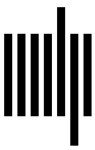
The Open Handbook of Linguistic Data Management

Edited by: Andrea L. Berez-Kroeker, Bradley McDonnell, Eve Koller, Lauren B. Collister

Citation:

The Open Handbook of Linguistic Data Management
Edited by: Andrea L. Berez-Kroeker, Bradley McDonnell, Eve Koller, Lauren B. Collister
DOI: 10.7551/mitpress/12200.001.0001
ISBN (electronic): 9780262366076
Publisher: The MIT Press
Published: 2022



The MIT Press

16 Data Management at the uOttawa Sociolinguistics Laboratory

Shana Poplack

1 Introduction

This chapter details the data management principles and practices used in my Sociolinguistics Lab (http://www.sociolinguistics.uottawa.ca/thelab.html), home to hundreds of hours and millions of words of recorded spontaneous speech. The lab is the repository of nineteen major spoken language corpora in a variety of languages and eight corpora of language mixing in typologically similar and distinct language pairs, all constructed by our team. These include large-scale computerized data sets on spoken Canadian French spanning an apparent time period of a century and a half, Quebec English spoken before and after the passage of the Charter of the French Language (1977) that made it a minority language, and diaspora varieties of African American English, among others, as well as the Sociolinguistic Archives, collected by students in Urban Dialectology field methods courses between 1982 and 2018. It also houses two major written corpora of speech surrogates (*Ottawa Repository of Early African American Correspondence*, Van Herk & Poplack 2003; *Recueil historique des grammaires du français*, Poplack et al. 2015).

Much of this work was initiated in the early 1980s, well before data management became a popular topic in circles outside of variationist sociolinguistics, so some of the methods described here will appear (and in fact are) utterly antiquated by today's standards. Nonetheless, forty years down the road, there seems to be no danger of these materials "expiring" or getting "used up" any time soon, despite thousands of uses in the form of articles, books, talks, workshops, theses, dissertations, class papers, and other tangible products, generated by me, my students, lab assistants and associates, as well as colleagues far and wide. This is because all of these corpora have been properly preserved, if only by the seat of their pants, and remain discoverable, accessible (Bird & Simons 2003), and to the extent permitted by ethical considerations, shareable. Ensuing sections describe how we approached the ever-present tensions between the ideal and the feasible to achieve this.

2 Corpus creation

Recognizing that the linguistic materials at one's disposal exert the strongest constraint on what can ultimately be investigated, our data management practices begin with data acquisition. This inevitably raises the question of what to collect and from whom. The corpora housed in the Sociolinguistics Lab are conceived first and foremost as archives of potential answers to specific research questions. In keeping with the sociolinguist's mandate to study language issues of particular import to society, these typically emerge from public stakeholder discourse, but with a special focus on those of theoretical linguistic interest.

A notable example is the *Ottawa-Hull French Corpus* (Poplack 1989), the cornerstone of a still ongoing project that seeks to elucidate the extent to which minority status affects a language's permeability to influence from a majority language. The nature, extent, and even existence of contact-induced change have been the subject of much controversy in linguistics. It is also a long-standing concern of Canadian francophones, who fear that intense contact with and influence from English is destroying the structural integrity of French. Such concerns, which until recently, enjoyed relatively little empirical validation, dictate our methodology for corpus construction. The National Capital Region of Canada, situated on both sides of the geographic, provincial, and linguistic border between Quebec (French majority language) and Ontario (French minority language) was selected as the study site. Twenty-four native francophones, stratified according to age and gender were

randomly sampled from each of five "French" neighborhoods, each differing according to intensity of contact with English at the local level. Another corpus (*Le français en contexte: Milieux scolaire et social*; Poplack 2015; Poplack & Bourdages 2005), originally built to test the popular belief that francophone youth are not acquiring "standard" French because their teachers do not master it themselves, is made up of 166 francophone high school students and their French language-arts teachers. Because these data were collected on the Quebec side of the National Capital Region twenty-five years later, this corpus contributes a real-time dimension. Together with a third corpus constructed from recordings made by folklorists in the 1940s and 1950s of rural Quebeckers born between 1846 and 1895 (*Récits du français québecois d'autrefois*; Poplack & St-Amand 2007), it extends the apparent time frame for the study of linguistic change to a century and a half, a span virtually unprecedented in the study of *speech*. The *Récits* also function as a precontact benchmark, crucial for any study of contact-induced change (Poplack & Levey 2010).

This comparative approach is also at the root of the *Early Black English* corpora (Poplack & Sankoff 1987; Poplack & Tagliamonte 1991). Inspired by the longstanding debate over the origins of African American Vernacular English (AAVE) as a prior creole or a dialect of English, these data address concerns of native speakers over the "quality" of their language, while responding to the caveat that any question of origins requires reference to an earlier stage of the language. The *Early Black English* corpora capitalize on speech recorded synchronically in three African American diaspora isolates settled between 1783 and 1824, with the goal of triangulating analysis of diagnostic linguistic features among them and relevant benchmarks to reconstruct the ancestor of contemporary AAVE (Poplack 2000; Poplack & Tagliamonte 2001).

