factors are known to influence complementizer *that* retention/omission, we annotated all occurrences of complementizer *that* retention or omission for a large number of language-internal predictors (e.g. embedded clause length, properties of the matrix verb, and so on) as well as for a limited number of language-external predictors (e.g. geography – in which county in Great Britain was the interview recorded?) (see Section 3 for details). Subsequently, we marshalled three modern statistical analysis techniques – mixed-effects regression modeling, random forest modeling, and conditional inference trees – to demonstrate how new ways of analyzing syntactic variation can shed fresh light on a *per se* well-researched alternation in the grammar of English. In summary, our study is a corpus-based exercise in variationist (socio)linguistics because it investigates linguistic choices as a function of language-internal and language-external constraints; it is concerned with knowledge, processing, and cognition thanks to the inclusion of factors such as the horror-aequi principle ("avoid identical structures in adjacency") or persistence (priming), which is about the tendency of language users to repeat previously heard or used linguistic material; and it has a dialectological twist, because it also considers the effect that geography has on linguistic choices.

This contribution is structured as follows. In Section 2, we review the literature on complementizer retention/omission in dialects of English and beyond. Section 3 describes the data source we use and the methods on which we rely. In Section 4, we present the empirical analyses. Section 5 offers a summary and some concluding remarks.

2. On complementizer retention/omission in English

The variation between retention and omission of the *that*-complementizer has been the subject of an abundance of previous research in various linguistic sub-disciplines, e.g. cognitive linguistics, sociolinguistics and psycholinguistics. Empirical research has predominantly dealt with post-predicate embedded clauses (e.g*., I know (that) she did it*), in which the omission of *that* is the more frequent option (McDavid 1964; Biber et al. 1999:681–682). A notable exception is Kaltenboeck's (2006a; 2006b) work on extraposed *that* clauses, e.g., *It is obvious (that) she did it* (2006a: 371, his example). The aim of recent work (Tagliamonte & Smith 2005; Jaeger 2006; Kolbe 2008; Torres Cacoullos & Walker 2009) has been to use multivariate analyses to study the significance and strength of the suggested determinants of this variation in naturalistic speech. To our knowledge, the only previous studies drawing on British English dialect data are Tagliamonte and Smith (2005) and Kolbe (2008). In the remainder of this

section, we will give an overview of research aimed at identifying possible determinants of the variation between retention and omission of the complementizer. The description here will remain brief, as we will provide more details of the individual determinants relevant to our analyses in the following sections.

Only a few studies of the variation in *that*-complementizer choice have included language-external factors such as social groups and region (Finegan & Biber 2001; Staum 2005; Kearns 2007; Kolbe 2008:90–129). As regards register and style, research drawing on corpora of spoken and written English has demonstrated that the omission of *that* is pervasive in conversation and informal language and more frequent in more subjective contexts (McDavid 1964; Finegan & Biber 1995; 2001; Biber et al. 1999:12,680; Storms 1966:262-265).

Most of the language-internal factors identified as determinants of the variation between retention and omission of *that* revolve around the processing load caused by the structure of matrix and embedded clause. According to McDavid (1964) and Hawkins (Hawkins 2004:154), in sentences such as *His mother knew (that) the boy liked homemade* bread (McDavid 1964:108, her example) the subject of the embedded clauses could also function as object of the main clause verb when the complementizer is omitted (*His mother knew the boy*). Thus the retention of the complementizer *that* would help to avoid structural ambiguity between nominal and clausal complements of transitive verbs by identifying the embedded clause. A number of psycholinguistic experiments and studies, however, have shown that this is not the case (Trueswell, Tanenhaus & Kello 1993; Garnsey et al. 1997; Ferreira & Dell 2000; Roland, Elman & Ferreira 2006). The retention of *that* does not necessarily avoid ambiguity because the linguistic context usually provides enough cues as to whether a verb is followed by a direct object or a *that* clause (Roland, Elman & Ferreira 2006).

The retention of *that* has also been claimed to prevent structural ambiguity when lexical material intervenes between the matrix verb and the subject of the embedded clause (as in *I think, however, that this is unnecessary*) by identifying the beginning of the embedded clause (Bolinger 1972:38; Elsness 1984:524). Bolinger (1972) also argues that, due to the derivation of the complementizer from the demonstrative pronoun, the retention of *that* achieves stronger anaphoric reference to the matrix clause which is needed especially when the complement clause is not adjacent. Hence, *that* is employed when there is greater semantic distance; an observation parallel to the importance of the cognitive strength of reference detected by Elsness (1984:529-531). *That* is omitted when less cognitive effort is needed to connect the complement clause to previous material, but retained when its reference is clear. The hypothesis that a greater semantic distance leads to the retention of *that* is also supported by Kaltenboeck (2006b) and Yaguchi (2001). Bolinger (1972) and Yaguchi (2001) perceive more objectivity when *that* is retained (Bolinger 1972) and more emotionality when *that* is omitted (Yaguchi 2001:1132–1140). However, experiments conducted by Kinsey, Jaeger and Wasow (2007) have shown that the variation between the retention and the omission of *that* does not imply a meaning distinction between less or more emotional utterances or a differentiation based on conceptual distance.

The cognitive complexity of an utterance has proved to be a strong influence on the retention of *that*. According to Rohdenburg's "complexity principle" (1996; 1999), more explicit grammatical options are typically preferred in cognitively more complex environments. *That* is often retained to reduce parsing load by explicitly marking the following clause as complement (Hawkins 2001; Hawkins 2003; Jaeger 2006), whereas in cognitively less complex structures explicit marking is not necessary. Therefore, the omission of *that* is most frequent in "brief and uncomplicated" clauses (Quirk et al. 1985:1050). Specific properties that influence and/or diagnose cognitive complexity in the retention/omission of *that* are represented by the predictors considered in the present study. These include the choice of matrix verb (Storms 1966:262–265; Biber et al. 1999:681–682; Dor 2005; Jaeger 2006:78–89; Torres Cacoullos & Walker 2009:20–23), linguistic material intervening between matrix and embedded clause (Bolinger 1972:38; Rohdenburg 1996:160; Hawkins 2003:178–179), morphosyntactic properties of the matrix clause (Rohdenburg 1996), whether the matrix clause is *I think* (Thompson & Mulac 1991), subject and length of the embedded clause (Rohdenburg 1999), disfluency markers (Jaeger 2006), and whether the subjects in matrix and embedded clause are coreferential (Elsness 1984:526; Finegan & Biber 2001:262). Yet another cognitive process that influences the alternation between the omission and retention of *that* is the persistence of grammatical structures in speech production: speakers tend to re-use syntactic patterns that they have heard or produced before (Ferreira 2003). These properties will be discussed in more detail in section 3.3.

In sum, a speaker's choice between retaining or omitting the *that* complementizer has been shown to depend on a multitude of factors, and our analysis in section 4 below aims at taking account of them.

3. Data and methods

3.1. Data source

This study taps into *the Freiburg Corpus of English Dialects* (henceforth: FRED) (see Hernández 2006; Szmrecsanyi & Hernández 2007 for the publicly available sampler version). The corpus spans approximately 2.5 million words of running text, which translates into about 300 hours of recorded speech. The texts are mainly transcribed so-called "oral history" interviews, most of which were recorded between 1970 and 1990; in most cases, a fieldworker interviewed an informant about life, work, etc. in former days. The informants sampled in the corpus are typically elderly people with a working-class background – so-called NORMS (non-mobile old rural males) (see Chambers & Trudgill 1998:29). The interviews were conducted in 163 different locations (that is, villages and towns) in 43 different pre-1974 counties in 9 major dialect areas: the Southeast of England, the Southwest of England, the English Midlands, the North of England, Wales, the Scottish Highlands, the Scottish Lowlands, the Isle of Man, and the Hebrides.

