Canadian Journal of Linguistics/Revue canadienne de linguistique, 62(4): 685–701, 2017
doi: 10.1017/cnj.2017.34

Variationist sociolinguistics and corpus-based variationist linguistics: overlap and cross-pollination potential

BENEDIKT SZMRECSANYI
KU Leuven
benedikt.szmrecsanyi@kuleuven.be

Abstract

The paper surveys overlap between corpus linguistics and variationist sociolinguistics. Corpus linguistics is customarily defined as a methodology that bases claims about language on usage patterns in collections of naturalistic, authentic speech or text. Because this is what is typically done in variationist sociolinguistics work, I argue that variationist sociolinguists are by definition corpus linguists, though of course the reverse is not true: the variationist method entails more than merely analyzing usage data, and not all corpus analysts are interested in variation. But that being said, a considerable and arguably increasing number of corpus linguists not formally trained in variationist sociolinguistics are explicitly concerned with variation and engage in what I call *corpus-based variationist linguistics* (CVL). I first discuss what unites or divides work in CVL and in variationist sociolinguistics. In a plea to cross subdisciplinary boundaries, I subsequently identify three research areas where variationist sociolinguists may draw inspiration from work in CVL: conducting multi-variable research, paying more attention to probabilistic grammars, and taking more seriously the register-sensitivity of variation patterns.

Keywords: variationist sociolinguistics, corpus linguistics, aggregation, register, probabilistic grammar

Résumé

Cet article explore le chevauchement entre la linguistique de corpus et la sociolinguistique variationniste. La linguistique de corpus est typiquement définie comme une méthodologie qui fonde ses affirmations linguistiques sur les régularités de l'usage émergeant des collectes de données orales ou textuelles naturalistes et authentiques. Puisque c'est ce qui se fait généralement en sociolinguistique variationniste, je soutiens que les sociolinguistes variationnistes sont par définition des linguistes de corpus, bien que l'inverse ne soit pas vrai: la méthode variationniste implique davantage que le seul fait d'analyser les données de

I am grateful to Jeroen Claes, Jason Grafmiller, Lars Hinrichs, Laurel MacKenzie, and two anonymous referees for helpful feedback on earlier versions of this paper. The usual disclaimers apply.

l'usage, et tous les analystes de corpus ne s'intéressent pas à la variation. Ceci étant dit, un nombre grandissant de linguistes de corpus n'ayant pas été formellement formés en sociolinguistique variationniste s'intéressent explicitement à la variation et s'investissent dans ce que j'appelle la *linguistique variationniste basée sur les corpus* (en anglais, *corpus-based variationist linguistics* ou CVL). Je discute d'abord ce qui unit et ce qui divise la linguistique variationniste basée sur les corpus et la sociolinguistique variationniste. Dans un appel visant à franchir les frontières des sous-disciplines, j'identifie ensuite trois domaines de recherche où les sociolinguistes variationnistes peuvent s'inspirer des travaux en linguistique variationniste basée sur les corpus: effectuer des recherches multi-variables, accorder plus d'attention aux grammaires probabilistes et prendre plus au sérieux la sensibilité au registre des modèles de variation.

Mots-clés: sociolinguistique variationiste, linguistique de corpus, agrégation, registre, grammaire probabiliste

1. Introduction

The goal of this paper is to encourage reflection on the methodologies and research questions that fuel work in corpus linguistics and variationist sociolinguistics.[1] That there is, as I will argue, considerable overlap between the two fields will be well-known to many (though maybe not all) readers. But then again, reminding practitioners in both fields of the common ground that they share, and of the cross-pollination potential that this common ground creates, is perhaps an exercise not entirely lacking in merit. Let us begin by reviewing definitions of what a corpus is. The community of corpus analysts that is of interest in the present paper defines corpus linguistics as a *methodology* that draws on collections of more or less naturalistic texts or speech for the sake of conducting some sort of linguistic analysis.

1. A corpus will be considered a collection of texts or parts of texts upon which some general linguistic analysis can be conducted (Meyer 2002: xi).
2. A corpus is a collection of non-elicited usage events […] a sample of spontaneous language use that is (generally) realized by native speakers (Tummers et al. 2005: 231).
3. A corpus can be defined as a body of naturally occurring language (McEnery et al. 2006: 4).

I would like to argue that to the extent that typical work in the variationist sociolinguistics, also known as the Language Variation and Change (henceforth: LVC) paradigm, is based on the analysis of fully transcribed sociolinguistic interviews or other naturalistic production data, LVC work is arguably based on corpora, albeit so-called *specialized* ones (McEnery et al. 2006: 5). And therefore the variationist method is a proper subset of the corpus-linguistic family of methods.

[1]Abbreviations used: ARCHER: A Representative Corpus of Historical English Registers; CVL : corpus-based variationist linguistics; CHILDES: Child Language Data Exchange System; FRED: Freiburg English Dialect Corpus; ICE: International Corpus of English; L2: second language; LVC: Language Variation and Change; MD: multi-dimensional.

https://doi.org/10.1017/cnj.2017.34 Published online by Cambridge University Press