For a full list of the lab's holdings, refer to Research Holdings (http://www.sociolinguistics.uottawa.ca/holdings.html) and the references therein. For the present purposes, suffice it to say that these are "unconventional" corpora (Beal, Corrigan, & Moisl 2007; Poplack 2007), insofar as they were specifically built to instantiate the conditions required to address a specific research question, rather than simply to represent some geographically or linguistically defined community, as is more often the case in (socio) linguistics. Thanks to the collection methods detailed herein, all of these data sets lend themselves readily to the study, both synchronic and diachronic, of just about any linguistic feature that occurs freely in speech and can be apprehended from an audio recording. In addition, by virtue of the criteria motivating their constitution, they offer the less common advantage that the behavior of each feature can also be interpreted in terms of a more general research question, an inestimable added benefit of corpora built on such principles.

3 Data collection

As sociolinguists, our primary interest lies in the study of spontaneous language use, and in particular, the linguistic variability that is its hallmark. This variability typically involves alternation between ratified forms and their non-standard, often overtly stigmatized, counterparts. The latter are emblematic of the *vernacular*, valued as the most regular and systematic form of language (Labov [1966] 2006). Eliciting the vernacular spontaneously in the course of the formal data collection process is no simple matter. It requires, first and foremost, creating a situation in which its use is deemed appropriate by the speaker. While this may be straightforward enough in small-scale social network studies, which often involve long-term participant observation of and attendant familiarity with members, it turns out to be quite a challenge when the researcher is also seeking to constitute a representative and numerically adequate speaker sample. The major corpora housed in the Sociolinguistics Lab, for example, are made up of 120–183 individuals, who are in most cases unrelated. What such large-scale surveys offer in breadth and range they often lose in depth. The typical result is speech data that rarely transcend the more formal poles of the stylistic continuum, where the linguistic features of interest are either rare or altogether absent. This constitutes a real liability for the study of the vernacular.

Our response to this problem has been to adopt the ethnographically inspired methods developed by Labov and his associates (ourselves included) to resolve the *observer's paradox* (Labov 1972) by encouraging conversation that is a better approximation to the vernacular than that typically afforded by face-to-face interviews. The major tool here is the methodological instrument known as the *sociolinguistic interview* (Labov 1984). The antithesis of a standard interview schedule, this is rather a guide to eliciting casual speech through introduction of a wide range of conversational topics of a largely informal nature. These are adapted to the interests and

concerns of the local populations by trained interviewers who, to the extent possible, are both linguists and community members themselves. To minimize the effect of the interview situation, including the power differential between interviewer and interviewee, the latter is encouraged to take the lead in the interaction, initiating transitions from topic to topic, as well as inclusion or exclusion of specific topics, all with limited intervention from the interviewer. The only exceptions involve the collection of metadata for each participant (section 4), which is introduced near the end of each recording session.

The results of these efforts, described in detail elsewhere (e.g., Poplack 1989; Poplack, Walker, & Malcolmson 2006), are a wealth of spontaneous speech, ranging in length from one to five hours per participant and featuring many narratives of personal experience, small-group discussions, and other highly informal discourse modes. These include the crucial vernacular variants targeted for study, in addition to (more easily accessible) careful speech styles. To the extent possible, we rely on linguists who are also community members to act as fieldworkers. They receive dedicated training, both in the lab and in the context of formal coursework, on how to administer the sociolinguistic interview, which is greatly enriched by their own insider knowledge of community mores. Perhaps our most successful endeavor of this kind is evidenced in the *African Nova Scotian English* corpus (Poplack & Tagliamonte 1991), a substantial compilation of largely informal speech collected by in-group members of tight-knit communities diglossic in Standard Canadian English and AAVE. Were it not for the local status—and skill!—of these fieldworkers, most of the vernacular grammatical features whose behavior we have since studied in detail would simply never have appeared during the recording sessions at all. This is because, where linguistic insecurity and at least some degree of diglossia coexist, "offending" forms are typically avoided with non-community members. This highlights the imperative to create appropriate conditions for data collection. Other notable examples of such prized but elusive linguistic features include multiword code-switching, or the French-Canadian informal register disparagingly referred to as *joual*. As observed in section 10, success at obtaining exemplars of the vernacular has often resulted in content too personal to disseminate publically, raising serious ethical concerns with respect to data sharing.

4 Metadata

Nowhere is the dictate that data cannot be divorced from their source(s) (Good, chapter 3, this volume; see also Holton, Leonard, & Pulsifer, chapter 4, this volume) more relevant than in sociolinguistic research. Recognizing that the speaker is a key source of the variability that characterizes language as well as the key agent of linguistic change, relevant speaker characteristics are explicitly incorporated into many of our analyses of linguistic structure. This section describes the metadata developed at the Sociolinguistics Lab to facilitate this endeavor. First, after each recording session, the fieldworker completes an *interview report* containing detailed demographic, sociological, and some language-related information about the participant(s) and their immediate families. At this time they also anonymize the interaction, providing each participant with a temporary pseudonym and speaker number, taking care to co-index these with every recording, interview report form, and any other documentation pertinent to the individual in question.[1] These are then transferred onto corpus-specific master metadata lists, which are continually updated as data acquisition continues, and even into preliminary analysis. Once data collection, transcription, and corpus construction have concluded, each participant is assigned a permanent unique number and pseudonym, consistently linking them to all their associated data. The result is a detailed inventory of potentially relevant characteristics. What is relevant can be expected to differ from variable to variable and community to community: one variable may be primarily socially conditioned, another may serve as a stylistic marker, and a third may display a distinct sociolinguistic profile depending on the speech community. Standard sociodemographic traits such as age, gender, neighborhood of residence, level of education, or approximate socioeconomic status are always taken into account, but we also capitalize on our knowledge of the community to incorporate local concerns into the analysis wherever feasible. Thus, speaker attitudes toward a majority language may prove explanatory in the analysis of variability in minority-language contexts (Poplack, Walker, & Malcolmson 2006); reported proficiency in a second language, or propensity to code-switch or to engage in lexical borrowing may be significant predictors in bilingual communities (Poplack 1989, 2018). These and other community-specific and more general factors are

operationalized as factors via codes that can be readily incorporated into multivariate analyses to assess their relative contribution to the linguistic choices speakers make.