3.2. Variable context and variable extraction

We use the variationist method (in the spirit of Labov 1982; Weiner & Labov 1983). Thus the dependent variable in our study is the binary choice between retention versus omission of the complementizer *that*. The study we report on here restricts attention to the textually frequent complement-taking predicates *think, say,* and *know* – the most common matrix verbs of embedded *that* clauses (Biber et al. 1999:668) – to obtain a large sample of finite complement clauses in an economical fashion. These verbs occur vastly more often than not without complementizer, but there still is variation to be explained. Drawing on a Perl script identifying these verbs in FRED, the beginning and the end of each embedded complement clause following these verbs was handcoded by the two authors. Tests for inter-coder reliability in samples of clauses showed that the two coders agreed in 83% of all cases.[2]

3.3. Variable extraction and annotation

A second Perl script subsequently extracted the complement clauses, along with relevant meta-data. This yielded a dataset with N = 5,296 observations ($N_{omission}$ = 4,820; $N_{retention}$ = 476). In what follows, we discuss the language-external and language-internal predictors (by and large the usual suspects, according to the literature) for which the dataset was annotated. The discussion here is limited to predictors that are significant in regression analysis (see Section 4.1.).[3]

3.3.1. Language-external predictors

The role of sociolinguistic predictors in complementizer retention/omission has been discussed in several studies. Finegan and Biber (2001:261–263) observe no difference in the percentage of the omission of *that* between social groups. Kearns (2007) and Staum (2005) notice regional variation in Australian and American English, but ascribe this variation to the influence of other factors. Kolbe (2008), however, detects significant dialectal differences in the choice between the retention and the omission of *that:* First of all, the dataset (NITCS, FRED) exhibits significantly more retention of *that* in the Scottish and Welsh data, and significantly less

[2] Differences between the assessments of clause length typically result from different perceptions on whether a chunk was still embedded in the previous clause or not.

[3] Factors that were annotated but failed to obtain significance in the regression analysis include MATRIX_NEGATION (whether or not the matrix verb is negated), ADV_BEGINNING (whether an adverbial *(in, because, cause, if, since, when, after, before, during*) occurs at the beginning of the embedded clause), INTERV_MATERIAL_MACL_EMBCL (the number of orthographic words between the matrix verb and the start of the embedded clause), SAME_VERB_IN_EMBD_CL (whether or not the exact same verb as the matrix verb occurs in the embedded clause), BETA_PERS_DISTANCE (the textual distance in words to the last occurrence of any *that* token, including e.g. demonstrative *that*), and TTRPASSAGE (the type-token ratio in a context of -50/+50 words around the matrix verb slot).

in the English North and Midlands. Based on the FRED metadata, we thus annotated for the variables TEXT, COUNTY, and SPEAKER.[4]

- SPEAKER renders the speaker's ID (as defined in the FRED manual). This information we use in regression analysis to approximate idiolectal differences via a random effect.
- TEXT indicates the corpus file (e.g. SFK_018 or IOM_002) where the token occurred. The predictor models non-repeatable properties of the interview situation via a random effect.

COUNTY specifies the county in which the speaker lived at the time of recording.

3.3.2. Language-internal predictors I: features of the matrix clause

- VERB LEMMA: The variable verb lemma specifies which matrix verb is used (*think, say,* or *know)*. While all three verbs are textually highly frequent and often function as matrix verbs of *that* clauses, *think* and *say* are by far the most frequent matrix verbs of *that* clauses in British English, and compared to *know* they strongly favor the omission of *that* (Biber et al. 1999:668, 681). Previous research (Torres Cacoullos & Walker 2009; Jaeger 2006) has shown that the choice of matrix verb strongly affects the likelihood of the retention of *that.* The more often a matrix verb is followed by an embedded *that* clause (with omitted or retained complementizer), the more predictable it becomes for language users that this particular verb will be followed by this particular type of clause. Thanks to this predictability, the embedded clause is in less need to be marked by explicit *that* as being embedded. Therefore, it is omitted more often after the most frequent matrix verbs of *that* clauses, i.e. after *think, know,* and *say* (McDavid 1964; Bolinger 1972:20–23; Elsness 1984:523; Thompson & Mulac 1991; Finegan & Biber 2001:263; Biber et al. 1999:681; Jaeger 2006:78–89). In Jaeger (2006:51–95), predictability turns out to be the most important predictor of the omission of *that*..
- VERBMORPH and MATRIX_AUXILIARY: Further features of the verb phrase in the matrix clause that affect a speaker's choice to omit or retain *that* are the morphology of the verb, and whether the verb phrase contains an auxiliary. The less complex the verb phrase is, the less likely *that* is to occur (Torres Cacoullos & Walker 2009:24–27; Biber et al. 1999:681–682; Thompson & Mulac 1991:246). Therefore, VERBMORPH codes whether the matrix verb occurs in its base form (2a), or whether the morphology is more complex: third person singular present tense (2b), past (tense or participle) (2c) , or *-ing* (2d).

(2) a. They got to **know** that I worked at Staverford's (FRED LEI_002)

[4] Information on speakers' age is unavailable for 1,124 cases in the database, or more than 20 percent. Since the inclusion of age would thus result in substantial data loss, we did not include speakers' age in the analysis. We also did not include information on speaker's sex, since the sample in FRED is skewed: of the 5,296 utterances, only a quarter (1,389 instances) is produced by female speakers. While this decision simplifies model building, it partly undermines our aim to study the influence of language-external factors, in omitting factors that play an important role in traditional variationist studies.

b. So they **says** __ they'd got no driver there (FRED NTT_001)

c. So they **thought** ___ it was better to get them from Holland (FRED NTT_001)

- d. I was **saying** ___ we used to push the lifeboat out (FRED YKS_008)

MATRIX_AUXILIARY determines whether the matrix verb is preceded by a modal auxiliary (e.g. *should, could, would, will, 'll, shall, must, can* + negated forms, as in (3)) or not.

(3) you **couldn't** say ___ we had any social life (FRED NTT_007)

- MATRIX_SUBJECT_TYPE: When the matrix clause subject is *I* (4a) or *you* (4b), the omission of *that* is said to become more likely (Thompson & Mulac 1991:242–243). The appearance of the matrix subject is determined by the predictor MATRIX_SUBJECT_TYPE, which distinguishes between *I, you,it* (as in (4a-c)) versus any other subject (as in (4d)).

(4) a. **I** knew ___ it was pretty hard as, as a child (FRED NTT_001)

b. **you** said ___ they lived in the pottery (FRED SAL_008)

c. **it** meant to say that you'd got to go down the garden (FRED SAL_023)

d. The **policeman** said ___ there was a poacher going up the other side (FRED SAL_020)

- I_THINK: This is a binary predictor ('yes' vs. 'no'). If the matrix clause is *I think* (as in (5)), retention of *that* is highly unlikely.

(5) but **I think** ___ it is one of the originals (FRED SAL_002)

I think as a matrix clause not only comprises all features that favor the omission of *that* (subject *I*, simple verb morphology, highly frequent verb lemma which very often controls *that* clauses), it also functions as a comment clause or epistemic parenthetical (see e.g. Thompson & Mulac 1991; Tagliamonte & Smith 2005). According to Tagliamonte and Smith (2005), the complementizer *that* is unlikely to be retained after *I think* once speakers recognize *I think* as an epistemic marker rather than as a matrix clause *I think*. Thompson and Mulac (1991) argue that the omission of *that* in sentences such as (5) yield the epistemic phrase *I think* which does not function as matrix clause (in contrast to *I think that*). In prosodic analyses of British English, Dehé and Wichmann (2010), however, have shown that sentence-initial *I think* often does function as matrix clause, not only as epistemic marker or phrase. Their analysis also reveals that the omission of the complementizer is not preliminary to epistemic function – *I think that* is a regular variant of the epistemic marker as well as of the matrix clause *I think*. For the purpose of our analysis, we consider sentence-initial *I think* as a potential matrix clause, but do account for the fact that retention of *that* after this matrix clause is highly unlikely by way of this variable.

