Grammatical variation

Daniela Kolbe-Hanna (University of Trier)
Benedikt Szmrecsanyi (University of Manchester)

1 A discussion of previous research in this area

In this article, we use a fairly liberal definition of "grammatical variation", including both genuinely variationist research – where grammatical variants are modeled as competing against each other – and text-linguistic research that explores variable text frequencies of particular grammatical constructions in corpora.

Corpus-based research on grammatical variation is a wide research area, so the review we are offering is somewhat selective. By and large, research along these lines can be categorized into five groups:

1. *Variationist sociolinguistics*: Researchers in this tradition are typically interested in how linguistic variation is conditioned by language-internal and language-external factors. Although early studies largely relied on the researcher's personal notes taken while listening to the audio recordings, state-of-the-art work is usually based on corpora consisting of fully transcribed sociolinguistic interviews (see Tagliamonte 2007). The bulk of this literature is concerned with phonological variation; however, grammatical variation has been subject to study since at least the 1980s (Torres Cacoullos and Walker 2009a; MacKenzie 2013; Poplack and Dion 2009; Poplack and Tagliamonte 1996; Scherre and Naro 1991; Tagliamonte and Temple 2005; Travis 2007; Weiner and Labov 1983)

2. *Diachronic linguistics*: In this department we find work exploring grammatical variation with regard to resultant grammatical change – short-term or long-term, completed or in progress – on the basis of historical corpora (Biber and Gray 2011; Gries and Hilpert 2010; Hinrichs and Szmrecsanyi 2007; Hundt 2004; Nevalainen 1996; Raumolin-Brunberg 2005; Taylor 2008; Wolk et al. 2013).

3. *Register/genre/text type analysis*: Work in this tradition is often focused on lexis, but many studies include features that come within the remit of grammatical variation (e.g. Biber 1988; Biber and Finegan 1989; Hundt and Mair 1999). Some analysts are exclusively concerned with grammatical variation (see Grafmiller to appear; and the papers in Dorgeloh and Wanner 2010). The Longman Grammar (Biber et al. 1999) backs up its comprehensive grammatical description of English with information about text type variation.

4. *Dialectology*: While dialectology, by and large, is phonology-focused and atlas-based, corpus-based research into grammatical variation within or between dialects, or regional varieties, of a language has become increasingly popular since relatively recently (Anderwald 2009; Grieve 2012; Steger and Schneider 2012; Szmrecsanyi 2013; Tagliamonte and Smith 2002, and the contributions in Kortmann et al. 2005; Hernandez, Kolbe, and Schulz 2012).

5. *Knowledge, processing, cognition*: Grammatical variation has also been the subject of corpus-based psycholinguistic research (Gries 2005; Jaeger 2006; Reitter,

Moore, and Keller 2010), of work in cognitive linguistics (De Smet and Cuyckens 2005; Grondelaers and Speelman 2007; Hilpert 2008), and of research concerned with the nature of linguistic knowledge (Bresnan et al. 2007; Adger and Smith 2010).

In what follows, we discuss five studies, each one exemplifying one of the categories defined above.

Variationist sociolinguistics: Weiner and Labov (1983) is a comparatively dated study, but it illustrates well the variationist sociolinguistic approach to grammatical variation. It draws on a corpus of agentless sentences produced in sociolinguistic interviews (this "corpus" may not live up to contemporary standards of corpus compilation, but it comes within the remit of our notion of corpus linguistics) The paper is concerned with a syntactic variable, the alternation between agentless passives, as in (1a), and generalized, "empty" actives, as in (1b), in spoken English:

(1a) The liquor closet got broken into.

(1b) They broke into the liquor closet. (Weiner and Labov 1983, 34)

As is customary in the variationist sociolinguistic literature, Weiner and Labov assume semantic interchangeability between the two variants. Using a private corpus containing sociolinguistic interviews attesting 1,489 relevant constructions, Weiner and Labov conduct, among other things, multivariate Variable Rule (Varbrul) analyses to determine the influence of several external and internal factors on the choice of generalized actives or agentless passives in their corpus. The language-external conditioning factors studied are careful vs. casual style, sex, and social class; the language-internal constraints subject to study are information status, parallelism in surface structure, and whether or not a passive was used anywhere in the five preceding clauses. Weiner and Labov find, in short, that the single most powerful factor influencing the choice of actives vs. passives is repetition of previous structure.

Diachronic linguistics: Gries and Hilpert (2010) draw on the syntactically parsed Corpus of Early English Correspondence (PCEEC) to explore morphological variation during the transition period, between 1417 and 1681, from third person singular *–(e)th* (as in *he giveth*) to *–(e)s* (as in *he gives*). From the parsed data source, Gries and Hilpert extract about 20,000 relevant observations of *–(e)th* or *–(e)s*. They use a variant of cluster analysis to derive periods in a bottom-up fashion, using text frequencies of the variant suffixes to derive a similarity measure. Thus, based on the structure of the dataset, Gries and Hilpert distinguish five intervals: 1417–1478, 1479–1482, 1483–1609, 1610–1647, and 1648–1681. This periodization is one of the explanatory variables that is subsequently fed into a mixed-effects regression analysis that models the change from *–(e)th* to *–(e)s* considering both language-external factors (such as bottom-up periods, or author gender) and language-internal constraints (e.g. does the following word begin in a fricative?) to predict the observable variation, which the model does with a 95% success rate. The model also takes care of author idiosyncrasies and verb lemma effects by treating these independent variables as random effects. Crucially, Gries and Hilpert check for interactions between the periodization they use and the language-internal constraints considered in the model to establish if and to what extent particular language internal constraints are fluctuating diachronically. Along these lines it turns out, for instance, that certain phonological factors play a role only in some periods but not in others.

Register/genre/text type analysis: Hundt and Mair (1999) explore the original Brown family of corpora, a set of four 1-million word parallel corpora sampling Standard written British and American English from the 1960s and the 1990s. The Brown corpora cover a variety of written text types, and Hundt and Mair propose a continuum of openness to innovation ranging from "agile" to "uptight" genres. Specifically, Hundt and Mair compare the press and academic prose sections of the Brown corpora, and demonstrate that the two genres differ in terms of innovativeness and conservativeness: press is an "agile" genre, and academic prose is "uptight". On the methodological plane, Hundt and Mair investigate features that are suspected to contribute to the growing "colloquialization" of the norms of written English. Such features include first and second person pronouns, sentence-initial conjunctions, contractions, phrasal and phrasal-prepositional verbs, passive constructions, abstract nouns ending in *–tion, -ment, -ness, -ity*, usage of the progressive construction, usage of bare infinitives after *help*, and usage of the preposition *upon*. For example, phrasal verbs (as in *he made the story up*) are considered the colloquial variant of more bookish Latinate verbs, such as *invent*. This is why Hundt and Mair explore text frequencies of phrasal verbs with *up*, implicitly assuming that frequency increases are at the expense of more bookish alternatives. It turns out that in press writing, both the type and token frequency of phrasal verbs has increased between the 1960s and the 1990s – in academic writing, by contrast, type and token frequencies are rather stable or even decreasing.