The resulting master catalogs are also useful in the creation of sub-corpora for specific studies. To name but one example, the question of whether copious code-switching leads to structural convergence could be addressed by comparing the behavior of a linguistic candidate for contact-induced change in a corpus made up of copious and sparse code-switchers identified from such metadata (Poplack, Zentz, & Dion 2012). Metadata also facilitate the construction of representative subsamples, which are particularly useful for pilot or partial studies. The smallest such subsamples are made up of equal numbers of participants with maximally contrasting social characteristics (e.g., oldest vs. youngest, high-contact vs. low-contact neighborhood of residence) and are gradually expanded to include intermediate categories if exploratory studies warrant it. From participants with the targeted profiles, those whose recordings are most "successful" (in terms of length, stylistic range, vernacular features, and such) are then culled. The resulting subsamples obviate the need for a major investment in time and resources at the initial stage, while at the same time correcting for any (naturally occurring) imbalances in the larger sample. For example, while younger speakers tend to be more bilingual than their older counterparts in the wider *Quebec English Corpus* (Poplack, Walker, & Malcolmson 2006), the sub-sample maintains a perfect balance between those with low and high levels of bilingualism for each age cohort. Analysis of different variables in these pre-established sub-corpora enhances the generalizability of results.

A final type of metadata is afforded by a "corpus article" detailing the rationale behind the overall project, along with relevant methodological details regarding community type, data collection procedures, sample selection criteria, participant characteristics, transcription protocol, type of data, and such (e.g. Poplack 1989; Poplack, Walker, & Malcolmson 2006; Poplack & St-Amand 2007). Because data and findings can only be fully apprehended in the context of the norms of the community within which they were collected, prospective users are expected to familiarize themselves with these publications prior to receiving access to the corpora.

5 Annotation

Before describing the transcription protocols applied to the lab's corpora, recall that much of the data were collected decades before the advent of the wide variety of annotation schemes available today (e.g., Han, chapter 6, this volume; Beal, Corrigan, & Moisl 2007). Viewed in comparison with them, the solution we have adopted—manual transcription using standard orthography!—is not only technologically archaic, but to all appearances ridiculously vanilla. Yet even as increasingly sophisticated methods have gained more traction over the years, we have largely stuck with our original protocols for subsequent corpora, experience having shown us that they served our needs well. Our understanding of annotation as a *gateway* to analysis, as opposed to an end in and of itself, has dictated these choices.

For starters, once we experienced firsthand the formidable investment of time and money required to build a principled spoken-language corpus from scratch, we resolved to partition our own limited resources such that data *analysis* (i.e., actual linguistic research) would take precedence over data management, yet without unduly sacrificing the latter. Like many other corpus builders, we were torn between the lure of representing every linguistically interesting token and the desire to begin analysis of the large quantities of data that were flooding in. We were also sensitive to the hard lessons learned by colleagues whose ultra-detailed multilevel annotation systems eventually resulted in only a handful of fully transcribed recordings, having thus robbed precious time from analysis to support data handling for corpora that never materialized beyond a few speakers. This motivated our decision to maximize data entry (by getting initial transcriptions as quickly as possible) and multiply correction phases (totaling between three and six, both manual and automated, depending on the corpus). This meant giving up on many bells and whistles, despite hundreds of hours of labor from teams of seasoned transcribers.

An annotation scheme is only as good as the purposes it serves. We noted earlier that the corpora we construct lend themselves to the study of a wide variety of linguistic phenomena. Some may be known to us at the outset, but most emerge as the various projects evolve. Having no way of predicting what these might be, we (reluctantly) conceded that it would be unreasonable, if not

impossible, to attempt to account for all of them during the transcription phase. As an example, because the research we carry out is largely based on detailed analyses of morphosyntactic variation, we resolved to ignore the myriad phonetic variants with which the data are rife, reasoning that any phonetician who wanted to exploit the lab's corpora from this perspective would prefer to analyze the raw recordings personally. The transcription protocol we adopted was largely determined by the goal of the automatic data manipulation phase of the project: construction of a computerized concordance in which the entire database is maximally accessible. To achieve this, the data must be rendered as faithfully and as consistently as possible. This is a particularly daunting challenge for spontaneous speech, which may include numerous and varied non-standard forms—sometimes, as in our case, in more than one language. Our objective was to preserve all the pertinent variation without needlessly multiplying entries, which has the effect of reducing retrieval.

The intended uses of the data, coupled with the determination to facilitate its automated treatment, led us to adopt an orthographic solution rather than a fine phonetic or prosodic transcription. This is described in detail in Poplack (1989), but in addressing the problem of how to treat each of the variant forms that a single word may have, the overall strategy has been to represent variation resulting from the operation of phonetic or phonological processes in standard orthography, regardless of the actual realization of the form. Thus, the main verb in (1), which was phonetically realized as [gɛɾɳ], was transcribed <getting>, while variant forms affecting an entire morpheme, as in (2) (where the final [s] of 'trunks' represents the plural marker), were represented exactly as produced.