3.3.2. Language-internal predictors II: features of the embedded complement clause

- EMB_CL_LENGTH specifies the number of words in the embedded clause (excluding the complementizer when present), since it is known that speakers use the explicit complementizer more often when the complement clause is comparatively long (Jaeger 2006:79–87; Rohdenburg 1996:164). For example, the embedded clause in (6) has a length of 8 orthographically transcribed words.

  (6) I think ___ **the coal was wheeled through the old kiln** (FRED SAL_002)

  Note that in regression analysis, the predictor was modeled logarithmically.

- ADV_AFTEREND indicates whether an adverbial (introduced by *in, because, cause, if, since, when, after, before, during*) occurs after the end of the embedded clause, as in (7).

- (7) I thought ___ he had died in the house where I lived, **in Fonthill Road** (FRED LND_004)COMPLEMENT_SUBJECT codes whether the first element in the embedded clause is a pronoun (as in (8)), in which case previous research has diagnosed a stronger tendency to omit *that* (Rohdenburg 1996:162; Torres Cacoullos & Walker 2009:24, 28).[5]

  (8) I think ___ **they** used to make butter in the afternoons (FRED CON_011)

- HORROR_AEQUI determines whether the embedded clause after the complementizer (explicit or zero) starts in *that*, as in (9).

  (9) I think ___ **that**'s away from the farming side of things. (FRED CON_011)

  There is evidence that speakers avoid to use identical forms (*I think that that's away* …) in adjacency (Kolbe 2008:217–218, 222–224).

3.3.3. Language-internal predictors III: features across clauses

- EHMS_ETC_NARROW: According to the literature, production difficulties increase the likelihood of *that* retention (Jaeger 2006:91–92). In terms of cognitive complexity this relates to the speaker's need to make the relationship between two clauses more explicit when complexity has already led to production difficulties. We thus count the number of speech perturbations in the immediate context from three words before the matrix verb to the end of the embedded clause. For example, in (10) we find one such speech perturbation.

  (10) **Ehr**, I think ___ he worked for a local cheese factory previously (FRED SOM_029)

  In regression analysis, this predictor was modeled logarithmically.

[5] When the subjects of matrix and embedded are coreferential, processing load is reduced, which leads to more omission of *that* (Elsness 1984:525–526; Biber et al. 1999:681). Ferreira and Dell (2000) ascribe this to the fact that pronominal referents are already available to the speaker. In previous multivariate research, however, this factor did not turn out to be significant, so we do not include it in our analyses.

- ALPHA_PERSISTENCE_50: As speakers tend to reuse grammatical patterns that they have used previously (Gries 2005; Szmrecsanyi 2005; Szmrecsanyi 2006), we take into account whether an explicit *that* complementizer occurs up to 50 words before the complement clause under analysis. This is the case in example (11), for instance, where the second complement clause (*say that they put so many eggs in*) is preceded by a complement clause with an explicit *that* complementizer (*say that there was certain people*).

  (11) Y' know *they used to say **that** there was certain people who used to put chunks of beef in*; and *different ones used to say **that** they put so many eggs in*, and all that kind of thing for finings. (FRED SOM_024)

4. Analysis

In this section, we present a series of three multivariate analyses of the dataset. In Section 4.1., we discuss a binary logistic regression model with mixed effects, a technique that is a modern reincarnation of classical fixed-effects logistic regression analysis, which has been customary – in the form of the Variable Rules (Varbrul) program (Cedergren & Sankoff 1974) – in variationist linguistics since the 1970s. In section 4.2., we present a random forest analysis, which is a fairly new, non-parametric multivariate analysis technique that can be utilized to gauge the relative importance of predictors. In section 4.3., we draw on a conditional inference tree to highlight how predictors interact to fuel variation in complementizer *that* retention or omission.

4.1. Mixed-effects regression modeling

Logistic regression modeling (Pampel 2000), which has been the workhorse statistical analysis technique in variation studies for decades, is the closest a linguist working with observational data can come to conducting a controlled experiment: The procedure models the combined contribution of all predictors considered, and systematically tests the effect of each predictor while holding the other predictors in the model constant. Mixed-effects logistic regression modeling (Pinheiro & Bates 2000)[6] is a recent improvement over more traditional regression modeling techniques, which is becoming increasingly popular in variation studies (see, for example, Gries & Hilpert 2010; Wolk et al. 2013; Hinrichs, Szmrecsanyi & Bohmann 2015). Mixed-effects models incorporate not only "classical" fixed effects but also so-called random effects, such as idiolectal differences, or differential propensities of different verb lemmas to retain or omit the complementizer. Consider for exemplification idiolectal differences: the dataset we analyze here samples speech from 331 speakers. Modeling idiolectal differences as a fixed effect is, for one thing, challenging technically because a fixed-effect predictor with 331 levels is unwieldy. On the other hand, these 331 speakers do (of course!) not

[6] We utilized the implementation of generalized linear mixed effects models in the lme4 library (R package version 0.999999-2) in R (R Development Core Team 2013).

exhaust the population of dialect speakers in Great Britain, which is why a different sample would in all likelihood sample different speakers – in other words, we are dealing here with a generalizability issue as idiolectal differences (in our case, the variable SPEAKER) are not necessarily repeatable and therefore, in statistical parlance, "random". For this reason, our analysis will model idiolectal differences as a random effect, rather than ignoring it and losing statistical power (see Tagliamonte & Baayen 2012:142–146 for an extended discussion of this issue). We would like to emphasize that the label 'random' or 'non-repeatable' does not necessarily imply that an effect is un-important – quite on the contrary. Random effects simply offer new ways to analyze constraints on variation.

4.1.1. Model design

The dependent (response) variable is the choice between the zero complementizer and its explicit form; the predicted odds are for retention of *that*. As FRED consists of spoken data only and "[i]n conversation, the omission of *that* is the norm, while the retention of *that* is exceptional" (Biber et al. 1999:680), the more frequent value of the complementizer is its omission, which occurs in 91% of all embedded clauses in our dataset. We included a number of fixed effects and four independent variables, or factors, as random effects: VERB LEMMA, SPEAKER, TEXT and COUNTY. Again, these factors are non-repeatable effects, as a second study relying on randomly chosen verbs, speakers, texts and counties would result in a different sample. VERB LEMMA can be seen as a classical by-item effect, whereas SPEAKER is the classical by-subject effect. TEXT and COUNTY are directly connected to SPEAKER, because they represent a particular interview with a speaker who lived in a specific county at the time. The random effects in our model are all intercept adjustments. As for the fixed effects, our model is a simple main effects model, as we do not have a theoretically justified reason to expect interaction effects, which is why we did not go fishing for them.[7]

4.1.2. Model fitting

We observed the customary steps (Baayen 2008; Crawley 2005) to obtain a minimal adequate regression model. We began by fitting the maximal model including all available predictors (including those listed in footnote 3). The model was simplified by removing predictors and interaction terms lacking significant explanatory power. We started the pruning process with the least significant effect, moving on to more significant ones. Explanatory power of main-effect categorical predictors with more than two levels was assessed via likelihood-ratio tests. The pruning process resulted in the deletion of the predictors detailed in footnote 3. Finally, we tested the justification of including the random effects (intercept adjustments) in the model by means of likelihood ratio tests. To guard against overfitting, the model was bootstrapped using the test discussed in Baayen (2008:283; sampling with replacement, 10 runs,

[7] We add that the conditional inference tree discussed in Section 4.3. may be utilized to generate testable hypotheses about interaction effects.

the confidence intervals did not include zero); also, the issue of excluding outliers was investigated and consequently rejected.

4.1.3. Model evaluation

The quality of the resulting, minimal adequate model is good. As for predictive accuracy, the model correctly predicts 92.4% of all outcomes, which is a modest but significant ($p = 0.01$) increase over the baseline percentage at 91.0%, which represents the percentage of zero-complementizers in the database. Although the model is thus validated, its predictive bonus is not exactly breathtaking. This may well be caused by the predominance of the zero variant, since a baseline percentage of already 91% is difficult to increase. In any event, the index of concordance *C*, which measures how well the model discriminates between retention and omission of *that*, is 0.89 (values > 0.8 are customarily interpreted as indicating a good fit). Multicollinearity is not a problem, as the model's condition number ($\kappa = 12.5$) is below the customary threshold of 15.