Dialectology: Szmrecsanyi (2013) explores the extent to which grammatical variation in British English dialects is structured geographically – and thus, is sensitive to the likelihood of social contact. The study draws on a methodology ("corpus-based dialectometry") that combines corpus-based variation studies with aggregative-dialectometrical analysis and visualization methods. It is proposed that this synthesis is desirable for two reasons. First, multidimensional objects, such as dialects, call for aggregate analysis techniques. Second, compared to linguistic atlas material, corpora yield a more trustworthy frequency signal. Against this backdrop, Szmrecsanyi calculates a joint measure of dialect distance based on the discourse frequency of dozens of morphosyntactic features, such as, e.g., multiple negation (e.g. *don't you make no damn mistake*), non-standard verbal *-s* (e.g. *so I says, What have you to do?*), or non-standard weak past tense and past participle forms (e.g. *they knowed all about these things*) in data from the 2.7-million word Freiburg Corpus of English Dialects, which covers dialect interviews all over Great Britain. The ultimate aim is to reveal large-scale patterns of grammatical variability in traditional British English dialects. The study shows, for example, that it is impossible to find a clearly demarcated Midlands dialect area on grammatical grounds, and that travel time is a better predictor of linguistic distance than as-the-crow-flies geographic distance.

Knowledge, processing, cognition: Bresnan et al. (2007) explore the well-known dative alternation, viz. the syntactic variation between the double object structure, as in (2a), and the prepositional dative structure, as in (2b)

(2a) He gave the children toys

(2b) He gave toys to the children

Bresnan et al. calculate several logistic regression models that all correctly predict more

than 90% of the actual dative outcomes in the Switchboard collection of recorded telephone conversations and in a corpus of Wall Street Journal texts. Bresnan et al.'s models draw on a wide range of explanatory variables to account for speakers' dative choices: semantic class, accessibility of the recipient, accessibility of the theme, pronominality of the recipient, pronominality of the theme, definiteness of the recipient, definiteness of the theme, animacy of the recipient, person of the recipient, number of the recipient, number of the theme, concreteness of the theme, structural parallelism in dialogue, and length difference between theme and recipient. In addition, the study employs bootstrapping techniques and mixed-effects modeling to investigate issues such as the role of idiolectal differences and the validity of cross-corpus generalizations. Besides showing that intuitions are a poor guide to understanding the dative alternation, Bresnan et al. conclude (i) that "linguistic data are more probabilistic than has been widely recognized in theoretical linguistics" (Bresnan et al. 2007, 91), (ii) that naturalistic corpus data can be married to sophisticated statistical analysis techniques to address issues in linguistic theory, and (iii) that probabilistically constrained linguistic variation is every bit as interesting as categorical patterns.

1.1 Discussion of methods

Most of the research in the variationist sociolinguistic tradition is based on studies of "private" (D'Arcy 2011: 55) corpora. These kinds of corpora comprise, data that were gathered specifically for a project, such as in Weiner and Labov (1983). The particular text type sampled is usually sociolinguistic interviews, which more often than not are conducted in order to analyze a specific variable. Unfortunately, the data are not usually made public or generally available to other researchers. Studies that draw on publicly available corpora, such as the Corpus of Early English Correspondence (CEEC) in Gries and Hilpert (2010), are arguably more reliable and replicable, as anyone interested in the topic is able to repeat the analysis, and to add and remove variables as fits their own research interest. This increases the transparency of research.

Rich grammatical annotation in corpora such as the parsed version of the CEEC employed in Gries and Hilpert (2010) enable researchers to conduct feature extraction and annotation more easily, relying on decisions made in the process of parsing. Dealing with corpora that are not parsed makes extraction and annotation more laborious. Any decisions on assigning a category are left solely to the researcher and are thus more subjective. Very often, however, there is no choice left if the corpus subject to analysis is not parsed. Sometimes it might also seem beneficial to ignore existing parsing to have control over all of the data input.

Traditionally the focus has been on determining the conditioning of grammatical variation, so studies such as Bresnan et al (2007) consider variation as explanandum: how can we explain regularities in variation? What are the factors that favor usage of one or the other variant – for example, which factors favor the prepositional dative structure? In another line of research, however, grammatical variation is seen as explanans, so that the aim is not to show the causes, but the effects of variation. For example, Szmrecsanyi (2013) (see Grieve 2012 for a smilar approach) is interested in how micro-variation in dozens of grammatical features "gangs up", as it were, to create the big picture – that is, dialect areas and dialect continua. In a similar vein, in much register-analytic research

(consider, for example, Biber 1988) the question is how grammatical variation, which is seen as having functional motivations, creates register differences. When studies observe the constraints and conditions of variation, some rely on univariate analyses that assess the effect of one independent factor at a time on the variable in question (e.g. Hundt 2004). Other (multivariate) studies analyze the joint impact of a multitude of independent variables on a dependent variable (e.g. Bresnan et al. 2007). Since many grammatical variables have proved to be conditioned by a multitude of factors, analysts should attempt to include as many of these factors as possible. For certain variables, however, this might not be possible (very little is known, for instance, about the choice between *if* and *whether* in interrogatives).

As far as analysis tools are concerned, Varbrul has dominated the market in Labov-type variationist sociolinguistics in particular. Varbrul is designed to calculate the stochastic influence of multiple independent variables on grammatical variation (e.g. Weiner and Labov 1983). Recently, however, the dominance of Varbrul in variationist linguistics has been challenged by competitors such as the open source software R, which makes possible sophisticated mixed-effect modeling (as implemented, e.g., in the package lme4), which has the following advantages over Varbrul: First,unlike Varbrul, it can take care of non-repeatable nuisance factors such as idiolectal differences; thus, it can control for the fact that individual speakers may deviate strongly from average behavior and are not representative of the population. Second,unlike Varbrul, R can handle continuous independent predictors, while predictors in Varbrul must be discrete and categorical. A continuous variable or predictor is, for instance, mean length (in orthographic characters) of words. By contrast, categorical variables consist of distinct (discrete) categories such as 'male' versus 'female'. In a Varbrul analysis, mean word length would have to be transposed into discrete categories such as < 3 characters vs. ≥ 4 characters – a kind of data reduction that is more often than not not desirable. Third, unlike Varbrul, R can also handle interaction effects, such as the interaction between speaker age and speaker sex, i.e. differences in behavior between older (or younger) men vs. women. Varbrul, however, can handle influence of a speaker's sex and a speaker's age separately, but not in interaction. We hasten to add that the Varbrul software has been refurbished in the form of Rbrul (Johnson 2008), but it is not clear to us why variationist sociolinguists would need an idiosyncratic software tool, given that lme4 is available and widely used in the social sciences at large.