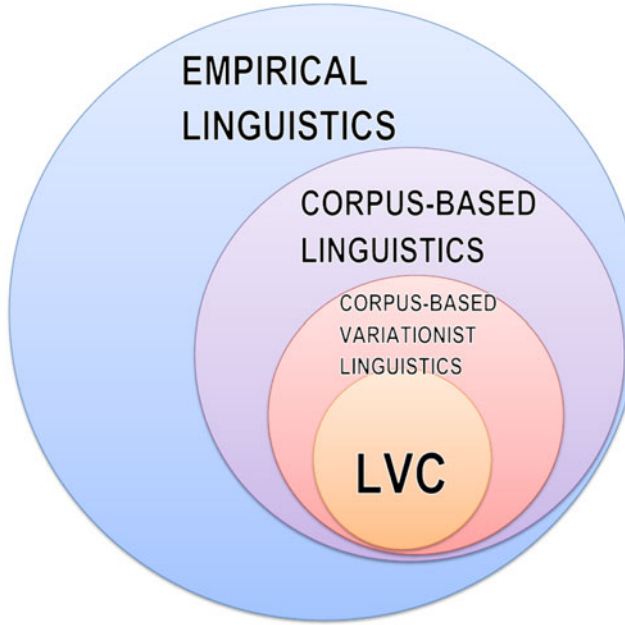

Figure 1: LVC versus other branches of empirical linguistics

Figure 1 explores the relationship between LVC and other branches of empirical linguistics from a corpus-linguistic perspective. While of course not all empirical linguists rely on the observational analysis of naturalistic corpus data (see, for example, Rosenbach 2005 for an experimental analysis of syntactic variation), corpora are a fairly popular data source. So corpus-based linguistics comes within the remit of empirical linguistics. But crucially, not all corpus-linguistic research is concerned with linguistic variation in the variationist sense – see, for example, Nesselhauf (2005) for corpus-driven, frequency-based collocation research which does not focus on variation per se. There are also corpus-linguistic approaches to sociolinguistics (e.g., discourse analysis) that are not necessarily variationist (e.g., Baker 2010, Friginal and Hardy 2014). Note further that there is plenty of corpus research that is concerned with variation in a broader sense (e.g., Biber 1988 and follow-up work) without using the variationist method in the narrower, LVC-inspired sense. This narrower sense would define corpus-based variationist linguistics (henceforth: CVL) as scholarship that meets the following three criteria:

1. CVL is concerned with "alternate ways of saying 'the same' thing" (Labov 1972: 188), so CVL researchers properly define variables and variants.
2. CVL research observes the Principle of Accountability (Labov 1969: 738), hence it is linguistic choice-making processes and not text frequencies that take centre stage.
3. CVL uses rigorous quantitative methodologies and statistical modeling (e.g., regression analysis) to explore the conditioning of linguistic variation.

https://doi.org/10.1017/cnj.2017.34 Published online by Cambridge University Press

This means that the remit of CVL as I define it here does *not* include research that is empirical but not corpus-based (consider, for example, experimental psycholinguistics in the spirit of Bock 1986), or research that is corpus-driven but not concerned with variation (consider, for example, multiword expression research à la Rayson et al. 2010), or research that is corpus-based and actually concerned with variation but that does not use the variationist method (consider, for example, seminal register-variation research in the tradition of Biber 1988). Recent CVL research that does fit the bill as defined above includes studies such as Gries (2005), Heylen (2005), Jaeger (2006), Bresnan et al. (2007), Grondelaers and Speelman (2007), Hinrichs and Szmrecsanyi (2007), Hilpert (2008), Lohmann (2011), De Cuypere and Verbeke (2013), Levshina et al. (2013), Schilk et al. (2013), Theijssen et al. (2013), Wolk et al. (2013), Claes (2014), Ehret et al. (2014), Grafmiller (2014), Pijpops and Van de Velde (2014), Wulff et al. (2014), and Shih et al. (2015). The authors listed here have mostly not received formal training in variationist sociolinguistics, but the studies cited all utilize the variationist method as defined above. In short, this work would probably qualify in principle to be published in the journal *Language Variation and Change*.

Work by LVC practitioners, needless to say, also meets the criteria as set forth above, and so Figure 1 takes the liberty – perhaps somewhat provocatively – of mapping LVC as a sub-field of CVL, the justification being that the LVC approach can be seen as a particular way of doing CVL. That being said, it is clear that work by LVC practitioners is set apart from other work in the CVL domain by a number of emphases and practices that we will discuss in more detail in the next section. One should also keep in mind that this exercise in subsetting is to some extent anachronistic, in that historically speaking work in CVL has drawn inspiration from work in LVC, and not the other way round. Moreover, I concede that an argument could be made that the LVC circle in Figure 1 should extend beyond the CVL area, to the extent that LVC practitioners use, for example, experimental methods. And finally I hasten to add that the size of the circles is not necessarily proportional to the relative importance of the fields.

The remainder of this contribution is structured as follows. Section 2 catalogues a number of differences between CVL and LVC. In Section 3, I identify research questions and methodologies with regard to which LVC work may draw inspiration from work that has been carried out in the CVL community. Section 4 offers some concluding remarks.

2. Some differences between LVC and CVL

In terms of research emphases and practices, and as regards theoretical commitments, there are six major differences that distinguish work in LVC from work in CVL:

1. Almost needless to say, much LVC work takes an interest in how demographic factors such as speaker age, gender, education, or regional provenance engender variation. To the extent that CVL work is interested in sociolinguistic factors at all, it is typically rather macro-sociological drifts such as "economization" (the tendency toward

https://doi.org/10.1017/cnj.2017.34 Published online by Cambridge University Press

informational compression) or "colloquialization" (use of informal language) that take centre stage (see, for example, Hinrichs and Szmrecsanyi 2007).

2. Since its inception, LVC as a field has taken a particular interest in phonetic variation (see Lavandera 1978 for discussion). The focus on phonetic variation is by no means exclusive (see, for example, Weiner and Labov 1983 on the passive/active alternation), but it is undeniable that phonetics is going strong in the LVC community. By contrast, phonetic variation plays a minor role in CVL work (but see, for example, Rosenfelder 2009), which instead tends to prioritize lexical (e.g., Heylen 2005) and, in particular, grammatical (e.g., Lohmann 2011) variation.

3. In the realm of language change, variationist sociolinguists have pioneered usage of the apparent-time construct (e.g., Bailey et al. 1991), which ingeniously allows the analyst to infer processes of change from synchronic datasets sampling multiple generations. Diachronically oriented CVL work, by contrast, prefers to track changes in real-time data (e.g., Szmrecsanyi et al. 2016a), drawing on increasingly massive historical corpora that cover, more often than not, a variety of written text types, such as A Representative Corpus of Historical English Registers (ARCHER) (Yáñez-Bouza 2011).

4. More culturally speaking, fieldwork plays a big role in LVC training and research, and it is fair to say that many practitioners like to maintain a close relationship to their data. In the corpus community, by contrast, many corpus analysts are more than happy to draw on existing, publicly available (but anonymous) corpora. We also note that LVC analysts do not shy away from meticulous manual data annotation, which is seen as positive because manual annotation helps the researcher to keep in touch with the dataset. CVL analysts, by contrast, tend to be more interested in (semi-)automatic retrieval and annotation procedures. This difference may also partly explain why CVL analysts tend to be rather more enthusiastic about big data projects – the community is now turning to multi-*billion* word corpora such as, for example, the Corpus of Global Web-based English (Davies and Fuchs 2015) – than LVC analysts are.