(1) And I said, "If things don't change around here, I'm getting out of here." (QEC.037.630)[2]

(2) That man had two trunks. Two trunk full of gold and silver and everything. Two trunk, big trunks. Full of gold and silver. (ANSE.NP.030.1323)

Our transcription protocol is generally consistent with accepted orthographic conventions in the language(s) in question, except where these violate our accessibility criteria. Thus, instead of using standard orthography for the recurrent Quebec French adverbial locution *à cette heure* 'now', containing the highly frequent (but in this context non-productive) function words *à* 'to' ($n = 32{,}521$) and *cette* 'this' ($n = 926$), we adopted the widespread dialect orthography *astheure* ($n = 1{,}222$) so as to enhance retrievability. In a small number of cases, we created idiosyncratic orthographies, mainly to distinguish certain high-frequency forms from already high-frequency homographs. Thus, English quotative and discourse *like* ($n = 37{,}449$), as in (3), was transcribed <lyke> to distinguish it from verbal and comparative homographs, rendered <like> ($n = 9{,}723$).

(3) And all of a sudden, I'm lyke, "Why am I feeling grouchy?" (QEC.308.2210)

Another important departure involved inserting spaces after apostrophes replacing elided vowels in French articles to detach them from the lexical item they were qualifying (e.g., <j' ai>, <l' amie>), or to separate productive English contracted forms (e.g., <n't> from <do n't>, <did n't>), as illustrated in (4) with excerpts from the *Quebec English Corpus*. This facilitates their location in the alphabetical concordance (see sentence display 4).

(4)

301	45	that way? [301] He did	n't	want to and the school
301	60	I was lyke, okay I do	n't	care. (laughter) Do n't
304	570	I think back then it was	n't	at all. [2] Do you have a
307	776	I think they just should	n't	be listening to music that
314	501	It's the second last, is	n't	it? [Oh it is, driving, you
315	1375	for, because he would	n't	stand there, 'cause it was
316	231	ahold of that and it does	n't	cost us anything, so we
318	445	remember that, I have	n't	thought of that for years. I
319	162	[1] Oh yeah? [319] I ca	n't	tell you the stories about.
319	671	never allowed to use "ai	n't	", so I mean, now
320	608	close. Now they were	n't	people that uhm, were on

As a rule, however, no other effort was made to modify the form of the material in any way. Syntax, lexical choice, deletions, insertions, and neologisms of all sorts, as well as code-switching and borrowing, were all scrupulously reproduced.

An as-yet-unresolved problem concerns the many forms that are regularly deleted in speech, affecting a wide variety of linguistic categories. Over the years spent transcribing and correcting the various lab corpora, successive teams have grappled repeatedly with the issue of how to render these deleted items so as to facilitate their eventual retrieval. But the daunting number of disparate forms that would have to be coded as null, coupled with the difficulty of finding a unique representation for each (one capable of distinguishing a null subject from a null complementizer or a null inflection, for instance), eventually led to a point of diminishing returns, leading us to revert to manual retrieval of such items. Another notable challenge involved language tagging. Because several of our corpora are bilingual, constructed with the express purpose of facilitating the linguistic analysis of language mixing phenomena such as code-switching and lexical borrowing, we wanted to flag other-language incorporations. This turned out to be relatively straightforward for multiword stretches, as in (5), where each change of language is introduced by a code (here, <A> for *anglais* 'English' and <F> for *français* 'French').

(5) Tu sais, <A> I helped them <F> à comprendre le français pis à le parler. (OH.014.569)
'You know, I helped them to understand French and to speak it.'

(6) On a faite nettoyer les sewers ici dans le projet parce c'était toute bloqué. (OH.027.2265)
'We got the sewers here in the projects cleaned because it was all clogged up.'

But, unsurprisingly in retrospect, our francophone transcribers were no more able to reliably identify many established English-origin *loanwords* in French, such as *sewers* in (6) (initially represented as *sours,* reproducing community norms for its realization, but contravening our orthographic conventions) than most anglophones would be to flag words such as *terrace* or *lawyer* as having originated in French. Accordingly, to enhance consistency in the quantitative analysis of lexical borrowing and integration (e.g. Poplack & Dion 2012; Poplack 2018), we again resort to manual retrieval for lone other-language items.

Transcription, perhaps more than any other aspect of data management, requires ongoing decision making, especially when the protocols adopted blend annotation and some degree of analysis, as in our case. This process was greatly facilitated by the fact that our transcription teams were made up of trained linguists who, as noted, were generally also members of the community under study. All decisions, arrived at jointly, were incorporated into a continually evolving transcription guide to ensure their consistent application, not only to future occurrences of the same phenomena, but also retroactively to previous ones. Transcribers consulted (and, when necessary, updated) this document regularly during transcription and again at each correction phase.