4.1.4. Fixed effects

Table 1 lists the fixed effects that turn out to be significant predictors of the choice between zero and explicit complementizers. The strength of each predictor is indicated by the value in the columns "Coefficient". These are the estimated coefficients of the respective predictors, to be added or subtracted from the intercept. Negative numbers indicate a negative (disfavoring) influence of this predictor on the use of explicit *that*; positive numbers indicate an increase in the likelihood of the retention of *that* if the respective predictor level applies. A larger value represents a stronger effect. The figures in the column "Odds Ratio" render the same influence in a different, more interpretable form. An odds ratio of 1 would mean that the odds for the use of *that* are 1:1 (i.e. no effect). If *that* is less likely to be used with a certain predictor, the odds ratio for this predictor ranges between 0 and 1 (e.g. 2:10 = 0.2), while odds ratios larger than 1 (e.g. 10:2 = 5) indicate that the complementizer is more likely to be retained. The significance codes refer not to the strength of a predictor but its statistical reliability in the model. For categorical variables (e.g. VERBMORPH or HORROR_AEQUI), the model relies on treatment coding, such that we consider a particular category the constant or default value, to which the other values are compared. The default category in VERBMORPH, for example, is "base" – the verb used in its base form (e.g. *say*). Let us now go through the three categories of language-internal predictors discussed in Section 3.3. to see how individual predictors in each category favor or disfavor retention of *that*.

Table 1. Fixed effects in the minimal adequate logistic regression model. Predicted odds are for the retention of *that*. Significance levels: . marginally significant ($p < 0.1$), * significant ($p < 0.05$), **very significant ($p < 0.01$), *** highly significant ($p < 0.001$)

	coefficient	odds ratio	
(Intercept)	-3.37	0.03	***
a. Language-internal predictors I: features of the matrix clause			
VERBMORPH (default: base form)			
VERBMORPH: 3sg	-0.73	0.48	.
VERBMORPH: past	-0.35	0.71	*
VERBMORPH: *ing*	1.07	2.92	***
MATRIX_AUXILIARY: yes	0.72	2.05	***
MATRIX_SUBJECT_TYPE (default: other)			
MATRIX_SUBJECT_TYPE: *I*	-1.48	0.23	***
MATRIX_SUBJECT_TYPE: *it*	1.42	4.13	**
MATRIX_SUBJECT_TYPE: *you*	-0.99	0.37	**
I_THINK: yes	-0.70	0.49	***
b. Language-internal predictors II: features of the embedded complement clause			
LOG(EMB_CL_LENGTH)	1.08	2.95	***
ADV_AFTEREND: yes	0.42	1.52	*
COMPLEMENT_SUBJECT: pronoun	-0.53	0.59	***
HORROR_AEQUI: yes	-1.04	0.35	**
c. Language-internal predictors III: features across clauses			
LOG(EHMS_ETC_NARROW+1)	0.79	2.21	**
ALPHA_PERSISTENCE_50: yes	1.07	2.92	***

Language-internal predictors I: features of the matrix clause (Table 1a). As for VERBMORPH, the use of *that* becomes less likely when the matrix verb is in third person singular present tense or in past tense. An *–ing* participle, however, increases the likelihood of omission. A matrix auxiliary (MATRIX_AUXILIARY) significantly favors *that* retention, and the form of the matrix subject (MATRIX_SUBJECT_TYPE) also has a significant effect: when the matrix subject is *I* or *you*, the retention of *that* is less likely than with *it* or the default category of any other subject. When *it* is the matrix subject, speakers use *that* more readily. Finally, speakers disfavor *that* retention after the sequence *I think* (I_THINK). As discussed in section 3.3.2 above, *I think* in sentence-initial position can function as matrix clause or comment clause/ epistemic marker; both functions allow the retention

of *that*.

Language-internal predictors II: features of the embedded complement clause (Table 1b). As expected, longer embedded clauses favor *that* retention; for every one-unit increase in the log of EMB_CL_LENGTH (which measures the number of orthographically transcribed words), the odds for *that* retention increase by a factor of 2.95. The presence of an adverbial at the end of the embedded clause (ADV_AFTEREND) likewise increases the odds for *that* retention. Conversely, a pronoun as the subject of the complement clause (COMPLEMENT_SUBJECT) disfavors *that* retention, and so does HORROR_AEQUI: when the embedded clause starts in *that,* speakers avoid the use of the explicit complementizer.

Language-internal predictors III: features across clauses (Table 1c). As expected, for every one-unit increase in the log of EHMS_ETC_NARROW (which measures the number of speech perturbations in the immediate context), the odds for *that* retention increase. Also as expected, recent usage of a *that* complementizer increases the odds for *that* retention by a factor of 2.92.

4.1.5. Random effects

The four random effects included in the model are summarized in Table 2. Column 'N' indicates how many levels are distinguished for each effect, and Column 'Variance' indicates the relative importance of the random effects. Thus we observe that the random effect SPEAKER is the most important random effect, while VERB LEMMA is – interestingly – the least important one.

Table 2. Random effects in the minimal adequate logistic regression model

	N (categories)	**variance**
VERB LEMMA	3	0.09
SPEAKER	331	0.54
TEXT	310	0.17
COUNTY	38	0.33

Let us have a closer look at the random effects in the model. As for VERB LEMMA – a by-item effect that captures differential propensities of particular verbs to retain the complementizer – *think* disfavors *that* retention most strongly of the three matrix verbs: of 3,394 instances of *think*, only 152, or 4.5%, control an explicit *that* clause, so the adjustment of the intercept is negative (-0.27) for *think*, while the intercept adjustment for *know* is 0.37 and the adjustment for *say* is 0.15. This is another way of saying that the matrix verb *know* is per se most likely to retain the complementizer.

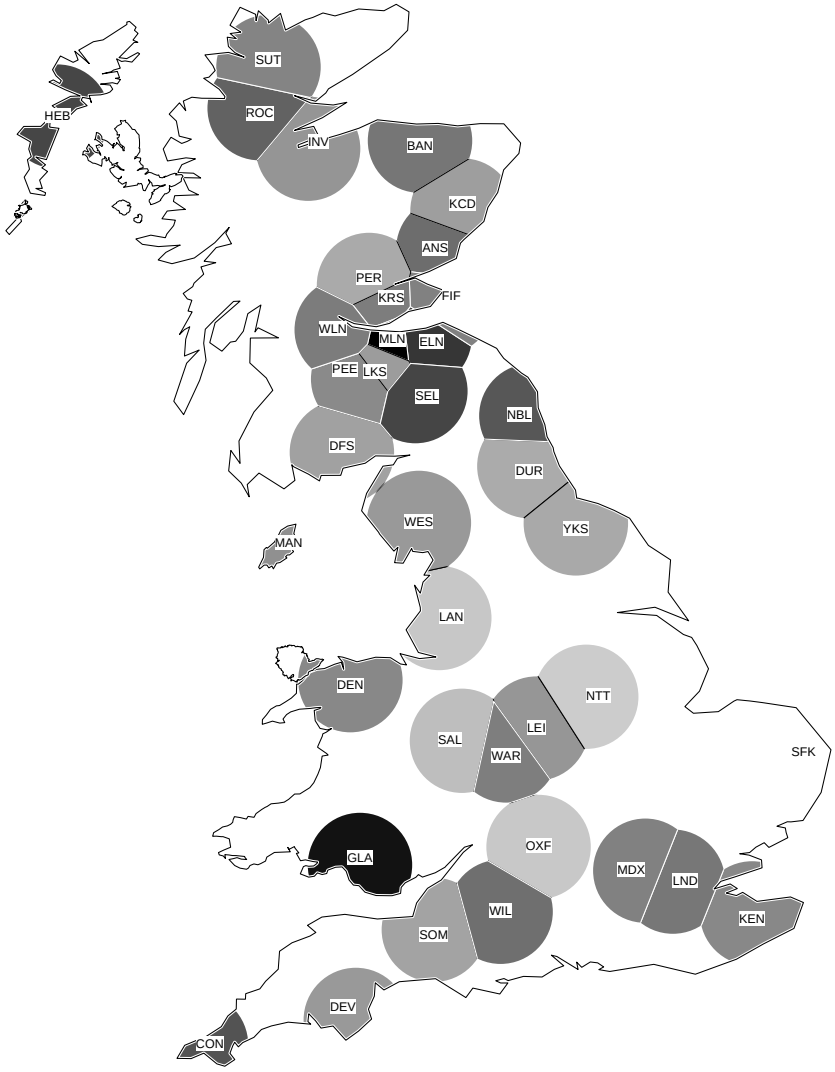
SPEAKER and corpus TEXT are two so-called by-subject effects that seek to take care of idiolectal variation and by-interview variation, respectively. It is not feasible to report all intercept adjustment for each of the 300+ speakers and texts. In the extremes, FRED speaker MlnJH disfavors explicit complementation most (intercept adjustment: -0.87; *that* retained in 4.3%), while speaker SRLM_HM is most favoring (1.33; 48% *that* retained). Similarly, FRED interview MLN_007 (in which speaker MlnJH is being interviewed) is least hospitable towards explicit *that* (-0.28; 4.3% retention), and interview KEN_010 is most hospitable (0.45, 24% retention).