5. CVL analysts are often more adventurous than LVC analysts when it comes to experimenting with innovative analytic techniques. Corpus linguists have been the driving force beyond the development and application of more robust statistical techniques, which offer various practical and theoretical advantages compared to plain binary logistic regression. These innovative techniques include, for example, polytomous or multinomial regression (e.g., Han et al 2013), multimodel inference (e.g., Grafmiller and Shih 2011), as well as more psychologically plausible models such as memory-based learning (e.g., Theijssen et al. 2013) and naïve discriminative learning (e.g., Baayen et al. 2013).

6. Lastly, as regards theoretical commitment(s), corpus linguists tend to fall into two camps. On the one hand are those who eschew theoretical commitments and/or believe that doing corpus-driven linguistics is actually a theoretical commitment in its own right (e.g., Teubert 2005: 2). On the other hand are those who consider corpus linguistics a *methodology* to be utilized to address theory-inspired research questions. The vast majority of CVL analysts belong to the second category. The exact theoretical motivation for conducting CVL research varies substantially – the spectrum covers, for example, cognitive linguistics (e.g., Grondelaers and Speelman 2007), psycholinguistics (e.g., Gries 2005), dialectology/geolinguistics (Kolbe-Hanna and Szmrecsanyi 2015), and second language acquisition research (e.g., Wulff et al. 2014) – but the least common denominator is that CVL analysts tend to view variation analysis as an exercise in usage-based linguistics, in the sense that grammar is seen as "the cognitive organization of one's

https://doi.org/10.1017/cnj.2017.34 Published online by Cambridge University Press

experience with language" (Bybee 2006: 711). While this view is very mainstream in the CVL camp, the matter of usage- and experience-basedness appears to be a more controversial issue in the LVC community (see Guy 2014 for discussion).

3. Cross-pollination potential

CVL is a comparatively young research orientation whose disciplinary roots are mainly in European-style corpus linguistics as represented in the work of scholars such as Geoffrey Leech, Christian Mair, and (to some extent) John Sinclair. But then again, historically speaking a major source of *methodical* inspiration for many people now engaging in CVL has been North American variationist sociolinguistics in the tradition of William Labov. Against this backdrop, this section investigates the extent to which inspiration may flow in the other direction as well: which features of CVL-style variation analysis may be of interest to LVC practitioners? In what follows, I discuss three areas with cross-pollination potential of this type: the tendency in CVL research to study more than one variable at a time (Section 3.1.), the more explicit attention paid to the nature of probabilistic knowledge (Section 3.2.), and the way CVL scholarship takes an interest in how register differences shape variation patterns (Section 3.3.).

3.1 More variables at a time

Much work in the LVC literature takes, in the parlance of Nerbonne (2009: 176), a "single-feature-based" approach, in which the conditioning of individual (socio)linguistic variables is explored by looking at one variable at a time. This does not mean, of course, that multi-feature studies are completely absent from the LVC literature, or that their importance is not recognized: consider, for example, Rickford and McNair-Knox (1994), who study style-shifting based on a joint analysis of a comparatively large number of variables; Guy (2013), who investigates correlations among four sociolinguistic variables to show that individuals do not always use variables as coherently as one would think; or Meyerhoff (2017), who argues that conducting work at the intersection of language description and language variation necessitates looking at multiple variables ("symphonies of variation"). That being said, by and large it seems fair to say that the traditional way of doing things in variationist sociolinguistics is the single-feature way. By contrast, multi-feature research designs are comparatively more common, as I shall argue, in CVL circles, even though here as well much work is single-feature-based.

Multi-feature analysis has been an important theme in the corpus linguistics literature for a long while, going back at least to the Multi-Dimensional (MD) approach pioneered by Biber (1988). This particular approach is all about text frequencies and not at heart variationist, but the recent corpus literature showcases a number of ways to conduct variationist multi-variable research that may possibly inspire work in LVC. Corpus linguists have two motivations to investigate multiple variables: either to obtain a more reliable, stronger, or more multi-faceted linguistic signal through aggregation, and/or to investigate the extent to which variables are related or are interacting.

https://doi.org/10.1017/cnj.2017.34 Published online by Cambridge University Press

As to the first goal (increasing reliability), relevant work in the CVL tradition includes, for example, Szmrecsanyi (2013), a study that charts aggregate grammatical variation patterns in Great Britain's dialect landscape, based on data from the *Freiburg English Dialect Corpus* (FRED). The study is an exercise in "corpus-based dialectometry", dialectometry being the branch of geolinguistics concerned with measuring, visualizing and analyzing aggregate dialect similarities or distances as a function of properties of geographic space. In this spirit, Szmrecsanyi (2013) investigates joint variation patterns of 57 grammatical variants (e.g., *would* versus *used to* as markers of habitual past, non-standard past tense *done* vs. standard *did,* and so on). Among other things, the analysis shows that linguistic gravity scores (see Trudgill 1974), defined as two dialects' population product divided by the square of travel time between them, is a better predictor of aggregate grammatical similarity than mere geographic as-the-crow-flies distance. In a somewhat similar spirit, Grieve et al. (2011) aggregate over 40 high-frequency lexical alternation variables (e.g., *if* versus *whether*, *maybe* versus *perhaps*) in a 26-million-word corpus of letters to the editor covering written language usage in 206 cities from across the United States, and find that lexically speaking, there are five major dialect regions. With a similar interest in lexical variation, Ruette et al. (2016) conduct a joint analysis of no less than 303 lexical alternation variables (e.g., *holiday* versus *trip*, *cost* versus *expense*) in the Brown family of corpora of standard written-edited-published English, with the goal of determining the extent to which variation in the corpus material is systematically related to three lectal dimensions. In the order of their overall importance for explaining variation, these are discourse type/genre (informative versus imaginative), standard variety (British English versus American English), and real time period (1960s versus 1990s). Interestingly, in contrast to the other dimensions the study fails to identify distinctive lexical variables for the real-time dimension. CVL studies such as the ones sketched above share the view that single-feature study is inadequate when the objective is not to study particular features but to characterize multidimensional objects such as (dia)lects. Because the next feature down the road can and often does behave differently from a given feature (see Nerbonne 2009; Szmrecsanyi 2013, chapter 7 for extended discussions), multi-feature analysis is needed to reliably characterize (dia)lects.