In any event, it has been a hallmark of variationist linguistics to study relative frequencies or usage rates of a variant vis-a-vis another variant that it competes with (e.g. Weiner and Labov 1983, Gries and Hilpert 2010); the level of granularity is such that individual linguistic choices take center stage. By contrast, corpus linguists often adopt a text-linguistic perspective, investigating absolute frequencies of a phenomenon per, for instance, a million words of running text (see, for example, Hundt and Mair 1999, Szmrecsanyi 2013). Biber, Egbert, Gray, Oppliger and Szmrecsanyi (submitted) offer an in-depth discussion of the differences between the two approaches.

In conclusion, corpus-based research has shown that grammatical variation, like phonological variation, can be sociolinguistically conditioned. Also, we now know that we can trace back the development from variation to language change in historical corpora, as do, for instance, Gries and Hilpert (2010). Further, although some scholars have suggested that grammatical variation is not sensitive to geography, recent research

(e.g. Szmrecsanyi 2013) has demonstrated that grammatical variation can provide a geolinguistic signal, especially when joining many variables at a time to paint a larger picture.

It also turns out that speakers implicitly know about (probabilistic) aspects of grammatical variation. For instance, in a series of experiments Joan Bresnan (Bresnan 2007) has shown that the intuitions native speakers have about the acceptability of dative structures (double object versus prepositional dative) in particular contexts correlate significantly with the probabilities that corpus models calculate (such as the ones reported in Bresnan et al. 2007), given the same co(n)text.

An interesting issue that remains to be resolved would be to link up the variation-as-explanandum approach and the variation-as-explanans-approach. An example will be provided in our case study in Section 2, which models constraints on complementizer *that* deletion (variation-as-explanandum), but then goes on to utilize part of the probabilistic output of this analysis to explore how complementizer *that* variation engenders dialectological differences (variation-as-explanans). Wolk (in preparation) is a study that more systematically explores this interface between the two approaches, integrating the aggregational approach to language variation as exemplified by dialectometry (for example, Szmrecsanyi 2013), and the probabilistic modeling of language variation customarily utilized by probabilistic grammarians (Bresnan et al. 2007). Another important aspect to keep in mind is that any results can only be as good as the corpora they are based on. When we use statistic models created for representative samples we must make sure to establish clear criteria of what counts as a representative corpus.

2 Case study: Variation in the use of the complementizer *that*

Our case study is an exercise in variationist analysis: we investigate grammatical variation in the use of the complementizer in *that* clauses, such as in (3), where speakers have the choice between retaining the explicit complementizer *that*, and omitting it. The latter option is also referred to as the use of the zero variant.

(3) We think (that) these worries are common.

While many factors have proved to be influential in variable *that* omission, only few studies have included language-external factors (e.g. Staum 2005). We add this additional perspective to previous research and explore how language-internal and language-external factors interact to engender linguistic variation. In this endeavor, we use data from FRED (the Freiburg English Dialect Corpus), which samples dialect speech from all over Great Britain. Following our typology in Section 1, our study is situated at the intersection between variationist sociolinguistics, dialectology, and research on knowledge, processing, and cognition.

On the technical plane, we employ multivariate analysis, in particular mixed-effects logistic regression as implemented in the lme4 package, to include as many factors as possible known to condition this variation, and to compare the strength of each factor with the strength of each other factor. We also account for the lack of repeatability in certain independent variables by treating them as random effects. This means rather than calculating the strength of influence of these variables, the model adjusts the calculation according to the bias of these variables (cf. Baayen 2008, 241-242). A typical

random effect is variation by subject, which, in our case, is equal to variation by speaker. A new sample of the same population would result in a different set of speakers with different idiosyncratic preferences, so the set of speakers in the dataset is random. Since individual speakers may have very strong preferences for *that* or its omission that are not representative of the whole population, by-subject variation may skew the results. It is therefore useful to fit a model that takes this skewing into account. In contrast, fixed variables, such as verb morphology, are repeatable and not random, because they would be the same in a different sample – one would always assign the same value (past, base form etc.) to the same form of a verb.

2.1 Previous research on the retention or omission of *that*

The variation in the use of the *that*-complementizer has been the object of an abundance of linguistic research. Research in cognitive linguistics as well as in psycholinguistics has shown that the cognitive complexity of an utterance plays an important role in a speaker's choice to use or not to use the explicit complementizer. This choice reflects speakers' effort to find a balance between explicitness and economy. While the retention of *that* explicitly marks the subsequent clause as an embedded clause and is thus more precise, the omission of the complementizer reduces the production effort by rendering a shorter utterance. Consequently, the omission of *that* is in general preferred in linguistically less complex environments, where less explicitness is needed to signal that the following linguistic material is a complement clause (Rohdenburg 1996, 1999, Jaeger 2006, 74-89, Hawkins 2003). Whereas cognitive complexity is a language-internal issue, Kolbe (2008, 90-129) shows that language-external factors such as age, sex and dialectal preference may also influence a speaker's choice between zero and *that* (cf. Staum 2005).

2.2 Data and methods

FRED consists of roughly 2.7 million words of dialect speech (Hernández 2006; see Szmrecsanyi and Hernández 2007 for the publicly available sampler version). As the retention or omission of *that* is determined by many cognitive factors, spoken language is a crucial resource to examine the influence of those factors. The texts in FRED derive from interviews with speakers from England, Wales and Scotland. The corpus files mostly consist of transcripts of oral history interviews and thus offer a style of speech that is casual and adapted to the interview format (Hernández 2006).

We chose to restrict attention to the 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 clauses. Note that we are fully aware of the fact that these verbs commonly occur with omitted *that*. Based on a Perl script identifying these verbs in FRED, the beginning and the end of each embedded *that* clause following these verbs was coded manually by the two authors. Tests for inter-coder reliability in samples of clauses showed that the two coders agreed in 83% of all cases.[1] We then used another Perl script to extract the *that* clauses identified in this manner, their matrix verb phrase(s) and the corresponding meta-data (speaker, local origin, file name, etc.).

[1] 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.