The dialectological meat of the current analysis is in the random effect COUNTY. The geographical distribution of complementizer choice according to county is illustrated in Map 1. This dialectometric map (Szmrecsanyi 2011; Szmrecsanyi 2013) translates our analysis into the sort of geolinguistic mapping customary in dialectology and dialectometry by projecting intercept adjustments to geography. It shades counties in Great Britain proportionally to the intercept adjustment that they receive in the regression model. In other words, Map 1 highlights how hospitable dialects are towards *that* retention under multivariate control. Darker shades indicate more hospitality towards explicit that (i.e. positive intercept adjustments), lighter shades indicate less hospitality towards explicit *that* (i.e. negative intercept adjustments). Thus, we see a stronger preference to omit *that* in Central England (including Lancashire), and a preference for retention of *that* in Southern Wales, in Edinburgh and to its South (East Lothian, Midlothian and Selkirkshire), and on the Outer Hebrides, which corresponds to the findings in Kolbe (Kolbe 2008:120–121). In general, we note that in Southern Great Britain *that* is less likely to be retained; notable exceptions are Southern Wales and Cornwall.

A tentative explanation of this geographical distribution would observe, first, that the areas where complementizer *that* is retained – parts of Scotland (including the Hebrides), Wales, and parts of the Southwest of England – are those where English has had a history as a Second Language to be acquired during adulthood and has had to compete, to varying degrees, with Celtic L1 languages. By contrast, the places where *that* tends to be omitted, according to the regression analysis, form an area where English has been long in use as an uncontested mother tongue. Note now that we know from SLA research that learners of English have a preference for retaining complementizer *that*:

> learners adopt a more conservative strategy with regard to complementizer omission, such that they only drop the complementizer under 'safe' circumstances, that is, in contexts that do not entail high processing cost and/or with verbs that are particularly highly associated with zero-*that*. (Wulff, Lester & Martinez-Garcia 2014:291)

Hence, we would like to suggest that Map 1 can be seen as a dialectological reflex of past Second Language Acquisition processes, and reflects to some degree the history and type of the dialects that are spoken in the areas investigated here.



*Map 1.*Projection of COUNTY intercept adjustments to geography. Darker shades indicate positive intercept adjustments (i.e. more hospitality towards explicit *that*), lighter shades indicate negative intercept adjustments (i.e. less hospitality towards explicit that).

4.1.6. Regression model: interim summary

Most of the language-internal predictors discussed here behave as they should, given

the literature. Observe, along these lines, that many of these predictors are concerned, in one way or another, with cognitive complexity (in the spirit of Rohdenburg 1996; 1999a). Thus we saw that the complementizer is more likely to be omitted in less complex environments, in which it is easier for the listener to infer that material following the matrix clause will be an embedded *that* clause.[8] Such cognitively less complex environments are typically the more frequent patterns, which makes the syntactic structure of the utterance more predictable (Roland, Elman & Ferreira 2006). That, for instance, *you think* will be followed by an embedded *that* clause is predictable for listeners because this is the most frequent case, so that the relationship between the clauses need not be explicated by the retention of *that* (in exemplar theoretic terms, one could say that frequent non-complex contexts may be stored as predictable chunks, which makes explicit marking unnecessary). By contrast, cognitively more complex contexts increase the need to mark the embeddedness of the following clause explicitly by retaining *that*. In our study this proves to be the case especially if the matrix subject is *it,* or when the embedded clause is long. That said, one predictor increasing the probability of explicit *that* is not related to cognitive complexity, namely the persistence of *that*. Speakers are more likely to use explicit *that* if they did so the last time they had a choice and if that choice occurred within 50 words of the present slot. Although this predictor is not related to the cognitive complexity of the current locus of variation, it is linked to a speaker's processing load in general, since *that* is repeated simply because it prevails in the speaker's working memory (see, for example, Gries 2005; Szmrecsanyi 2006).

Some of our results do come as a surprise, however. It was not to be expected that the occurrence of an adverbial at the end of the embedded clause would turn out to be a significant predictor, whereas an adverbial between matrix and embedded clause did not.

Methodologically, the discussion of the random effects has shown that these are not necessarily linguistically uninteresting "nuisance factors" (think of by-subject effects such as SPEAKER and TEXT), but may very well reveal more crucial insights about the observable variation, as we saw in the discussion of the dialectologically highly relevant random effect COUNTY.

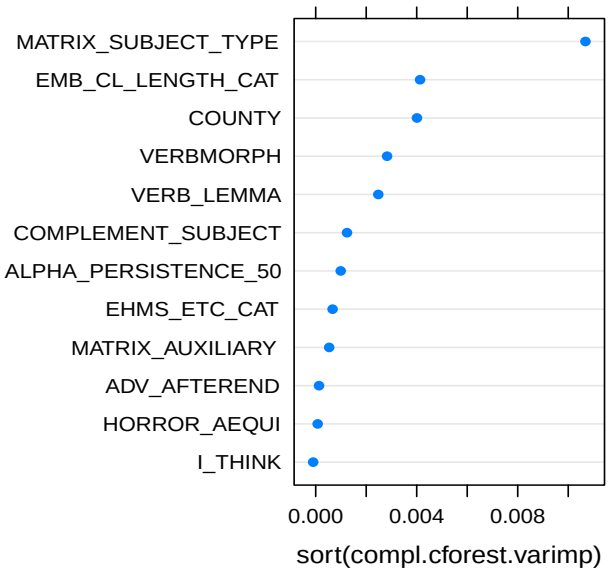
4.2. Random forest modeling

The regression analysis of the fixed effects in the previous section was instructive in terms of the *effect directions* and *effect sizes* that predictors have – so, for example, we have seen that recent usage of a *that* complementizer increases the odds for *that* retention by a factor of 2.92. By contrast, in this section we are concerned with *the variable importance* of factors for predicting complementizer *that* retention or omission. Note here that effect size and variable importance are different things. Both smoking and being hit by lightning adversely

[8] If sentence-intial *I think* functions as epistemic marker, the following clause is a main clause. According to Thompson & Mulac (1991 a, b), comment clause function derives from a reanalysis caused by the omission of *that* which is due to the high frequency of *I think*. Thus, the shift in function originates in precisely the high frequency and predictability we have outlined here.

affect life expectancy, but the effect size of lightning is much larger than the effect size of smoking. Yet on the whole, because lightning strikes are very rare, in a comprehensive model that predicts life expectancy in a large population, smoking will presumably be vastly more important explanatorily than lightning.