As for the second motivation, examples of studies going multi-variable in order to investigate the extent to which variables are related or are interacting include Wolk et al. (2013), who investigate the English genitive, as in (1), and dative alternation, as in (2), in ARCHER, which covers written English in the Late Modern English period.

(1) a. The president's speech (the *s*-genitive)

b. The speech of the president (the *of*-genitive)

(2) a. The linguist sent the president a letter (the ditransitive dative variant)

b. The linguist sent a letter to the president (the prepositional dative variant)

Among other things, Wolk et al. aim to determine the extent to which the two alternations exhibit cross-constructional parallels. What they find is that the animacy constraint – animate possessors tend to favor the *s*-genitive, animate recipients tend to

https://doi.org/10.1017/cnj.2017.34 Published online by Cambridge University Press

favor the ditransitive dative variant – started to weaken in both alternations at around the same point in real time. This raises the possibility that this set of grammatical changes had not only language-internal but also language-external causes, for the study observes (Wolk et al. 2013: 413) that the decline of the animacy constraint is in some genres paralleled by an increasing usage frequency of expressions referring to inanimate and collective entities.

A second example for CVL research looking into intercorrelations between variables is Hinrichs et al. (2015), a study that investigates the relative importance of prescriptivism vis-à-vis other cultural developments that potentially constrain variation and change in written styles. Utilizing a variationist research design, the study is specifically concerned with relativizer choice in restrictive relative clauses with inanimate antecedents, as in (3):

(3) a. This is the house **that** Jack built.
b. This is the house **which** Jack built.
c. This is the house _____ Jack built.

In written/edited/published British and American English from the 1961–1992 period as sampled in the Brown family of corpora, alternation among relativizers is undergoing a massive shift from *which*, as in (3b), to *that*, as in (3a), and the study takes an interest in the factors driving this change. On the one hand, style guides – in particular American ones – prescribe relative *that* and proscribe relative *which* in restrictive relative clauses; on the other hand, *that* is the more informal option compared to bookish *which* (see e.g., Tagliamonte et al. 2005). To investigate whether prescriptivism or colloquialization is responsible for the shift, Hinrichs et al. (2015) annotated tokens for a variety of language-internal (e.g., relative clause length) and language-external (e.g., genre) factors. Crucially, their annotation also included information about additional variables and features regulated by prescriptivism, such as the proscribed use of stranded prepositions and the prescribed avoidance of the passive voice in the corpus texts under analysis – the assumption being that if the *that*-shift is a prescriptivism-driven change, then those writers who use *that* at the expense of *which* should also be those who avoid, for example, preposition stranding and the passive voice. Using regression analysis Hinrichs et al. show, among other things, that avoidance of *which* correlates with avoidance of the passive voice (a formal feature) and a preference for the active voice (which happens to also be the colloquial option), but not with other prescriptive rules advocating formal variants. This, Hinrichs et al. interpret as evidence that we are dealing with a colloquialization change that, unusually, is backed up by the infrastructure of prescriptivism.

In summary, the CVL research reviewed in this section would seem to suggest that variationist analysis can profit from multi-feature research designs. More specifically, I have argued that the study of joint variation patterns yields, on the one hand, a more robust and reliable characterization of multidimensional (dia)lects – or community grammars, for that matter. On the other hand, multi-feature analysis can help us better understand how different features and variables interact, and to state, in the words of Nerbonne (2009: 177), "general laws of linguistic variation".

https://doi.org/10.1017/cnj.2017.34 Published online by Cambridge University Press

3.2 More attention to probabilistic knowledge

As befits a community that is in the business of conducting variationist *socio*linguistics, LVC analysts are, as we have seen in Section 2, more often than not particularly interested in how language-external factors such as speaker age, gender, or education constrain variation patterns. This is not to say that LVC analysts are not also interested in the role that language-internal constraints play – on the contrary, constraint rankings and related measures play a key role in variationist/comparative sociolinguistics (Tagliamonte 2001).

But recent CVL work suggests that there may be even more (theoretical) mileage to be gained from taking a close look at language-internal constraints, and from considering both what these reveal about the nature of (implicit) linguistic knowledge, and how this knowledge is dynamic across space, time, and speakers' lifetimes. Take for example the variation-centered, usage- and experience-based probabilistic grammar framework developed by Joan Bresnan and collaborators (Bresnan 2007, Bresnan and Ford 2010). This research program marshals essentially variationist analysis methods to investigate syntactic variation phenomena such as the English dative alternation in naturalistic corpus data with optional experimental back-up (see below). The goal is to infer the extent and nature of grammatical knowledge from how language-internal constraints regulate language variation. Three assumptions inform this line of research: First, grammatical variation is sensitive to multiple and sometimes conflicting probabilistic constraints, which influence linguistic choice-making in subtle ways that may remain invisible unless analyzed quantitatively. Second, speakers have robust predictive capacities, and thus grammatical knowledge must have a probabilistic component. Third, this probabilistic component is derived in large part from language experience, and so is subtly, but fluidly (re)constructed throughout speakers' lives.