Drawing on previous research on the choice between the omission and the retention of *that* (see above), we identified as independent variables those aspects of each embedded clause and its matrix clause that are likely to influence a speaker's choice between explicit and zero *that*. Szmrecsanyi (2006) shows that grammatical variants tend to persist in speech, so that speakers are more likely to use the variant they have used before. We therefore investigate in how far this is true for the use of explicit *that*. The following section (2.2.1) provides a description of all variables used in this study.

2.2.1 Variables

The dependent variable in our study is binary: retention or omission of the complementizer *that*. The independent variables in our annotation layer are detailed in the following.

Language external factors

As Kolbe (2008) and Staum (2005) observe, sociolinguistic factors may influence a speaker's choice of complementizer. Based on the FRED metadata (see Hernández 2006 and Szmrecsanyi and Hernández 2007), we included the variables TEXT, AREA, COUNTY, and SPEAKER.

- TEXT indicates the corpus file (e.g. SFK_018 or IOM_002) where the token occurred.
- AREA codes the nine dialect area as specified in FRED. These are Hebrides, Isle of Man, English Midlands, Northern England, Scottish Highlands, Scottish Lowlands, Southeast England, Southwest England, and Wales.
- COUNTY specifies the county in which the speaker lived at the time of recording.
- SPEAKER renders the current speaker's ID (as defined in the FRED manual).

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.

Language-internal factors

According to Rohdenburg's "complexity principle" (1996, 1999) more explicit variants are preferred in cognitively more complex environments. Most of our independent linguistic variables are related to the cognitive complexity of an utterance.

1. *Features of the matrix clause*

As previous research has shown (Torres Cacoullos and Walker 2009b, Jaeger 2006, 88-89, Dor 2005, Biber 1999, 681), the choice of matrix verb strongly affects the likelihood of the retention of *that*: speakers tend to omit the complementizer after using a matrix verb that frequently controls embedded *that* clauses. The variable VERB thus specifies which matrix verb is used (*think, say* or *know)*. All three verbs are highly frequent and they frequently function as matrix verbs of *that* clauses (Biber et al. 1999,

668). However, *think* and *say* are by far the most frequent matrix verbs of *that* clauses in British English and their use strongly favors the omission of *that* (Torres Cacoullos and Walker 2009b, 19-20, Biber et al. 1999, 681, Thompson and Mulac 1991, 244-245).

Further features of the verb phrase in the matrix clause that affect a speaker's choice between zero and *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* occurs (see Torres Cacoullos and Walker 2009b, 24-27, Biber et al. 1999, 681-682, Rohdenburg 1996, 161, Thompson and Mulac 1991, 246). These features are captured in the variables VERBMORPH, MATRIX_NEGATION, and MATRIX_AUXILIARY.

- VERBMORPH codes whether the matrix verb occurs as base form, as third person singular present tense, as past (tense or participle), or as *–ing* form.
- MATRIX_NEGATION specifies whether the matrix verb is negated.
- MATRIX_AUXILIARY states whether the matrix verb is preceded by a modal auxiliary (e.g. *should, could, would, will, 'll, shall, must, can* + negated forms).

A third feature of the matrix clause that has an impact on a speaker's choice of complementizer is its subject. When the matrix clause subject is *I* or *you*, the omission of *that* becomes more likely (Torres Cacoullos and Walker 2009b, 24-25, Thompson and Mulac 1991, 242-243). The appearance of the matrix subject is determined by MATRIX_SUBJECT_TYPE, which distinguishes between *I, you, it* and any other subject.

If the matrix clause is *I think*, the retention of *that* is highly unlikely. This clause does not only comprise all features that favor the omission of *that* (subject *I*, simple verb morphology, highly frequent verb lemma that very often controls *that* clauses), it also functions as a comment clause or epistemic parenthetical (see, for example, Thompson and Mulac 1991). For reasons that will be discussed in more detail in section 2.4, we consider it a matrix clause, but we use the variable I_THINK to distinguish this matrix clause from other ones.

The variables MORPHID and VERBID combine features of the matrix verb that are potentially relevant to grammatical persistence as discussed in Szmrecsanyi (2005, 2006). They check whether a matrix verbs has the same morphology, respectively verb lemma, as the previous verb or not.

2. *Features of the embedded* that *clause*

Whether speakers choose to retain or omit a *that* complementizer has also proved to depend on features of the embedded clause itself. The cognitive complexity of the embedded clause is gauged by means of the following variables:

- EMB_CL_LENGTH specifies the number of words in the embedded clause (excluding the complementizer when present), since speakers use the explicit complementizer more often in longer clauses (Jaeger 2006, 85, Rohdenburg 1999, 164). In the analysis, we use a logarithmic transformation, LOG_EMB_CL_LENGTH, to reduce skewing and outliers in the data, as is customary in quantitative modeling (see Baayen 2008: 31).
- Material between matrix verb and beginning of the embedded clause: *that* is more often retained if any linguistic material occurs between matrix verb and embedded clause (Rohdenburg 1996, 160, Hawkins 2003, 178–179). We distinguish between INTERV_MATERIAL_MACL_EMBCL, which specifies the number of orthographic words between the matrix verb and the start of the embedded clause, and

ADV_BEGINNING, which specifies whether an adverbial *(in, because, cause, if, since, when, after, before, during*) occurs at the beginning of the embedded clause. We also included ADV_AFTEREND, which shows whether an adverbial (*in, because, cause, if, since, when, after, before, during*) occurs after the end of the embedded clause.

- COMPLEMENT_SUBJECT states whether the first element in the embedded clause is a pronoun, in which case previous research has shown a stronger tendency to omit *that* (Torres Cacoullos and Walker 2009b, 24, 28, Rohdenburg 1999, 162).
- SAME_VERB_IN_EMBD_CL indicates if the exact same verb as the matrix verb occurs in the embedded clause, which we assume to be a possible factor to decrease cognitive complexity.
- HORROR_AEQUI checks whether the embedded clause after the complementizer (explicit or zero) starts with *that*, as in *I think that that man is his father*. There is evidence from non-finite clauses that speakers avoid to use identical forms consecutively (Kolbe 2008, 217-218, 222-224).