To gauge the variable importance of the language-internal predictors we have considered so far, we turn to random forest analysis (Breiman et al. 1984; Breiman 2001), which is a cutting-edge analysis technique that is becoming increasingly popular in linguistics (Shih 2011; Tagliamonte & Baayen 2012; de Swart, Eckhoff & Thomason 2012; Schneider 2014).[9] We skip a discussion of the technicalities and refer the reader to the very accessible introduction in Tagliamonte and Baayen (2012) instead. What is important here is that random forest modeling seeks to predict, much as logistic regression analysis does, a linguistic outcome – e.g. retention or omission of *that* – given a set of predictor variables. However, unlike regression analysis, random forest modeling is a non-parametric analysis technique (which is another way of saying that the technique makes fewer assumption about the data than e.g. regression analysis). The technique can be applied to so-called "small *n* large *p*" datasets (that is, datasets with comparatively few observations but much annotation); multicollinearity is never a problem; and random forest modeling is immune to overfitting and small cell counts (Tagliamonte & Baayen 2012:161–163). What is especially appealing from the point of view of the present study is that the procedure can also be used to rank predictors according to their overall explanatory importance.



[9] We utilize the cforest() function of the random forest implementation available in the party package (R package version 1.0-8) in R (R Development Core Team 2013). To rank predictors, we use the conditional variable importance measure implemented in the party package's varimp() function (Strobl et al. 2008) with the parameter 'conditional' set to 'FALSE'.

Figure 1. Variable importance of variables predicting complementizer retention and omission according to random forest analysis.

A variable importance ranking of the predictors in our dataset, based on the random forest, is presented in Figure 1. Note that for the sake of keeping the analysis computationally feasible – random forest modeling is computationally very intensive – we restricted attention to those language-internal predictors that proved significant in the foregoing regression analysis (Section 4.1).[10] Note further that we categorized the variable EMB_CL_LENGTH into three groups (short, medium, long), and the variable EHMS_ETC_NARROW into two groups (speech perturbations present versus absent).

What we can see from Figure 1, then, is that the subject type of the matrix clause is, by a wide margin, the overall most important predictor. Recall from Section 4.1. that *I* and *you* as matrix subjects disfavor *that* retention, and as it turns out this (dis)preference is overall most crucial for predicting *that* retention in our dataset. After some distance, length of the embedded clause (thanks to the complexity principle, longer clauses favor *that* retention) and COUNTY are tied for second place. The strong showing of COUNTY – which indicates where in Great Britain the interview was recorded – underlines the important role that dialectologically/geolinguistically conditioned variation plays in our dataset. Verb morphology is the third-ranked predictor. The fourth-ranked predictor is VERB LEMMA.

4.3. Conditional inference trees

Let us now investigate how the language-internal predictors reviewed in the previous sections interact to create linguistic outcomes. To this purpose, we consider in this section a single conditional inference tree (rather than an entire forest of such trees, as in Section 4.2). Bernaisch, Gries and Mukherjee (2014:14) succinctly sum up the essence of conditional inference trees as follows:

> Conditional inference trees are a recursive partitioning approach towards classification and regression that attempt to classify / compute predicted outcomes / values on the basis of multiple binary splits of the data. Less technically, a data set is recursively inspected to determine according to which (categorical or numeric) independent variable the data should be split up into two groups to classify / predict best the known outcomes of the dependent variable […] This process of splitting the data up is repeated until no further split that would still sufficiently increase the predictive accuracy can be made, and the final result is a flowchart-like decision tree.

[10] Due to limitations of computing resources, the factors SPEAKER and TEXT – which we modeled as random effects in regression analysis – could not be included in the analysis.

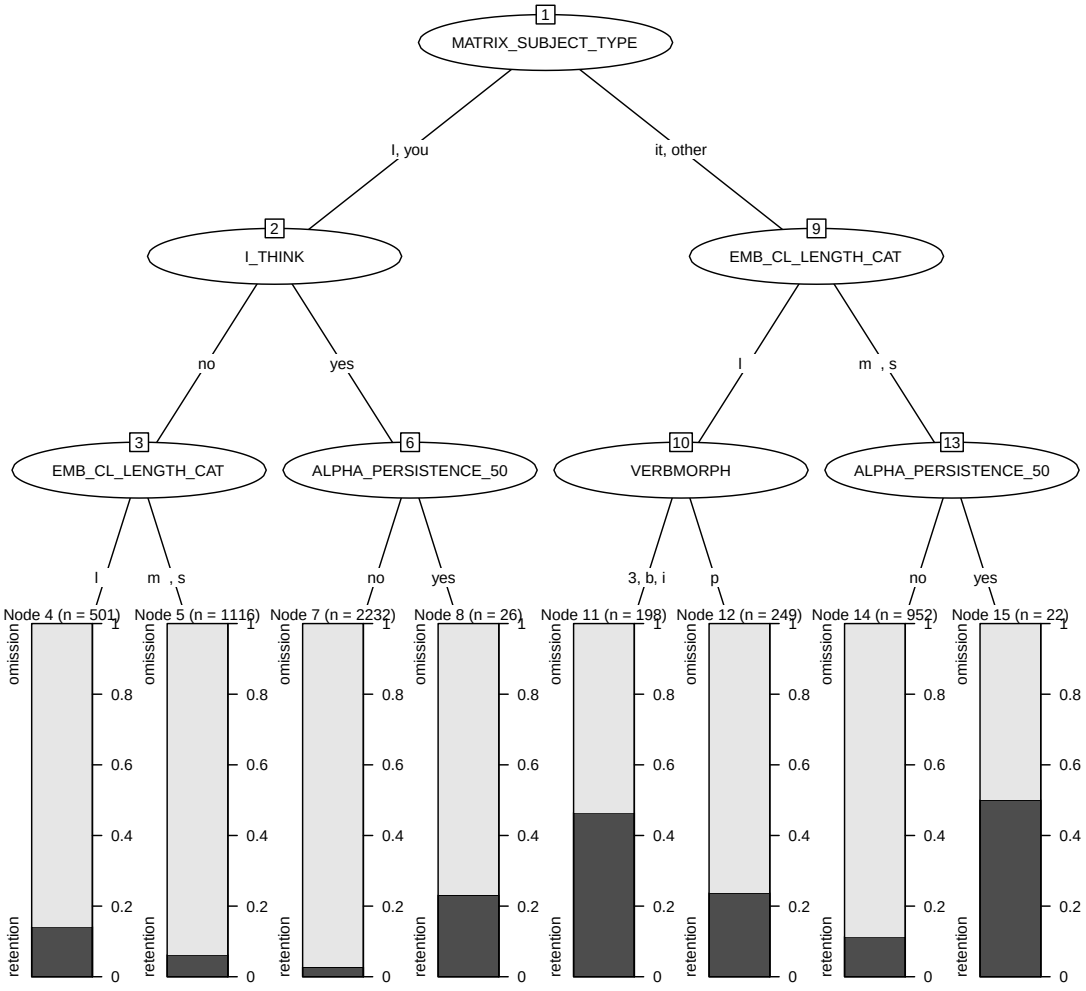


Figure 2. Conditional inference tree. Abbreviations: EMB_CL_LENGTH – 'l' long, 'm' medium, 's' short; VERBMORPH – '3' 3rd person singular, 'b' base, 'I' ing, 'p' past.

A conditional inference tree[11], based on the exactly same model formula that was fed into the random forest analysis in the previous section, is shown in Figure 2. To interpret the tree, we start at the top node (node 1) and move down in the tree. Terminal nodes (bars) indicate the relative frequency of *that* retention/omission, given the configuration of predictors at higher nodes; for example, node 4 (the leftmost bar in the tree) shows that there are 501 cases in the

[11] We used the ctree() function in the package party (R package version 1.0-8) in R (R Development Core Team 2013). To keep the tree structure manageable, we used the parameter maxdepth=3.

dataset where

- the matrix subject type is *I* or *you*, and
- the matrix clause is not *I think*, and
- the embedded clause is long.

Under such circumstances, *that* retention occurs in approximately 15% of all cases.