It would seem that this theoretical orientation is a better fit for the LVC enterprise – in which probabilistic modeling has, needless to say, played a huge role right from the outset (e.g., Labov 1969, Cedergren and Sankoff 1974) – than certain more formal, less stochastic and more categorical theoretical orientations. To highlight the potential of the framework for understanding variation, consider for example Bresnan and Hay (2008), who investigate two publicly accessible corpora (Switchboard and the Origins of New Zealand English corpus) with an interest in how well probabilistic grammars designed to predict American English dative choices between sentences like those in (2) generalize to the syntactic choices that speakers of New Zealand English make. The study finds that for speakers of New Zealand English, animacy of the recipient is a more powerful determinant of these choices than it is for speakers of American English: "non-animate recipients are more likely to be used in the double object construction in the NZ than in US spoken data" (Bresnan and Hay 2008: 252). Along similar lines, de Marneffe et al. (2012) investigate materials drawn from the Child Language Data Exchange System (CHILDES) to study whether children's dative choices are sensitive to the same constraints that regulate adults' dative choices, and report that there is "usage-based continuity between child and adult grammars" (de Marneffe et al.

https://doi.org/10.1017/cnj.2017.34 Published online by Cambridge University Press

2012: 26). A third example for this line of research is Szmrecsanyi et al. (2016b) , who investigate three grammatical alternations (the genitive alternation, as in (1) above, the dative alternation, as in (2) above, and the particle placement alternation, as in ex. (3), in the International Corpus of English (ICE):

(4) a. The linguist looked up the word. (verb-particle-object order)
b. The linguist looked the word up. (verb-object-particle order)

The goal of the study was to determine, among other things, the extent to which probabilistic constraints (and, by inference, the knowledge that language users have about these constraints) is stable as opposed to culturally malleable across native (e.g., British English) and postcolonial (e.g., Indian English) varieties of English. Analysis shows that there are subtle but significant "probabilistic indigenization" effects, defined as "the process whereby stochastic patterns of internal linguistic variation are reshaped by shifting usage frequencies in speakers of post-colonial varieties" (Szmrecsanyi et al. 2016b: 133). For example, in the particle placement alternation, increasing length of the direct object consistently favours the verb-particle-object order, but the effect is stronger in native varieties of English than in indigenized L2 varieties of English. The generalization is that the strength of such indigenization effects is proportional to the extent to which variables/alternations are tied to particular lexical anchors, such as particle verbs.

An interesting feature of research in this spirit is that corpus-based variation analysis can be optionally supplemented by experiments. Bresnan (2007: 76–84), for example, used a scalar rating task based on corpus materials (transcriptions of spoken dialogue passages) as stimuli to model subjects' responses regarding the naturalness of dative variants in contexts. These responses were compared to the predictions of the dative alternation regression model reported in Bresnan et al. (2007). The experiment showed that the likelihood of finding a particular linguistic variant in a particular context in a corpus tends to correspond to the intuitions that speakers have about the acceptability of the variants – hence, speakers' implicit knowledge about language must be to some extent probabilistic.

The research designs reviewed in this section all take an interest in how mutable the effect of language-internal constraints is, in the (implicit) knowledge of this mutability, and in how this sort of mutability can be interpreted as a function of geographic, social, and/or cultural factors. This focus may be attractive to LVC analysts thanks to its compatibility with the variationist method, and because it offers a "balanced diet" (Guy 2014: 59) consisting of both (abstract) constraints on variation plus a healthy dose of usage- and experience-based reasoning. Variationist datasets are rich data sources to further study the intersection between knowledge and variation, and LVC analysts may consider taking advantage of this.

3.3 More registers

It is clear that style and style-shifting have been of core interest to variationist sociolinguistics for as long as the field has existed, and interest in these concepts is not abating (see e.g., Bell 1984, Rickford and McNair-Knox 1994, and the articles in

https://doi.org/10.1017/cnj.2017.34 Published online by Cambridge University Press

Eckert and Rickford 2001). At the same time, though, LVC analysts tend to be especially interested in one particular register: vernacular speech as manifested in sociolinguistic interviews. Tagliamonte, for example, writes in her 2012 textbook that "variation in language is most readily observed in the vernacular of everyday life" (page 2; see also Chambers 2003: 6). To some extent, of course, this focus on vernacular speech is linked to the research community's interest in phonetic variation (see Section 2): there is no point in considering written genres if one is interested in pronunciation features. But beyond that, vernacular speech does have a special status in LVC theorizing. It is in this style where variation is thought to be at its best, that is, most systematic: it is "the style in which the minimum attention is given to the monitoring of speech" (Labov 1972: 208). What is more, there is a sense in the LVC community that variation grammars are stable in the face of style/genre/register-shifting, and that therefore variation patterns observed in vernacular speech can be generalized to other styles and registers (Guy 2005; see also Labov 2010: 265 and Rickford 2014: 596):

> For the most part, stylistic variation is quantitatively simple, involving raising or lowering the selection frequency of socially sensitive variables without altering other grammatical constraints on variant selection; indeed, it is commonly assumed in VR [Variable Rule] analyses that the grammar is unchanged in stylistic variation.. Guy 2005: 562

At least to outsiders it would seem that this is a theory-driven a priori assumption rather than an empirical truth (but see Travis and Lindstrom 2016). Corpus linguists in particular would feel that this is an issue begging for more data-driven investigation and an open mind. Corpus linguists, after all, tend to emphasize that register variation is rampant in human language (Ferguson 1983: 154), and so it is not surprising that register variation is an important topic – perhaps even the most important topic – in the corpus linguistics literature. And to many corpus linguists and CVL analysts it is just not immediately obvious that register-induced variation should be quantitatively simple, and that constraints on variation should be stable across registers.

Much previous corpus work on register differences has focused primarily on the text frequencies of particular linguistic features in particular registers: how often or rarely do we find particular linguistic features, such as passive constructions or causative verbs, in particular registers? The flagship method in this line of research is once again the Multi-Dimensional (MD) approach developed by Douglas Biber (1988), which measures co-occurrence patterns of linguistic features via frequency analysis. MD analysis has produced a number of invaluable results, but this line of research represents just one way of thinking about the difference that register makes. An alternative, variationist way of approaching register variation would ask not how frequently particular features are used in particular registers, but the following question: *when speakers can choose between different ways of saying the same thing, what is the extent to which they draw on different choice-making processes in different registers?*

This issue, while under-researched, should be dear to the hearts of both LVC and CVL practitioners. Indeed, preliminary corpus work suggests that linguistic choice-making is more sensitive to the situational context than many analysts of variation

https://doi.org/10.1017/cnj.2017.34 Published online by Cambridge University Press

would suspect. Thanks to the fact that many major corpora are multi-genre corpora (for example, the well-known British National Corpus samples dozens of spoken and written registers), corpus linguists have had no choice other than to take register variation seriously, and have been developing ever more sophisticated research designs to control for register-induced variation (see e.g., Gries 2015). And these sophisticated designs tend to show that choice-making processes may not be the same in different registers. Consider Grafmiller (2014), who studies a number of publicly accessible corpora such as the Switchboard corpus and the Boston University Noun Phrase Corpus. The target phenomenon is the English genitive alternation exemplified in (1) above, which is conditioned by numerous semantic, syntactic and phonological constraints. Grafmiller models the influence of these factors across six registers/genres (conversation, learned writing, non-fiction, general fiction, western diction, press) using a regression design. His analysis uncovers substantial interactions between external, stylistic constraints and the probabilistic weights that language-internal constraints on genitive variation have. In model 1 (p. 482), for example, no less than five (possessor animacy, possessor givenness, possessor/possessum length, type-token ratio, possessor text frequency) of the nine language-internal constraints under study turn out to have significantly different effect sizes across registers. In short, Grafmiller (2014) demonstrates that the probabilistic grammar of genitive choice is, in fact, massively sensitive to register effects.