6 Data correction

The goal of transcription is to yield a faithful reflection of what was actually said. As anyone who has worked with spontaneous speech data will attest, achieving it is perhaps the most arduous and time-consuming aspect of corpus construction. Our strategy of maximizing data entry, coupled with our requirements for retrievability, made efficient but effective correction even more of a priority, and one in which we invested considerable effort and experimentation. Teams of correctors, armed with the transcription protocol, alternated manual correction passes of the full transcript (at least one involving re-listening to the original audio recordings) with read-throughs of word lists and concordances. Automated "cleanup" programs targeting hundreds of recurrent transcription errors and spelling fluctuations (e.g., *favour* vs. *favor*) located additional inconsistencies. Interchanging documents among correctors at various correction stages further enhanced reliability. Progress was tracked on spreadsheets occupying large portions of the walls surrounding the workstations, where a dedicated column represented each correction phase for each recording. This enabled the team to ascertain at a glance where in the process each file was located, facilitating alternation of correctors and ensuring the full complement of correction passes, while minimizing the risk of (unplanned) duplication of effort. The result (described in detail in Poplack 1989) is a series of rather pristine corpora that can confidently be used to study morphosyntactic and lexical phenomena without recourse to original audio recordings.[3] And the simplicity of the single-level

transcription protocol lends itself well to adaptation to other tools (e.g., concordance [section 7] or forced alignment [Mielke 2013]).

7 Data retrieval

The generalizations variationists make about language tend to be derived from large-scale quantitative analyses of actual linguistic behavior. Depending on the size of the data set and the corpus frequency of the linguistic phenomenon in question, our studies may be based on data ranging from only a few hundred tokens (e.g., English relative clauses; Poplack et al. 2006; Lealess & Smith 2011) to tens of thousands (e.g., the expression of negation in French [$n = 85{,}447$]; Poplack 2015). Automated data handling is therefore essential. As explained in section 5, our corpora are not tagged for part of speech, nor indeed, in any other way beyond language and speaker. Instead we rely heavily on word concordance software. Many of these are readily available in the public domain, but few meet all of our specific needs. These include the imperative to associate each word with the specific speaker who produced it, while excluding those uttered by non-sample members (i.e., interviewers or other individuals present at the time of recording). We also want to ensure that extraneous elements present in the transcripts (metadata such as name of the interviewer or date, time, and location of the interview, and extralinguistic indications such as "(laughter)") do not figure in word counts or other calculations. Our current tool of choice is Concorder X (Edwards 2006), an in-house program based on existing open-source software (Concorder Pro; Fahrenbacher 2003) that was substantially modified to meet our conditions. Key among these is our requirement that all legal speech material (i.e., that produced by bona fide sample members) be properly associated with the speaker who produced it. Thanks to the basic format of the corpora going in, the data could be retrofitted to the requirements of the software with little to no conversion.

Concorder X is a versatile instrument that efficiently creates word lists and concordances in different configurations and orders (e.g., alphabetical, frequency). These can be efficiently generated for a single speaker, a corpus in its entirety, or a specified subsample thereof. The ability to consult a single word list or concordance (in contrast to as many such documents as there are speakers in the sample) dramatically reduces the amount of time required to locate and extract relevant data. This is particularly valuable for quick-and-dirty feasibility assessments of the frequency (or even existence) in a corpus of targeted linguistic features and to facilitate identification of the speakers who make use of them. As illustrated in (4), the concordance displays each lexical item as a keyword in its immediately preceding and following linguistic context, along with speaker identifier and address in the transcript. Clicking on the keyword takes the user to its original location in the corpus, enabling access to the entire wider context.

Variationist analyses often take the form of determining why one competing variant is chosen over another in a previously specified *variable context*.[4] The variants in question may include such disparate forms as subjunctive versus indicative mood under subjunctive-selecting matrices, modal versus periphrastic versus simple present expressions of future temporal reference, or imperfect versus conditional tenses in protases of hypothetical *if*-complexes, among myriad others. The context view for each token displayed by the concordance is often sufficient to enable the analyst to code it according to factors hypothesized to affect variant choice (e.g., proximity in the future, polarity of the utterance, presence of intervening material between main and complement clauses, and so on). An important caveat, however, is that to locate a token, a lexical signpost must be queried, and the output of the search may be overspecified or fall outside the variable context. Thus, searching *que* 'that' will turn up a lot of complement clauses, but not only subjunctive-selecting ones, searching *si* 'if' will return *if*-clauses, but not only hypothetical ones. Outliers, which will differ according to the parameters of the variable context under investigation, must be detected and disposed of manually. As noted in section 5, location of relevant tokens is further complicated by the fact that many function words such as *que* or *that* are often deleted altogether, as are subjects, copulas, and prepositions, among many others. Some of these may be variants of the variable under study, and therefore must be considered alongside their overt counterparts. Retrieval is therefore based on a combination of automated searches (for forms with overt lexical representations) and manual extraction (for null elements and syntactic variants such as relativization or word-order alternation). Admittedly, manual retrieval is incredibly onerous, especially

with large-scale corpora, but the silver lining is that it enables researchers to uncover the full set of variants of a given variable, a sine qua non for variationist analysis. Crucially, these may include variants that were not recognized or identified at the outset (e.g., selection of the conditional under French subjunctive-selecting matrices or absorption of prepositions in French relative clauses). Manual retrieval also forces researchers to continually (re)familiarize themselves with the data they are analyzing, which become exponentially more abstract as a function of the amount of annotation applied to them. In so doing, we respect another core tenet of the variationist paradigm, which is that linguistic elements must be studied in the *contexts* in which they occur.

However they are located, the extracted tokens are then *coded* according to a series of factors (themselves instantiations of hypotheses about what motivates variant choice) in preparation for statistical analysis. Data coding begins by transferring relevant tokens directly into Excel spreadsheets. Thousands of such tokens can be copied at a time, each automatically accompanied by speaker and line number, already split across columns with the keyword identified in bold red font. This not only ensures correct attribution of each token, but also enhances its visibility, which is particularly useful for correction purposes. We have marshalled many built-in features of Excel (e.g., filters, sorting and tabulating functions, the "hide columns" and subtotaling features) to facilitate coding and improve reliability. Here again, the simplicity of the original transcription protocol lends itself well to these efforts. The resulting strings encoding the targeted linguistic and social aspects of each token can then be fed into multivariate or other statistical analyses to assess the significance, relative magnitude, and direction of their effects. These results constitute the backbone of our analyses.