3. *Features across clauses*

The scope of the following variables goes beyond clause boundaries:

- Speech perturbations: Jaeger shows an effect of disfluency on a speaker's choice of complementizer (2006, 91–92). "Production difficulties" increase the likelihood of *that*-retention. 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. Jaeger speculates that speakers use the complementizer to signal production difficulties. In our study, EHMS_ETC_NARROW counts the number of speech perturbations in the immediate context from three words before the matrix verb to the end of the embedded clause. EHMS_ETC_BROAD counts the number of speech perturbations in the wider context from 100 words before the matrix verb to the next finite verb.
- Grammatical persistence: As speakers tend to reuse a form they have used previously (Szmrecsanyi 2005, 2006), we take into account grammatical persistence by including three variables: ALPHA_PERSISTENCE_50 checks whether an explicit *that*-complementizer occurs up to 50 words before the complementizer, ALPHA_PERS_DISTANCE specifies the textual distance (in words) to the last occurrence of an explicit *that*-complementizer (from matrix verb to matrix verb) and BETA_PERS_DISTANCE indicates the textual distance (in words) to the last occurrence of any *that.*
- Increased complexity of a sentence may also derive from a high type-token ration, so the variable TTRPASSAGE renders the type-token ration divided by 10 in a context of -50/+50 words around the matrix verb slot.

2.3. Determinants of the choice between explicit and zero-complementizer

In order to explore the influence of the independent variables mentioned above on a speaker's choice between explicit *that* and its omission ('zero'), we analyzed their effects in a logistic regression analysis with mixed effects, i.e. one that takes into account both

random and fixed effects for the reasons stated above (see Baayen 2008: 195–208 and 278–284).

The dependent variable is the choice between the zero complementizer and its explicit form. 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 predominant value of the complementizer is its omission, which occurs in 91% of all embedded clauses in the database.

We included four independent variables, or factors, as random effects (i.e. adjustments to the intercept), since their influence is non-repeatable. VERB lemma, SPEAKER, TEXT and COUNTY are non-repeatable effects, as a second study relying on randomly chosen verbs, speakers, texts and counties would result in a different sample. VERB 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 model adjusts the intercept for each of these non-repeatable effects, to avoid skewing the results in the direction of their deviation.

At first, we created a maximal model that included all independent variables listed in Section 2.2.1. Subsequently, the model was simplified by removing factors lacking significant explanatory power (such as, e.g., MORPHID and VERBID). We started the pruning process with the least significant factors, moving to more significant ones in a stepwise fashion. Explanatory power of categorical factors with more than two levels was assessed via likelihood-ratio tests. Our final model (the "minimal adequate model") comprises the minimal amount of factors showing maximal results and the best possible fit to the data. It 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.[2] Although the model is thus validated, the predictive bonus of the model is not exactly breathtaking. This may well be caused by the predominance of the zero variant, since a baseline percentage of already 91% is very difficult to increase. Somers' Dxy, a rank correlation coefficient between predicted outcome probabilities and observed binary outcomes, is 0.77, which indicates that the model discriminates fairly between complementation types.

The fixed effects that turned out to be significant predictors of the choice between zero and explicit complementizers concern ten variables:

- matrix verb morphology (VERBMORPH),
- the subject of the matrix clause (MATRIX_SUBJECT_TYPE),
- the presence of an auxiliary in the matrix verb phrase (MATRIX_AUXILIARY),
- whether the embedded clause is controlled by *I think* (I_THINK),
- HORROR_AEQUI,
- the presence of an adverbial after the end of the embedded clause (ADV_AFTEREND),
- whether the subject of the embedded clause is a pronoun or not (COMPLEMENT_SUBJECT),
- the logarithmically transformed length of the embedded clause

[2] Multicollinearity is not an issue, as the model's condition number (κ = 12.5) is below the customary threshold of 15. The model was bootstrapped (sampling with replacement, 10 runs, the confidence intervals did not include zero) and the exclusion of outliers was analyzed and consequently rejected.

(LOG(EMB_CL_LENGTH)),

- the occurrence of *that* within 50 words before the complementizer (ALPHA_PERSISTENCE_50),
- the number of speech perturbations in the immediate context of the complementizer (logarithmically transformed as LOG(EHMS_ETC_NARROW +1)).

Table 1. contains a detailed summary of the model. The strength of each of the predictors is indicated by the value in the columns "Coefficient". These are the estimated coefficients of the respective factors, to be added or subtracted from the intercept. Negative numbers show a negative (disfavoring) influence of this predictor on the use of explicit *that*; positive numbers represent an increase in the likelihood of the retention of *that* if the respective predictor level applies. A larger estimate's value represents a stronger effect. The figure 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, or 10:10. 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 column "*p*" ('probability') indicates statistical significance, which does not refer to the strength of a predictor but its statistically reliability in the model. The lower *p* is for any value, the more reliable the effect is. A value of *p* that is less than 0.5 conventionally denotes statistical significance. For categorical variables (e.g. VERBMORPH or HORROR_AEQUI), the model relies on identifying one category as the constant or default value, with which the other values are compared. The default category in VERBMORPH is "base" – the verb used in its base form (e.g. *say*).

Within VERBMORPH the use of *that* becomes less likely when the matrix verb is in third person singular present tense or in past tense (estimates -0.73 resp. -0.35) than when it is used in its base form (the default value). Although the influence of the third present singular form is not significant at the conventional threshold of 0.05, we consider it to be marginally significant because its *p*-value at 0.1 means that there is a 90 percent chance that its effect is not due to chance. An *–ing* participle, however, increases the likelihood of the use of zero *that* (1.07).

The form of the matrix subject (MATRIX_SUBJECT_TYPE) also has a significant effect on the retention of *that*. When the matrix subject is *I* or *you* (-1.48 and -0.99 respectively), 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 (1.42).

Moreover, the form of the subject of the embedded clause influences a speaker's choice significantly. When it is or begins with *that* speakers avoid the use of the explicit complementizer, to avoid the sequence *that that* (*horror aequi*, estimate -1.04). In addition, speakers tend to omit *that* if the subject of the embedded clause is a pronoun (-0.53), as well as after *I think* (-0.70).

The following conditions increase the probability of the retention of the complementizer *that*: the persistence of a *that*-complementizer, i.e. when it has been used within 50 words before this slot (1.07), an auxiliary in the matrix verb phrase (0.72), and an adverbial at the end of the embedded clause (0.42). The longer the embedded clause is and the more speech perturbations there are between matrix and embedded clause, the more likely speakers choose *that* (LOG(EMB_CL_LENGTH): 1.08;

(LOG(EHMS_ETC_NARROW + 1): 0.79).

The strongest factor increasing the likelihood of a *that* is noted for the matrix subject *it* (1.42). The strongest factor decreasing the probability of the retention of *that* is the use of *I* as matrix subject (-1.48). Another way to express the influence of these factors is in odds ratios, which are calculated by raising *e* to the power of the regression coefficient. An odds ratio (OR) of 1 indicated that a predictor has no effect whatsoever. An OR larger than 1 indicates that a predictor increases the odds for explicit *that* (there is no upper limit). An OR from 0 to 1indicator that a predictor decreases the odds that *that* will be retained. Compared to the default category "other", which comprises the pronouns *he, she, we, they,* and any non-pronominal matrix subject, retention of *that* turns out to be over four times more likely if the matrix subject is *it* (OR 4.13). If the matrix subject is *I,* however, the odds for the retention of *that* decrease by a factor of 0.23, i.e. by 77%.