Why should variationist sociolinguists care? Under the reasonable assumption that English genitive variation is not entirely atypical, massive register sensitivity à la Grafmiller (2014) raises important theoretical and methodological questions about the nature and scope of variation, and about the interaction of this variation with socioculture: how different do probabilistic effects have to be before we can speak of entirely different grammars – do we need different effect *directions*, or will significantly different effect *strengths* do? And if we do find that different registers come with different register grammars, is that another way of saying that we are essentially dealing with polyglossic or multilingual behavior, as suggested by Guy (2015)? In the mind of many corpus analysts, these are questions worth asking, but to do so we need more empirical/variationist work to explore the differences that register makes. This means that we need to go beyond sociolinguistic interviews and vernacular speech (see also Rickford and Wasow 1995: 128–129, D'Arcy and Tagliamonte 2015), to embrace, in a corpus-linguistic spirit, the full range of registers that we see when language is used.

4. Conclusion

I have argued in this article that the research questions and research methods in corpus-based variationist linguistics (CVL) are very close to those in the Language Variation and Change (LVC) community, and vice versa: variationist sociolinguists typically base claims about variation patterns on transcribed collections of naturalistic (vernacular) production data, and in so doing in fact utilize corpus-linguistic methods. On the other hand, CVL analysts adhere to the variationist method

https://doi.org/10.1017/cnj.2017.34 Published online by Cambridge University Press

(definition of variables/alternations, accountability, statistical modeling) when they analyze corpus data. The fact that there are, as we have seen, a number of emphases and practices that set LVC work apart from work in CVL does not negate the fact that there is substantial overlap.

So what? The upshot is that it would seem unreasonable to set up a rigid dichotomy between CVL and LVC. This is no surprise to the vast majority of CVL practitioners, who historically speaking have drawn much inspiration from work in variationist sociolinguistics, methodologically and otherwise. In the LVC community, there are likewise many people who understand the extent of the cross-pollination potential. Having said that, I believe that there is comparatively more potential in the LVC community to more fully embrace the opportunities that inter-subdisciplinarity would afford. It is against this backdrop that I have sketched in this paper three ways in which CVL work may help inspire new ways of analyzing variation in LVC. CVL practices with cross-pollination potential along these lines include first, multi-feature designs, to increase empirical reliability and to facilitate the study of intercorrelations between variables; second, methodologies to explore the intersection between knowledge and variation (how malleable are language-internal constraints on variation?); and third, a commitment to consider the full breadth of registers that we find in corpora of naturally occurring language, for the sake of understanding better the nature and scope of variation.

REFERENCES

Baayen, R. Harald, Peter Hendrix, and Michael Ramscar. 2013. Sidestepping the combinatorial explosion: An explanation of n-gram frequency effects based on naive discriminative learning. *Language and Speech* 56(3): 329–347. <doi:10.1177/0023830913484896>.

Bailey, Guy, Tom Wikle, Jan Tillery, and Lori Sand. 1991. The apparent time construct. *Language Variation and Change* 3(3): 241–264. <doi:10.1017/S0954394500000569>.

Baker, Paul. 2010. *Sociolinguistics and corpus linguistics*. Edinburgh: Edinburgh University Press.

Bell, Allan. 1984. Language style as audience design. *Language in Society* 13(2): 145. <doi: 10.1017/S004740450001037X>.

Biber, Douglas. 1988. *Variation across speech and writing*. Cambridge: Cambridge University Press.

Bock, Kathryn. 1986. Syntactic persistence in language production. *Cognitive Psychology* 18(3): 355–387.

Bresnan, Joan. 2007. Is syntactic knowledge probabilistic? Experiments with the English dative alternation. in *Roots: Linguistics in search of its evidential base*, ed. Sam Featherston and Wolfgang Sternefeld, 75–96. Berlin: Mouton de Gruyter.

Bresnan, Joan, Anna Cueni, Tatiana Nikitina, and Harald Baayen. 2007. Predicting the dative alternation. In *Cognitive foundations of interpretation*, ed. Gerlof Boume, Irene Krämer, and Joost Zwarts, 69–94. Amsterdam: Royal Netherlands Academy of Science.

Bresnan, Joan, and Marilyn Ford. 2010. Predicting syntax: Processing dative constructions in American and Australian varieties of English. *Language* 86(1): 168–213. <doi: 10.1353/lan.0.0189>.

Bresnan, Joan, and Jennifer Hay. 2008. Gradient grammar: An effect of animacy on the syntax of *give* in New Zealand and American English. *Lingua* 118(2): 245–259. <doi:10.1016/j.lingua.2007.02.007>.

Bybee, Joan. 2006. From usage to grammar: The mind's response to repetition. *Language* 82(4): 711–733.

Cedergren, Henrietta, and David Sankoff. 1974. Variable rules: Performance as a statistical reflection of competence. *Language* 50(2): 333–355.

Chambers, J.K. 2003. *Sociolinguistic theory: Linguistic variation and its social significance*. 2nd ed. Oxford: Blackwell.

Claes, Jeroen. 2014. A cognitive construction grammar approach to the pluralization of presentational *haber* in Puerto Rican Spanish. *Language Variation and Change* 26(2): 219–246. <doi10.1017/S0954394514000052>.