8 Data preservation

The data housed in the Sociolinguistics Lab were collected at different points in time and in a few cases (e.g., the *Récits du français québécois d'autrefois*, the *Ex-Slave Recordings*) by unaffiliated researchers. The material was therefore recorded using vastly different technologies resulting in different formats (ranging from reel-to-reel and cassettes to digital recordings). All recordings have been preserved in their original medium and subsequently digitized.[5] Digitized audio files were saved in the format most common at the time of digitization and are updated as required. All of the data, from transcripts to token files, exist in multiple copies (digital and physical) stored in multiple secure locations, both within the lab's premises and without (i.e., on the university server).

9 Data life cycle

The large-scale research carried out at the Sociolinguistics Lab may span months, years, or even decades. Analyses and associated token files are frequently revisited, whether to alter them by recoding or eliminating factors that have not proved revealing, to update them by incorporating additional factors to test new hypotheses, to extend the original coding protocol to novel data sets, or any combination of these. Such tasks may sound straightforward, but as any researcher who has attempted them knows, this is far from the case—especially after a considerable amount of time has elapsed. To complicate matters, several researchers may be involved in different aspects of a single project simultaneously, and because turnover among them is not uncommon (students are enrolled for a set amount of time, the tenure of a postdoc or visiting scholar is by definition limited, and so on), rigorous record keeping and documentation at every stage are imperative for the survival and continuing utility of the data. Accordingly, we have invested a good deal of time and effort in establishing protocols to ensure this outcome. These range from labeling conventions identifying file types (e.g., *trans* for transcriptions, *ci* for coding instructions), designated abbreviations for variables (e.g., *NEG* for negation), or standardized cross-study coding for social characteristics, lexical items, and more. Dedicated file management procedures ensure that the relevant versions of project documents (including coding instructions, token files, or outputs of statistical analyses) can be readily accessed. All such documents are relabeled and dated each time they are worked on, taking care to retain older versions in case of data corruption, handling accidents, or simply for reference. They are stored in folders where they can be sorted chronologically or alphabetically. These and other practices have contributed enormously to our goals of replicating and/or reproducing earlier research.

10 Ethical considerations

All of the data housed in the Sociolinguistics Lab have been collected, handled, and stored in compliance with

the ethical considerations outlined by both the relevant granting agencies (http://www.pre.ethics.gc.ca/eng/policy-politique_tcps2-eptc2_2018.html) and the University Research Ethics Board. Indeed, approval must be obtained before the project is even initiated. Our only departure from their requirements involves the principle that fully informed consent should precede data collection. Predictably enough, we have found that initiating interaction with a prospective participant by presenting the linguistic details of the project (which, according to some research ethics board members, should include describing the researcher's interest in use of the subjunctive or constraints on bilingual code-switching, for example), and the attendant reading, discussing, and signing of release forms, is antithetical to creating an atmosphere conducive to obtaining exemplars of unreflecting speech data, let alone any appreciable use of the vernacular. Instead, we explain the purpose of the interview in more general terms, but always including our interest in language use, and address the requirement of fully informed consent by debriefing participants immediately following the recording session (as stipulated by the research ethics board). Although participants are assured at the outset that consent may be withdrawn at any time with no penalty, this has never once occurred in the hundreds of interviews we have conducted. This is testimony to the skill and professionalism of the fieldworkers, all the more so considering that participation is wholly motivated by interest, no monetary incentive having ever been offered.

As noted, privacy and confidentiality of the data are ensured in several ways. The identity of participants is anonymized by means of pseudonyms and speaker numbers, and the material they provided, be it recorded or transcribed, is stored in a secure location under the supervision of the lab's research coordinator. Because of the personal nature of much of the material and the different ethical requirements to which the corpora are subject,[6] they have not been posted on the internet or otherwise distributed publicly. Instead, with the express consent of the participants, on-site access to the material is granted to applicants whose formal request to use the corpora has been approved by the lab's director. Approval is contingent upon the applicant's commitment to respect certain ethical conditions, including the following:

a. No information enabling identification of participants may be included in any paper or presentation based on the corpus.
b. Materials contained in the corpus will not serve as the basis for personal judgments about the opinions, personality, or language of the participant.
c. Materials contained in the corpus will only be cited verbatim in the interest of illustrating a linguistic point, and the content of any such citation must comply with conditions (a) and (b).

Such precautions are particularly important when the linguistic variety in question is non-standard or socially stigmatized, as is the case for much of the data housed at the Sociolinguistics Lab.

11 Data citation

As noted by Conzett and De Smedt (chapter 11, this volume), citation of data is not yet general practice. At the Sociolinguistics Lab, access to corpora is additionally contingent on the commitment to proper attribution, not only to the corpus from which the data were drawn (usually via citation of the published "corpus paper"), but also to the specific speaker who produced it. Any utterance that is reproduced must be attributed to its source via codes identifying the corpus name, speaker number, and utterance address, as exemplified in (1)–(3). The first citation of such an example must be accompanied by a footnote explaining the attribution codes (illustrated here in note 2) and referencing the appropriate corpus name and bibliographic reference, along with the clarification that examples are reproduced verbatim from speaker utterances. Such requirements make the data and associated claims readily verifiable, enhancing the reproducibility and accountability of any study based on them (Berez-Kroeker et al. 2018).