Fixed effects:

	Coefficient	**Odds Ratio**	***p***	
(Intercept)	-3.37	0.03	<0.001	***
VERBMORPH (default: base form)				
VERBMORPH: 3sg	-0.73	0.48	0.071	.
VERBMORPH: past	-0.35	0.71	0.022	*
VERBMORPH: *ing*	1.07	2.92	<0.001	***
MATRIX_SUBJECT_TYPE (default: other)				
MATRIX_SUBJECT_TYPE: *I*	-1.48	0.23	<0.001	***
MATRIX_SUBJECT_TYPE: *it*	1.42	4.13	0.005	**
MATRIX_SUBJECT_TYPE: *you*	-0.99	0.37	0.001	**
HORROR_AEQUI: yes	-1.04	0.35	0.001	**
ADV_AFTEREND: yes	0.42	1.52	0.032	*
MATRIX_AUXILIARY: yes	0.72	2.05	<0.001	***
COMPLEMENT_SUBJECT: pronoun	-0.53	0.59	<0.001	***
I_THINK: yes	-0.70	0.49	<0.001	***
ALPHA_PERSISTENCE_50: yes	1.07	2.92	<0.001	***
LOG(EMB_CL_LENGTH)	1.08	2.95	<0.001	***
LOG(EHMS_ETC_NARROW+1)	0.79	2.21	0.001	**

Random effects:

	N	**Variance**	**Standard deviation**

SPEAKER	331	0.54	0.74
VERB LEMMA	3	0.09	0.30
TEXT	310	0.17	0.42
COUNTY	38	0.33	0.57

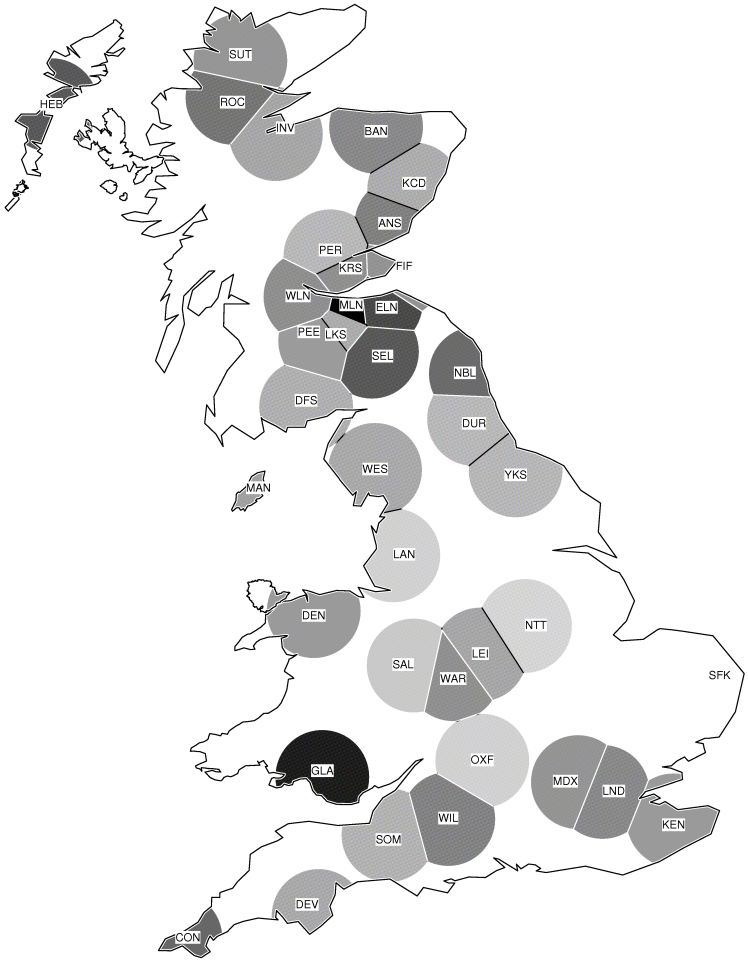
Model summary		
Somers' Dxy	0.77	
Correctly predicted	92.4%	(91.0 baseline)

Significance levels:
. marginally significant ($p < 0.1$), * significant ($p < 0.05$), **very significant ($p < 0.01$), *** highly significant ($p < 0.001$)

Table 1. Logistic regression model with fixed and random effects in complementation choice. Predicted odds are for the retention of the *that* – complementizer.

As regards the random effects, SPEAKER and COUNTY result in the strongest adjustments (standard deviations of 0.74 and 0.57 respectively). *Think* favors zero-complementizers most strongly of the three matrix verbs (of 3394 instances of *think*, only 152, or 4.5% control an explicit *that* clause), so the adjustment of the intercept is negative (-0.27) for this matrix verb, while the intercept adjustment for *know* is 0.37 and the adjustment for *say* is 0.15. It is not feasible to report all intercept adjustment for each of the 300+ speakers and texts. In the extremes, FRED speaker MlnJH favors explicit complementation least (intercept adjustment: -0.87; *that* retained in 4.3%), while speaker SRLM_HM likes it most (1.33; 48% *that* retained). FRED interview MLN_007 is least hospitable towards explicit *that* (-0.28; 4.3% *that*), and interview KEN_010 is most hospitable (0.45, 24% *that*).

The geographical distribution of complementizer choice according to county is illustrated in Map 1. This map, which translates our analysis into the sort of geolinguistic visualization customary in dialectology and dialectometry, projects intercept adjustments to geography: it shows how frequently complementizer *that* is retained in particular counties in Great Britain. In other words, Map 1 highlights how hospitable dialect speakers are towards *that* in which county. 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 more omission of *that* in Central England (including Lancashire) and more 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 (2008, 112, 120–121). In general, in Southern Great Britain *that* is less likely retained; notable exceptions are Southern Wales and Cornwall.



*Map 1.*Projection of 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).

2.4 Summary of Findings

Most of the factors that significantly influence a speaker's choice between zero and explicit *that* are concerned with cognitive complexity. 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. These cognitively less complex environments are typically the more frequent patterns, which makes the syntactic structure of the utterance more predictable (Roland et al. 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*. 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 the longer the embedded clause is.