De Cuypere, Ludovic, and Saartje Verbeke. 2013. Dative alternation in Indian English: A corpus-based analysis. *World Englishes* 32(2): 169–184. <doi:10.1111/weng.12017>.

D'Arcy, Alexandra, and Sali A. Tagliamonte. 2015. Not always variable: Probing the vernacular grammar. *Language Variation and Change* 27(3): 255–285. <doi:10.1017/S0954394515000101>.

Davies, Mark, and Robert Fuchs. 2015. Expanding horizons in the study of World Englishes with the 1.9 billion word Global Web-Based English Corpus (GloWbE). *English World-Wide* 36(1): 1–28.

Eckert, Penelope, and John R. Rickford. 2001. *Style and sociolinguistic variation*. Cambridge: Cambridge University Press. <http://dx.doi.org/10.1017/CBO9780511613258>.

Ehret, Katharina, Christoph Wolk, and Benedikt Szmrecsanyi. 2014. Quirky quadratures: On rhythm and weight as constraints on genitive variation in an unconventional data set. *English Language and Linguistics* 18(2): 263–303. <doi:10.1017/S1360674314000033>.

Ferguson, Charles A. 1983. Sports announcer talk: Syntactic aspects of register variation. *Language in Society* 12(2): 153–172.

Friginal, Eric, and Jack A. Hardy. 2014. *Corpus-based sociolinguistics: A guide for students*. New York: Routledge, Taylor and Francis Group.

Grafmiller, Jason. 2014. Variation in English genitives across modality and genres. *English Language and Linguistics* 18(3): 471–496. <doi:10.1017/S1360674314000136>.

Grafmiller, Jason, and Stephanie Shih. 2011. New approaches to End Weight. Paper presented at Variation and Typology: New trends in syntactic research. Helsinki.

Gries, Stefan Th. 2005. Syntactic priming: A corpus-based approach. *Journal of Psycholinguistic Research* 34(4): 365–399.

Gries, Stefan Th. 2015. The most under-used statistical method in corpus linguistics: Multilevel (and mixed-effects) models. *Corpora* 10(1): 95–125. <doi:10.3355/cor.2015.0068>.

Grieve, Jack, Dirk Speelman, and Dirk Geeraerts. 2011. A statistical method for the identification and aggregation of regional linguistic variation. *Language Variation and Change* 23(2): 193–221. <doi:10.1017/S095439451100007X>.

Grondelaers, Stefan, and Dirk Speelman. 2007. A variationist account of constituent ordering in presentative sentences in Belgian Dutch. *Corpus Linguistics and Linguistic Theory* 3(2): 161–193. <doi:10.1515?CLLT.2007.010>.

Guy, Gregory R. 2005. Letters to *Language*. *Language* 81(3): 561–563. <doi:10.1353/lan.2005.0132>.

Guy, Gregory R. 2013. The cognitive coherence of sociolects: How do speakers handle multiple sociolinguistic variables? *Journal of Pragmatics* 52(1): 63–71. <doi:10.1016/j.pragma.2012.12.019>.

Guy, Gregory R. 2014. Linking usage and grammar: Generative phonology, exemplar theory, and variable rules. *Lingua* 142: 57–65. <doi:10.1016/j.lingua.2012.07.007>.

Guy, Gregory R. 2015. Coherence, constraints and quantities. Paper presented at NWAV44, Toronto.

Han, Weifeng, Antti Arppe, and John Newman. 2013. Topic marking in a Shanghainese corpus: From observation to prediction. *Corpus Linguistics and Linguistic Theory*. Published online June 5, 2013-06-05. <doi:10.1515/cllt-2013-0014>.

Heylen, Kris. 2005. A quantitative corpus study of German word order variation. In *Linguistic evidence: Empirical, theoretical and computational perspectives*, ed. Stephan Kepser and Marga Reis, 241–264. Berlin: Mouton de Gruyter.

Hilpert, Martin. 2008. The English comparative: Language structure and language use. *English Language and Linguistics* 12(3): 395–417. <doi:10.1017/S1360674308002694>.

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(3): 437–474. <doi:10.1017/S1360674307002341>.

Hinrichs, Lars, Benedikt Szmrecsanyi, and Axel Bohmann. 2015. Which-hunting and the Standard English relative clause. *Language* 91(4): 806–836. <doi:10.1353/lan.2015.0062>.

Jaeger, T. Florian. 2006. Redundancy and syntactic reduction in spontaneous speech. Doctoral dissertation, Stanford University.

Kolbe-Hanna, Daniela, and Benedikt Szmrecsanyi. 2015. Grammatical Variation. In *The Cambridge handbook of English corpus linguistics*, ed. Douglas Biber and Randi Reppen, 161–179. Cambridge: Cambridge University Press.

Labov, William. 1969. Contraction, deletion, and inherent variability of the English copula. *Language* 45(4): 715–762. <doi:10.2307/412333>.

Labov, William. 1972. *Sociolinguistic patterns*. Philadelphia: University of Philadelphia Press.

Labov, William. 2010. *Principles of linguistic change, vol. 3: Cognitive and cultural factors*. Malden, MA: Wiley-Blackwell.

Lavandera, Beatriz R. 1978. Where does the sociolinguistic variable stop? *Language in Society* 7(2): 171–182. <doi: 10.2307/4166996>.

Levshina, Natalia, Dirk Geeraerts, and Dirk Speelman. 2013. Towards a 3D-grammar: Interaction of linguistic and extralinguistic factors in the use of Dutch causative constructions. *Journal of Pragmatics* 52: 34–48. <doi:10.1016/j.pragma.2012.12.013>.

Lohmann, Arne. 2011. Help vs. help to: A multifactorial, mixed-effects account of infinitive marker omission. *English Language and Linguistics* 15(3); 499–521. <doi:10.1017/S1360674311000141>.

de Marneffe, Marie-Catherine, Scott Grimm, Inbal Arnon, Susannah Kirby, and Joan Bresnan. 2012. A statistical model of the grammatical choices in child production of dative sentences. *Language and Cognitive Processes* 27(1): 25–61. <doi:10.1080/01690965.2010.542651>.

McEnery, Tony, Richard Xiao, and Yukio Tono. 2006. *Corpus-based language studies: An advanced resource book*. New York: Routledge.