In all – and in accordance with the random forest analysis – the conditional inference tree suggests that MATRIX_SUBJECT_TYPE is the overall most crucial predictor. If the subject is *I* or *you*, the outcome depends on whether the matrix clause is actually *I think*; if it is not, the share of *that* retention is lower when the length of the embedded clause is short or medium than when it is long; and so on. The big picture that emerges is that complementizer retention is most frequent (terminal node 15; about 50% retention) when the following three predictors interact in the following way:

- MATRIX_SUBJECT_TYPE = '*it*' or 'other'
- EMB_CL_LENGTH_CAT = 'medium' or 'short'
- ALPHA_PERSISTENCE_50 = 'yes' (that is, when an overt complementizer has been used recently).

By contrast, *that* retention is least likely (terminal node 7; less than 5% retention) when

- MATRIX_SUBJECT_TYPE = '*I*' or '*you*'
- I_THINK = 'yes' (see sections 3.3.2 and 4.1.4 above and Dehé and Wichmann (2010) for a more detailed discussion of this sequence)
- ALPHA_PERSISTENCE_50 = 'no' (that is, when an overt complementizer has not been used recently).

4.4. Random forest modeling and conditional inference trees: interim summary

This is a good opportunity to recapitulate what the conditional inference tree analysis (Section 4.3.) and the random forest analysis (Section 4.2.) have taught us that we did not already know from the regression analysis (Section 4.1.). The random forest has ranked the predictors under consideration in this study according to their overall importance (not according to effect size, which is what the coefficients in regression models measure). The conditional inference tree analysis has indicated in which corners of the data particular linguistic outcomes are particularly frequent. Both random forest analysis and conditional inference tree analysis are rather exploratory techniques that look at data from slightly different angles – and as we all know, changing one's perspective never hurts.

5. Summary and concluding remarks

In this paper, we have explored how a number of language-internal and language-external predictors constrain variation between complementizer *that* retention and omission in a corpus sampling dialect speakers all over Great Britain. Adopting the variationist method, we created a richly annotated dataset and marshaled three multivariate analysis techniques to investigate variation in complementizer *that* retention versus omission. We drew on mixed-effects logistic regression analysis (Section 4.1) to gauge, for one thing, the effect direction and effect size of a number of fixed-effect constraints. It turned out that most of these predictors have the effects reported in the literature, and that the complexity principle (Rohdenburg 1996; 1999) motivates many of them – for example, the fact that longer embedded clauses favor explicit *that*. We also saw a number of frequency effects (e.g. the frequent phrase *I think* triggering complementizer omission), and we would like to encourage future students of this variation to investigate exemplar-theoretic explanations for these effects.

In the dialectology department, we projected by-county intercept adjustments to geography and saw that as a rule of thumb, *that* is more likely to be retained in areas where English has been in contact with other languages and where it has had a history of Second Language Acquisition. We argued that this may be due to the fact that learners demonstrably (Wulff, Lester & Martinez-Garcia 2014) have a preference for retaining complementizers.

In Section 4.2., we grew a random forest to assess the variable importance of the predictors that turned out as significant in the foregoing regression analysis. The three factors that are overall most crucial for predicting complementizer retention/omission are (1) the type of the subject in the matrix clause, (2) the length of the embedded clause, and (3) the county where the interview was recorded. Finally, in Section 4.3., we inspected a conditional inference tree and learned that according to this particular statistical technology, particular pairings of contexts relating to the subject in the matrix clause, the length of the embedded clause, and persistence (among other predictors) tend to yield characteristically low or high rates of complementizer retention.

In conclusion, it would seem that our analysis of grammatical variation in English (dialects) is innovative in two ways. On the one hand, thanks to the dataset we use and the predictors we investigate, the study is situated at the intersection of variationist (socio)linguistics, dialectology, and research on the nature of linguistic knowledge, processing, and cognition. On the other hand, on more methodical grounds we utilize three new analysis techniques (mixed-effect logistic regression analysis, random forest modeling, and conditional inference trees) to investigate our dataset from different angles. We believe that new ways of analyzing grammatical variation like ours, which crosses sub-disciplinary boundaries using new analysis techniques, can yield new insights about per se well-researched grammatical alternations.

References

Baayen, R. Harald. 2008. *Analyzing linguistic data: a practical introduction to statistics using R*.

Cambridge, New York: Cambridge University Press.
Bernaisch, Tobias, Stefan Th. Gries & Joybrato Mukherjee. 2014. The dative alternation in South Asian English(es): Modelling predictors and predicting prototypes. *English World-Wide* 35(1). 7–31. doi:10.1075/eww.35.1.02ber.
Biber, Douglas, Stig Johansson, Geoffrey Leech, Susan Conrad & Edward Finegan. 1999. *Longman grammar of spoken and written English*. Harlow: Longman.
Bolinger, Dwight. 1972. *That's that*. The Hague / Paris: Mouton.
Breiman, Leo. 2001. Random Forests. *Machine Learning* 45(1). 5–32.
Breiman, Leo, Jerome H. Friedman, Richard Olshen & Charles J Stone. 1984. *Classification and regression trees*. Belmont: Wadsworth International Group.
Cedergren, Henrietta J & David Sankoff. 1974. Variable rules: Performance as a statistical reflection of competence. *Language* 50(2). 333–355.
Chambers, Jack K. & Peter Trudgill. 1998. *Dialectology*. 2nd ed. Cambridge, New York: Cambridge University Press.
Crawley, Michael J. 2005. *Statistics: an introduction using R*. Chichester, West Sussex, England: J. Wiley.
Dehé, Nicole & Anne Wichmann.2010. Sentence-initial *I think (that)* and *I believe (that):* Prosodic evidence for use as main clause, comment clause and discourse marker*. Studies in Language* 34(1): 36–74.
Dixon, Robert M. W. 1991. *A new approach to English grammar, on semantic principles*. Oxford: Clarendon.
Dor, Daniel. 2005. Toward a semantic account of that-deletion in English. *Linguistics* 43(2). 345–382.
Elsness, Johan. 1984. That or zero? A look at the choice of object clause connective in a corpus of American English. *English Studies* 65. 519–533.
Ferreira, Victor S. 2003. The persistence of optional complementizer production: Why saying a "that" is not saying "that" at all. *Journal of Memory and Language*(48). 379–398.
Ferreira, Victor S. & Gary S. Dell. 2000. Effect of ambiguity and lexical availability on syntactic and lexical production. *Cognitive Psychology* 40. 296–430.
Finegan, Edward & Douglas Biber. 1995. That and zero complementisers in Late Modern English: exploring ARCHER from 1650--1990. In Bas Aarts & Charles F. Meyer (eds.), *The verb in contemporary English: Theory and description*, 241–257. Cambridge: Cambridge University Press.
Finegan, Edward & Douglas Biber. 2001. Register variation and social dialect variation: The register axiom. In Penelope Eckert & John R. Rickford (eds.), *Style and sociolinguistic variation*, 235–267. Cambridge: Cambridge University Press.
Garnsey, Susan M., Neal J. Pearlmutter, Elizabeth Myers & Melanie Lotocky. 1997. The contributions of verb bias and plausibility to the comprehension of temporarily ambiguous sentences. *Journal of Memory and Language* 37. 58–93.
Gries, Stefan Th. 2005. Syntactic Priming: A Corpus-based Approach. *Journal of Psycholinguistic Research* 34(4). 365–399. doi:10.1007/s10936-005-6139-3 (27 January, 2013).
Gries, Stefan Th. & Martin Hilpert. 2010. Modeling diachronic change in the third person singular: a multifactorial, verb- and author-specific exploratory approach. *English Language and Linguistics* 14(03). 293–320. doi:10.1017/S1360674310000092 (26 January, 2013).
Hawkins, John A. 2001. Why are categories adjacent? *Journal of Linguistics* 31. 1–34.
Hawkins, John A. 2003. Why are zero-marked phrases closer to their heads? In Günter Rohdenburg & Britta Mondorf (eds.), *Determinants of grammatical variation in English*,