As *I think* is a very frequent comment clause that can occur nearly everywhere in a sentence (see, for example, Kaltenböck 2008), it is disputable whether this clause actually functions as matrix in sentence-initial position. It is nearly exclusively followed by zero-*that* clauses in our data (in 97%, viz. 66 out of 2,258 occurrences). The variable I_THINK captures this distribution so that the negative influence of the matrix subject *I* on the use of *that* mostly concerns the use of *I* as subject of other verbs, such as *say, know, said, knew,* but also in clauses such as *I would think* or *I was thinking*, which have similar discourse functions as *I think* (van Bogaert 2010). In addition, clauses such as *I don't know* take on discourse functions (e.g. as "utterance launcher", see Biber et al. 1999: 1002-1004). It therefore seems impossible to maintain a distinction between "grammatical" matrix clauses and constructions with discourse functions that are not actual matrix clauses.

One factor increasing the probability of explicit *that* is not related to cognitive complexity of the current clauses, 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 factor 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, Szmrecsanyi 2005, 2006).

Some of our results come as a surprise: 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 factor, whereas an adverbial between matrix and embedded clause did not.

Many of the independent variables accounted for in our original data set (for example, INTERV_MATERIAL_MACL_EMBCL) have proved to be less influential factors than the ones actually included in the final, minimal model model. This does not mean that they are not relevant to the choice between zero and explicit *that* at all, but that for the speakers in our data set they are less decisive than other factors. Further similar studies on grammatical should take into account as many variables as feasible, since the interplay of determinants of variation needs to be re-examined for each data set.

In sum, our little case study is a corpus-based exercise in variationist (socio)linguistics because it investigates linguistic choices between the use of an overt complementizer and zero, drawing on modern multivariate analysis techniques; it is concerned with knowledge, processing, and cognition thanks to the inclusion of factors such as horror_aequi; and it falls within the remit of dialectology because it considers the effect that geography has on linguistic choices (see Map 1).

References

Adger, David, and Jennifer Smith. 2010. Variation in agreement: A lexical feature-based approach. *Lingua* 120: 1109--1134. doi:10.1016/j.lingua.2008.05.007.

Anderwald, Lieselotte. 2009. *The morphology of English dialects: Verb-formation in non-standard English*. Cambridge: Cambridge University Press.

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

using R. Cambridge: Cambridge University Press.
Biber, Douglas. 1988. *Variation across speech and writing*. Cambridge: Cambridge University Press.
Biber, Douglas, and Edward Finegan. 1989. Drift and the evolution of English style: A history of three genres. *Language* 65: 487-517.
Biber, Douglas, and Bethany Gray. 2011. Grammatical change in the noun phrase: The influence of written language use. *English Language and Linguistics* 15: 223-250. doi:10.1017/S1360674311000025.
Biber, Douglas, Stig Johansson, Geoffrey Leech, Susan Conrad, and Edward Finegan. 1999. *Longman grammar of spoken and written English*. Harlow: Longman.
Biber, Douglas, Jesse A. Egbert, Bethany Gray, Rahel Oppliger, and Benedikt Szmrecsanyi. to appear. Variationist versus text-linguistic approaches to grammatical change in English: Nominal modifiers of head nouns. *Handbook of English Historical Linguistics*.
Bresnan, Joan. 2007. Is syntactic knowledge probabilistic? Experiments with the English dative alternation. In S. Featherston and W. Sternefeld (eds.), *Roots: Linguistics in search of its evidential base*, 75-96. Berlin; New York: Mouton de Gruyter.
Bresnan, Joan, Anna Cueni, Tatiana Nikitina, and Harald R. Baayen. 2007. Predicting the dative alternation. In G. Boume, I. Kraemer, and J. Zwarts (eds.), *Cognitive foundations of interpretation*, 69-94. Amsterdam: Royal Netherlands Academy of Science.
D'Arcy, Alexandra. 2011. Corpora: Capturing language in use. In W. Maguire and A. McMahon (eds.), *Analysing variation in English*, 49-72. Cambridge: Cambridge University Press.
Dor, Daniel. 2005. Toward a semantic account of *that*-deletion in English. *Linguistics* 43: 345-382.
Dorgeloh, Heidrun, and Anja Wanner, eds. 2010. *Syntactic variation and genre*. Berlin; New York: De Gruyter Mouton.
Grafmiller, Jason. to appear. Variation in English genitives across modality and genre. *English Language and Linguistics*.
Gries, Stefan Th. 2005. Syntactic priming: A corpus-based approach. *Journal of Psycholinguistic Research* 34: 365-399. doi:10.1007/s10936-005-6139-3.
Gries, Stefan Th., and Martin Hilpert. 2010. Modeling diachronic change in the third person singular: A multifactorial, verb- and author-specific exploratory approach. *English Language and Linguistics* 14 (3): 293-320. doi:10.1017/S1360674310000092.
Grieve, Jack. 2012. A statistical analysis of regional variation in adverb position in a corpus of written standard American English. *Corpus Linguistics and Linguistic Theory* 8. doi:10.1515/cllt-2012-0003. Downloaded from http://www.degruyter.com/view/j/cllt.2012.8.issue-1/cllt-2012-0003/cllt-2012-0003.xml.
Grondelaers, Stefan, and Dirk Speelman. 2007. A variationist account of constituent ordering in presentative sentences in Belgian Dutch. *Corpus Linguistics and Linguistic Theory* 3. doi:10.1515/CLLT.2007.010. Downloaded from http://www.degruyter.com/view/j/cllt.2007.3.issue-2/cllt.2007.010/cllt.2007.010.xml.
Hawkins, John A. 2003. Why are zero-marked phrases closer to their heads? In G. Rohdenburg and B. Mondorf (eds.), *Determinants of grammatical variation in English*, 175-204. Berlin; New York: Mouton de Gruyter.