12 Applications

In keeping with our understanding that the utility of a corpus lies to a large extent in the versatility of uses to which it can be put, Sociolinguistics Lab corpora have lent themselves to the study of a wide variety of research questions. These include constraints on different manifestations of language contact (lexical borrowing, code-switching, grammatical convergence), issues involving language change (across the life span, contact-induced, resistance to by linguistic isolates, the role of the media in), grammaticalization (in English, French, and pan-Romance), heritage language maintenance, prescription

versus praxis, and the origins of AAVE, among others. Linguistic variables marshalled to illustrate these issues include, but are not limited to, such disparate phenomena as copula deletion, dative alternation, question formation, word order variation, preposition stranding, relativization strategies, auxiliary alternation, the variable expression of present, past, and future temporal reference, mood alternation, gender assignment, and plural marking. For references to these and other publications based on lab corpora, see the lab's Publications page (http://www.sociolinguistics.uottawa.ca/publications.html).

13 Epilogue

In the current academic climate, the data-driven research that corpora enable is often denigrated as theoretically uninteresting linguistics—or even *not* linguistics—when not outright penalized. Outside of the narrow domain of (variationist) sociolinguistics, researchers are rarely (if ever) credited for the extreme efforts we undertake to collect, transcribe, organize, and share the vast quantities of precious spontaneous speech data that constitute many corpora. On the contrary, we are often chastised for the wonky distributions, empty cells, and sometimes less-than-optimal quantities of rare variants that characterize natural speech. It is not uncommon for leading linguistics journals, viewed as mouthpieces of the field, to reject or request major revisions of quantitative work reporting sparse or disproportionate data distributions, even when entire massive data sets have been systematically combed to yield those tokens. This while encouraging papers pushing complex and wide-ranging theories based on no production data at all beyond native speaker intuitions, often with little prospect of replicability or reproducibility. Linguists accustomed to painstaking analysis of language *as it is spoken* recognize that uneven distributions are the rule rather than the exception. But widespread unfamiliarity with the facts of actual spontaneous speech data, and the growing penchant for substituting it with more easily accessible surrogates (e.g., internet language), whose specific provenance we may know little or nothing about, have conspired to obscure these core characteristics. It is to be hoped that the practices outlined in this chapter, most of which have been staples of the methodology of variationist sociolinguistics since its inception over half a century ago, will contribute, along with the remainder of this volume, to rectifying this imbalance.

Acknowledgments

The work reported here was generously supported by the Social Sciences and Humanities Research Council of Canada through its Canada Research Chairs program and numerous research grants, as well as by the Killam Foundation, the Pierre Elliott Trudeau Foundation, the Ontario Ministry of Research and Innovation, the Canada Foundation for Innovation, and the Ontario Innovation Trust. I was first introduced to the concept of data management by the teachings and the example of William Labov. Respect for data and the speakers who provided it was at the core of his famous LIN 560 class at the University of Pennsylvania. Much of what I learned there formed the basis for my own data collection and handling practices at the Sociolinguistics Lab at the University of Ottawa, which I founded and have directed since the early 1980s, as well as for the Urban Dialectology courses we have been running here since then. My efforts have been immeasurably aided, and in fact surpassed, by generations of smart, committed, enthusiastic (and above all, incredibly organized!) students and associates, who have continued to translate those teachings into ever more productive and efficient ways of dealing with data. If I have been able to boast that we can reproduce or replicate an analysis ten years after the fact, it is thanks to them. I am also grateful to two anonymous reviewers for comments that enriched this chapter.

Notes

1. Real names are stored in a secure and confidential location during corpus construction and are disposed of when anonymization has been completed.

2. Codes in parentheses refer to corpus name, speaker number, and utterance address in the *Quebec English Corpus* (Poplack, Walker, & Malcolmson 2006) in (1) and (3), the *African Nova Scotian English Corpus* (Poplack & Tagliamonte 1991) in (2), and the *Ottawa-Hull French Corpus* (Poplack 1989) in (5) and (6). Examples are reproduced verbatim from speaker utterances.

3. One analysis of a 3.5-million-word corpus showed an average of one error per 520 words (including errors of all magnitudes, ranging from a missing space to a misspelled word), an excellent record by the standards of speech corpora.

4. The *variable context* is defined as the context in which variants alternate with no change in referential meaning.

5. With the exception of the older Sociolinguistic Archives recordings collected by students.

6. These have changed appreciably over the duration, with the result that different corpora are bound by different ethical conditions.

References

Beal, Joan, Karen Corrigan, and Hermann Moisl. 2007. *Creating and Digitizing Language Corpora*. Volume 1: *Synchronic Databases*. Houndmills: Palgrave-Macmillan UK.

Berez-Kroeker, Andrea L., Lauren Gawne, Susan Smythe Kung, Barbara F. Kelly, Tyler Heston, Gary Holton, Peter Pulsifer, et al. 2018. Reproducible research in linguistics: A position statement on data citation and attribution in our field. *Linguistics* 56 (1): 1–18.

Bird, Steven, and Gary Simons. 2003. Seven dimensions of portability for language documentation and description. *Language* 79 (3): 557–582.

Edwards, James. 2006. *Concorder X: Program and Documentation*. Ottawa: University of Ottawa Sociolinguistics Laboratory.