Meyer, Charles F. 2002. *English corpus linguistics: An introduction*. Cambridge: Cambridge University Press.

Meyerhoff, Miriam. 2017. Writing a linguistic symphony: Analysing variation while doing language documentation. *Canadian Journal of Linguistics* 62(4). <10.1017/cnj.2017.28>

Nerbonne, John. 2009. Data-driven dialectology. *Language and Linguistics Compass* 3(1): 175–198.

Nesselhauf, Nadja. 2005. *Collocations in a learner corpus*. Amsterdam: John Benjamins. <http://www.jbe-platform.com/content/books/9789027294739>.

Pijpops, Dirk, and Freek Van de Velde. 2014. A multivariate analysis of the partitive genitive in Dutch: Bringing quantitative data into a theoretical discussion. *Corpus Linguistics and Linguistic Theory*. Published online April 10, 2014. <doi:10.1515/cllt-2013-0027>.

Rayson, Paul, Scott Piao, Serge Sharoff, Stefan Evert, and Begoña Villada Moirón. 2010. Multiword expressions: Hard going or plain sailing? *Language Resources and Evaluation* 44(1–2): 1–5. <doi:10.1007/s10579-009-9105-0>.

Rickford, John R. 2014. Situation: Stylistic variation in sociolinguistic corpora and theory. *Language and Linguistics Compass* 8(11): 590–603.

Rickford, John R., and Faye McNair-Knox. 1994. Addressee- and topic-influenced style shift: A quantitative sociolinguistic study. In *Perspectives on register: Situating register variation within sociolinguistics*, ed. Douglas Biber and Edward Finegan, 235–276. Oxford: Oxford University Press.

Rickford, John R., and Thomas A. Wasow. 1995. Syntactic variation and change in progress: Loss of the verbal coda in topic-restricting *as far as* constructions. *Language* 71(1): 102–131. <doi:10.2307/415964>.

Rosenbach, Anette. 2005. Animacy versus weight as determinants of grammatical variation in English. *Language* 81(3): 613–644.

Rosenfelder, Ingrid. 2009. Sociophonetic variation in educated Jamaican English: An analysis of the spoken component of ICE-Jamaica. Doctoral dissertation, University of Freiburg.

Ruette, Tom, Katharina Ehret, and Benedikt Szmrecsanyi. 2016. A lectometric analysis of aggregated lexical variation in written Standard English with semantic vector space models. *International Journal of Corpus Linguistics* 21(1): 48–79. <doi:10.1075/ijcl.21.1.03rue>.

Schilk, Marco, Joybrato Mukherjee, Christopher Nam, and Sach Mukherjee. 2013. Complementation of ditransitive verbs in South Asian Englishes: A multifactorial analysis. *Corpus Linguistics and Linguistic Theory* 9(2): 187–225. <doi:10.1515/cllt-2013-0001>.

Shih, Stephanie, Jason Grafmiller, Richard Futrell, and Joan Bresnan. 2015. Rhythm's role in genitive construction choice in spoken English. in *Rhythm in cognition and grammar*, ed. Ralf Vogel and Ruben Vijver, 207–234. Berlin: De Gruyter. <http://www.degruyter.com/view/books/9783110378092/9783110378092.207/9783110378092.207.xml>.

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

Szmrecsanyi, Benedikt, Douglas Biber, Jesse Egbert, and Karlien Franco. 2016a. Toward more accountability: Modeling ternary genitive variation in Late Modern English. *Language Variation and Change* 28(1): 1–29. <doi:10.1017/S0954394515000198>.

Szmrecsanyi, Benedikt, Jason Grafmiller, Benedikt Heller, and Melanie Röthlisberger. 2016b. Around the world in three alternations: Modeling syntactic variation in varieties of English. *English World-Wide* 37(2): 109–137.

Tagliamonte, Sali. 2001. Comparative sociolinguistics. In *Handbook of language variation and change*, ed. J. K. Chambers, Peter Trudgill, and Natalie Schilling-Estes, 729–763. Oxford: Blackwell.

Tagliamonte, Sali. 2012. *Variationist sociolinguistics: Change, observation, interpretation*. Malden, MA: Wiley-Blackwell. <http://public.eblib.com/EBLPublic/PublicView.do?ptiID=819316>.

Tagliamonte, Sali, Jennifer Smith, and Helen Lawrence. 2005. No taming the vernacular! Insights from the relatives in Northern Britain. *Language Variation and Change* 17(1): 75–112.

Teubert, Wolfgang. 2005. My version of corpus linguistics. *International Journal of Corpus Linguistics* 10(1): 1–13. <doi:10/1075/ijcl.10.1.01.teu>.

Theijssen, Daphne, Louis ten Bosch, Lou Boves, Bert Cranen, and Hans van Halteren. 2013. Choosing alternatives: Using Bayesian networks and memory-based learning to study the dative alternation. *Corpus Linguistics and Linguistic Theory* 9(2): 227–262. <doi: 10.1515/cllt-2013-0007>.

Travis, Catherine E., and Amy M. Lindstrom. 2016. Different registers, different grammars? Subject expression in English conversation and narrative. *Language Variation and Change* 28(1): 103–128. <doi:10.1017/S0954394515000174>.

Trudgill, Peter. 1974. Linguistic change and diffusion: Description and explanation in sociolinguistic dialect geography. *Language in Society* 3(2): 215–246.

Tummers, Jose, Kris Heylen, and Dirk Geeraerts. 2005. Usage-based approaches in cognitive linguistics: A technical state of the art. *Corpus Linguistics and Linguistic Theory* 1(2): 225–261. <doi:10.1515/cllt.2005.1.2.225>.

Weiner, Judith, and William Labov. 1983. Constraints on the agentless passive. *Journal of Linguistics* 19(1): 29–58.

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* 30(3): 382–419. <doi:10.1075/dia.30.3.04wol>.

Wulff, Stefanie, Nicholas Lester, and Maria T. Martinez-Garcia. 2014. *That*-variation in German and Spanish L2 English. *Language and Cognition* 6(2): 271–299. <doi: 10.1017/langcog.2014.5>.

Yañez-Bouza, Nuria. 2011. ARCHER past and present (1990–2010). *ICAME Journal* 35: 205–236.