Hernández, Nuria. 2006. User's guide to FRED. Downloaded from http://www.freidok.uni-freiburg.de/volltexte/2489/pdf/Userguide_neu.pdf.
Hernández, Nuria, Daniela Kolbe, and Monika Edith Schulz, eds. 2011. *A comparative grammar of British English dialects*, v. 2□: *Modals, pronouns and complement clauses*. Berlin; Boston: De Gruyter.
Hilpert, Martin. 2008. *Germanic future constructions: A usage-based approach to language change*. Amsterdam□; Philadelphia: John Benjamins.
Hinrichs, Lars, and Benedikt Szmrecsanyi. 2007. Recent changes in the function and frequency of standard English genitive constructions: A multivariate analysis of tagged corpora. *English Language and Linguistics* 11: 437-474. doi:10.1017/S1360674307002341.
Huber, Magnus. 2012.Syntactic and variational complexity in British and Ghanaian English relative clause formation in the written parts of the International Corpus of English. In B. Kortmann and B. Szmrecsanyi (eds.), *Linguistic Complexity*, 218-242. Berlin, Boston: De Gruyter.
Hundt, Marianne. 2004. Animacy, agentivity, and the spread of the progressive in Modern English. *English Language and Linguistics* 8: 47-69. doi:10.1017/S1360674304001248.
Hundt, Marianne, and Christian Mair. 1999. "Agile" and "uptight" genres: The corpus-based approach to language change in progress. *International Journal of Corpus Linguistics* 4: 221-242.
Jaeger, T. Florian. 2006. Redundancy and syntactic reduction in spontaneous speech. PhD Thesis, Stanford University.
Johnson, Daniel Ezra. 2008. Getting off the GoldVarb Standard: Introducing Rbrul for mixed-effects variable rule analysis. *Language and linguistics compass* 3: 359-383.
Kolbe, Daniela. 2008. Complement clauses in British Englishes. PhD Thesis, University of Trier.
Kortmann, Bernd, Tanja Herrmann, Lukas Pietsch, and Susanne Wagner, eds. 2005. *A comparative grammar of British English dialects*, v. 1*: Agreement, gender, relative clauses*. Berlin; New York: Mouton de Gruyter.
MacKenzie, Laurel. 2013. Variation in English auxiliary realization: A new take on contraction. *Language Variation and Change* 25: 1-25. doi:10.1017/S0954394512000257.
Nevalainen, Terttu. 1996. Gender difference. In T. Nevalainen and H. Raumolin-Brunberg (eds.), *Sociolinguistics and language history: Studies based on the Corpus of Early English Correspondence*, 77-91. Amsterdam: Rodopi.
Poplack, Shana, and Nathalie Dion. 2009. Prescription vs. praxis: The evolution of future temporal reference in French. *Language* 85: 557-587. doi:10.1353/lan.0.0149.
Poplack, Shana, and Sali Tagliamonte. 1996. Nothing in context: Variation, grammaticization and past time marking in Nigerian Pidgin English. In P. Baker and A. Syea (eds.), *Changing meanings, changing functions: Papers relating to grammaticalization in contact languages*, 71-94. Westminster, UK: University Press.
Raumolin-Brunberg, Helena. 2005. The diffusion of subject YOU: A case study in historical sociolinguistics. *Language Variation and Change* 17: 55-73.
Reitter, David, Johanna D. Moore, and Frank Keller. 2010. Priming of syntactic rules in task-oriented dialogue and spontaneous conversation. *Proceedings of the 28th Annual*

Conference of the Cognitive Science Society: 685-690. Downloaded from http://csjarchive.cogsci.rpi.edu/Proceedings/2006/docs/p685.pdf.
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 and 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: 245-272.
Scherre, Maria, and Anthony Naro. 1991. Marking in discourse: "Birds of a feather". *Language Variation and Change* 3: 23-32.
De Smet, Hendrik, and Hubert Cuyckens. 2005. Pragmatic strengthening and the meaning of complement constructions: The case of like and love with the to-infinitive. *Journal of English Linguistics* 33: 3-34. doi:10.1177/0075424204273959.
Staum, Laura. 2005. When stylistic and social effects fail to converge: A variation study of complementizer choice. MS, Stanford University.
Steger, Maria, and Edgar W. Schneider. 2012. Complexity as a function of iconicity: The case of complement clause constructions in New Englishes. In B. Kortmann and B. Szmrecsanyi (eds.), *Linguistic complexity*, 156-191. Berlin; Boston: De Gruyter.
Szmrecsanyi, Benedikt. 2005. Language users as creatures of habit: a corpus-linguistic analysis of persistence in spoken English. *Corpus Linguistics and Linguistic Theory* 1: 113-150.
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. 2013. *Grammatical variation in British English dialects: A study in corpus-based dialectometry*. Cambridge, [England]; New York: Cambridge University Press.
Szmrecsanyi, Benedikt, and Nuria Hernández. 2007. Manual of information to accompany the Freiburg Corpus of English Dialects sampler ("FRED-S"). Downloaded from: urn:nbn:de:bsz:25-opus-28598, URL: http://www.freidok.uni-freiburg.de/volltexte/2859/. Freiburg.
Sali A. Tagliamonte. 2007. Representing real language: Consistency, trade-offs and thinking ahead! In J. Beal, K. Corrigan and H. Moisl (eds). *Using unconventional digital language corpora*. Volume 1: *Synchronic corpora*, 205-240. Basingstoke: Palgrave Macmillan.
Tagliamonte, Sali, and Helen Lawrence. 2000. *I used to dance, but i don't dance now*: The habitual past in English. *Journal of English Linguistics* 28: 324-353.
Tagliamonte, Sali, and Jennifer Smith. 2002. *Either it isn't or it's not*: NEG/AUX contraction in British dialects. *English World Wide* 23): 251-281.
Tagliamonte, Sali, and Rosalind Temple. 2005. New perspectives on an ol' variable: (t,d) in British English. *Language Variation and Change* 17: 281-302.
Taylor, Ann. 2008. Contact effects of translation: Distinguishing two kinds of influence in Old English. *Language Variation and Change* 20: 341-365. doi:10.1017/S0954394508000100.

Thompson, Sandra A. and Anthony Mulac. 1991. Discourse uses of *that* in English. *Journal of Pragmatics* 15: 237-251. doi: 10.1515/LING.2009.001

Torres Cacoullos, Rena, and James A. Walker. 2009a. The present of the English future: grammatical variation and collocations in discourse. *Language* 85: 321-354. doi:10.1353/lan.0.0110.

Torres Cacoullos, Rena and James A. Walker. 2009b. On the persistence of grammar in discourse formulas: A variationist study of *that*. *Linguistics* 47: 1-43.

Travis, Catherine E. 2007. Genre effects on subject expression in Spanish: Priming in narrative and conversation. *Language Variation and Change* 19: 101-135. doi:10.1017/S0954394507070081.

Van Bogaert, Julie. 2010. A constructional taxonomy of *I think* and related expressions: accounting for the variability of complement-taking mental predicates. *English Language and Linguistics* 14: 399-427. doi:10.1017/S1360674310000134.

Weiner, Judith, and William Labov. 1983. Constraints on the agentless passive. *Journal of Linguistics* 19: 29-58.

Wolk, Christoph (in preparation). Integrating Aggregational and Probabilistic Approaches to Dialectometry and Language Variation. PhD dissertation, University of Freiburg.

Wolk, Christoph, Joan Bresnan, Anette Rosenbach, and Benedikt Szmrecsanyi. 2013. Dative and genitive variability in Late Modern English: Exploring cross-constructional variation and change. *Diachronica*.