Fahrenbacher, Matt. 2003. *Concorder Pro 1.0: A Text-Analysis Tool for Mac OS X*. N.p.: Humongous Elephants and Tigers.

Labov, William. (1966) 2006. *The Social Stratification of English in New York City*. 2nd ed. Cambridge: Cambridge University Press.

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

Labov, William. 1984. Field methods of the project on linguistic change and variation. In *Language in Use: Readings in Sociolinguistics*, ed. John Baugh and Joel Sherzer, 28–54. Englewood Cliffs, NJ: Prentice Hall.

Lealess, Allison V., and Chelsea Smith. 2011. Assessing contact-induced language change: The use of subject relative markers in Quebec English. *Ottawa Papers in Linguistics* 36: 20–38.

Mielke, Jeff. 2013. Ultrasound and corpus study of a change from below: Vowel rhoticity in Canadian French. *University of Pennsylvania Working Papers in Linguistics* 19 (2): article 16.

Poplack, Shana. 1989. The care and handling of a mega-corpus. In *Language Change and Variation*, ed. Ralph Fasold and Deborah Schiffrin, 411–451. Amsterdam: Benjamins.

Poplack, Shana, ed. 2000. *The English History of African American English*. Oxford: Blackwell Publishers.

Poplack, Shana. 2007. Foreword. In *Creating and Digitizing Language Corpora*, ed. Joan Beal, Karen Corrigan, and Hermann Moisl, ix–xiii. Houndmills: Palgrave-Macmillan UK.

Poplack, Shana. 2015. Norme prescriptive, norme communautaire et variation diaphasique. Variations diasystématiques et leurs interdépendances. In *Travaux de linguistique romane*, ed. Kristen Kragh and Jan Lindschouw, 293–319. Strasbourg: Société de linguistique romane.

Poplack, Shana. 2018. *Borrowing: Loanwords in the Speech Community and in the Grammar*. Oxford: Oxford University Press.

Poplack, Shana, and Johanne Bourdages. 2005. *Le français en contexte: Milieux scolaire et social*. Ottawa: University of Ottawa. (Social Sciences and Humanities Research Council of Canada research grant #410-2005-2108.)

Poplack, Shana, and Nathalie Dion. 2012. Myths and facts about loanword development. *Language Variation and Change* 24 (3): 279–315.

Poplack, Shana, Lidia-Gabriela Jarmasz, Nathalie Dion, and Nicole Rosen. 2015. Searching for "Standard French": The construction and mining of the *Recueil historique des grammaires du français*. *Journal of Historical Sociolinguistics* 1 (1): 13–56.

Poplack, Shana, Adrienne Jones, Allison V. Lealess, Martine Leroux, Chelsea T. Smith, Yukiko Yoshizumi, Lauren Zentz, and Nathalie Dion. 2006. Assessing convergence in contact languages. Paper presented at New Ways of Analyzing Variation 35. Ohio State University, Columbus, November 9–12.

Poplack, Shana, and Stephen Levey. 2010. Contact-induced grammatical change. In *Language and Space—An International Handbook of Linguistic Variation*. Volume 1: *Theories and Methods*, ed. Peter Auer and Jürgen E. Schmidt, 391–419. Berlin: Mouton de Gruyter.

Poplack, Shana, and David Sankoff. 1987. The Philadelphia Story in the Spanish Caribbean. *American Speech* 62 (4): 291–314.

Poplack, Shana, and Anne St-Amand. 2007. A real-time window on 19th century vernacular French: *The Récits du français québécois d'autrefois*. *Language in Society* 36 (5): 707–734.

Poplack, Shana, and Sali Tagliamonte. 1991. African American English in the diaspora: Evidence from old-line Nova Scotians. *Language Variation and Change* 3 (3): 301–39.

Poplack, Shana, and Sali Tagliamonte. 2001. *African American English in the Diaspora*. Oxford: Basil Blackwell.

Poplack, Shana, James A. Walker, and Rebecca Malcolmson. 2006. An English "Like no other"?: Language contact and change in Quebec. *Canadian Journal of Linguistics* 51 (2/3): 185–213.

Poplack, Shana, Lauren Zentz, and Nathalie Dion. 2012. What counts as (contact-induced) change. *Bilingualism: Language and Cognition* 15 (2): 247–254.

Van Herk, Gerard, and Poplack, Shana. 2003. Rewriting the past: Bare verbs in the Ottawa Repository of Early African American Correspondence. *Journal of Pidgin and Creole Languages* 18 (2): 231–266.

This book was set in Stone Serif and Stone Sans by Westchester Publishing Services.

Library of Congress Cataloging-in-Publication Data

Names: Berez-Kroeker, Andrea L., editor. | McDonnell, Bradley James, editor. | Koller, Eve, editor. | Collister, Lauren B., editor.
Title: The open handbook of linguistic data management / edited by Andrea L. Berez-Kroeker, Bradley McDonnell, Eve Koller and Lauren B. Collister.
Description: Cambridge, Massachusetts : The MIT Press, [2021] | Series: Open handbooks in linguistics series | Includes bibliographical references and index.
Identifiers: LCCN 2020044363 | ISBN 9780262045261 (hardcover)
Subjects: LCSH: Computational linguistics. | Natural language processing (Computer science) | Data mining.
Classification: LCC P98 .O64 2021 | DDC 410.285—dc23
LC record available at https://lccn.loc.gov/2020044363