175–204. (Topics in English Linguistics). Berlin / New York: Mouton de Gruyter.
Hawkins, John A. 2004. *Efficiency and complexity in grammars*. Oxford New York: Oxford University Press.
Hernández, Nuria. 2006. *User's Guide to FRED.* Freiburg: University of Freiburg. URN:nbn:de:bsz:25-opus-24895, URL: http://www.freidok.uni-freiburg.de/volltexte/2489/.
Hinrichs, Lars, Benedikt Szmrecsanyi & Axel Bohmann. 2015. Which-hunting and the Standard English Relative Clause. *Language* 91(4).
Jaeger, Florian T. 2006. Redundancy and Syntactic Reduction in Spontaneous Speech. Stanford University PhD Thesis.
Kaltenboeck, Gunther. 2006a. `... that is the question´: Complementizer omission in extraposed that-clauses. *English Language and Linguistics* 10. 371–396.
Kaltenboeck, Gunther. 2006b. Zur Verwendung von that und Asyndeton in extraponierten Subjektsätzen des Englischen: Eine korpuslinguistische Untersuchung. In Bernhard Kettemann & Georg Marko (eds.), *Planing, gluing and painting corpora*, 69–99. Frankfurt am Main: Peter Lang.
Kearns, Kate. 2007. Epistemic verbs and zero complementiser. *English Language and Linguistics* 11. 475–505.
Kinsey, Rafe H., Tim Florian Jaeger & Thomas Wasow. 2007. *What does THAT mean? Experimental evidence against the principle of no synonymy*.
Kolbe, Daniela. 2008. Complement clauses in British Englishes. University of Trier PhD Thesis.
Kolbe-Hanna, Daniela & Benedikt Szmrecsanyi. in press. Grammatical variation. In Douglas Biber & Randi Reppen (eds.), *The Cambridge Handbook of English Corpus Linguistics*. Cambridge: Cambridge University Press.
Labov, William. 1982. Building on empirical foundations. In Winfred Lehmann & Yakov Malkiel (eds.), *Perspectives on Historical Linguistics*, 17–92. Amsterdam, Philadelphia: Benjamins.
McDavid, Virginia. 1964. The alternation of "that" and zero in noun clauses. *American Speech* 39. 102–113.
Pampel, Fred. 2000. *Logistic Regression. A Primer*. (Quantitative Applications in the Social Sciences). Thousand Oaks: Sage Publications.
Pinheiro, José C. & Douglas M. Bates. 2000. *Mixed-Effects Models in S and S-PLUS*. New York: Springer.
Quirk, Randolph, Sidney Greenbaum, Geoffrey Leech & Jan Svartvik. 1985. *A comprehensive grammar of the English language*. London: Longman.
R Development Core Team. 2013. *R: A Language and Environment for Statistical Computing*. Vienna, Austria. http://www.R-project.org/.
Rohdenburg, Günter. 1996. Cognitive complexity and increased grammatical explicitness in English. *Cognitive Linguistics* 7. 149–182.
Rohdenburg, Günter. 1999. Clausal complementation and cognitive complexity in English. In F.-W. Neumann & S. Schülting (eds.), *Anglistentag 1998 Erfurt*, 101–112. Trier: Wissenschaftlicher Verlag.
Roland, Douglas, Jeffrey L. Elman & Victor S. Ferreira. 2006. Why is that? Structural prediction and ambiguity resolution in a very large corpus of English sentences. *Cognition* 98(3). 245–272. doi:10.1016/j.cognition.2004.11.008 (27 August, 2013).
Schneider, Ulrike. 2014. Frequency, Chunks and Hesitations: A Usage-based Analysis of Chunking in English. University of Freiburg PhD dissertation. http://www.freidok.uni-freiburg.de/volltexte/9793/.

Shih, Stephanie. 2011. Random Forests for Classification Trees and Categorical Dependent Variables: an informal Quick Start R Guide. *Random forests for categorical dependent variables: an informal quick start R guide.[Online] Available from http://www. stanford. edu/∼stephsus/R-randomforest-guide. pdf [Accessed 25th July 2012]*.

Staum, Laura. 2005. When stylistic and social effects fail to converge: A variation study of complementizer choice. Stanford University, ms.

Storms, G. 1966. That-clauses in modern English. *English Studies* 47. 249–270.

Strobl, Carolin, Anne-Laure Boulesteix, Thomas Kneib, Thomas Augustin & Achim Zeileis. 2008. Conditional Variable Importance for Random Forests. *BMC Bioinformatics* 9(1). 307. doi:10.1186/1471-2105-9-307 (28 August, 2013).

Swart, Peter de, Hanne M. Eckhoff & Olga Thomason. 2012. A Source of Variation: A corpus-based study of the choice between APO and EK in the NT Greek Gospels. *Journal of Greek Linguistics* 12(1). 161–187. doi:10.1163/156658412X649760 (8 May, 2015).

Szmrecsanyi, Benedikt. 2005. Language users as creatures of habit: A corpus-based analysis of persistence in spoken English. *Corpus Linguistics and Linguistic Theory* 1(1). 113–150. doi:10.1515/cllt.2005.1.1.113 (27 May, 2011).

Szmrecsanyi, Benedikt. 2006. *Morphosyntactic persistence in spoken English: a corpus study at the intersection of variationist sociolinguistics, psycholinguistics, and discourse analysis*. Berlin, New York: Mouton de Gruyter.

Szmrecsanyi, Benedikt. 2011. Corpus-based dialectometry: a methodological sketch. *Corpora* 6(1). 45–76. (27 May, 2011).

Szmrecsanyi, Benedikt. 2013. *Grammatical variation in British English dialects: a study in corpus-based dialectometry*. Cambridge, New York: Cambridge University Press.

Szmrecsanyi, Benedikt & Nuria Hernández. 2007. *Manual of Information to accompany the Freiburg Corpus of English Dialects Sampler ("FRED-S")*. Freiburg: University of Freiburg. URN: \texttturn:nbn:de:bsz:25-opus-28598, URL: http://www.freidok.uni-freiburg.de/volltexte/2859/.

Tagliamonte, Sali & R. Harald Baayen. 2012. Models, forests, and trees of York English: Was/were variation as a case study for statistical practice. *Language Variation and Change* 24(02). 135–178. doi:10.1017/S0954394512000129 (27 August, 2013).

Tagliamonte, Sali & Jennifer Smith. 2005. No momentary fancy! The zero "complementizer" in English dialects. *English Language and Linguistics* 9. 289–309.

Thompson, Sandra & Anthony Mulac. 1991. The discourse conditions for the use of the complementizer that in conversational English. *Journal of Pragmatics* 15(3). 237–251. doi:10.1016/0378-2166(91)90012-M (27 August, 2013).

Torres Cacoullos, Rena & James A. Walker. 2009. On the persistence of grammar in discourse formulas: a variationist study of that. *Linguistics* 47(1). 1–43. doi:10.1515/LING.2009.001 (26 August, 2013).

Trueswell, John C., Michael K. Tanenhaus & Christopher Kello. 1993. Verb-specific constraints in sentence processing: Separating effects of lexical preference from garden paths. *Journal of experimental psychology: Learning, Memory, and Cognition* 19. 528–553.

Weiner, Judith & William Labov. 1983. Constraints on the agentless passive. *Journal of Linguistics* 19. 29–58.

Wolk, Christoph, Joan Bresnan, Anette Rosenbach & Benedikt Szmrecsanyi. 2013. Dative and genitive variability in Late Modern English: Exploring cross-constructional variation and change. *Diachronica* 30(3). 382–419. doi:10.1075/dia.30.3.04wol (16 November, 2013).

Wulff, Stefanie, Nicholas Lester & Maria T. Martinez-Garcia. 2014. That-variation in German and Spanish L2 English. *Language and Cognition* 6(02). 271–299. doi:10.1017/langcog.2014.5

(28 April, 2015).
Yaguchi, Michiko. 2001. The function of non-deictic that in English. *Journal of Pragmatics* 33. 1125–1155.