44 Managing Data in a Formal Syntactic Study of an Under-Investigated Language (Uzbek)

Vera Gribanova

1 Introduction

In this data management use case, I describe the considerations and decision points that arose in the course of organizing and making accessible certain parts of the data I collected through fieldwork on the morphosyntax of Uzbek, an under-studied Turkic language of Central Asia. The discussion will focus on one such data set—on the syntax of cleft constructions—that draws on, and documents the results of, empirical studies that formed the basis of my work on the syntax and morphosyntax of Uzbek between 2009 and 2019. The kind of fieldwork involved in this study involves native speaker[1] judgments about the grammaticality of sentences or morpheme combinations in words and occasionally judgments about interpretation as well; this approach will be more or less familiar to most theoretical linguists and involves fieldwork practices that are already very well described in textbooks and articles on fieldwork methodology (e.g., Bowern 2008; Vaux, Cooper, & Tucker 2007, among many others). I therefore do not focus especially on this aspect of the data management process in this chapter, except where it touches on fieldwork choices I made that were specific to Uzbek or to my investigation of it. Rather, my focus is on what motivated the organizational choices I made when it comes to managing data sets involving numerous moving parts. Secondarily, I will also discuss the motivations behind my decision to make publicly accessible the results of these studies and the various logistical decisions I faced when doing so.

As will become clear in the second and third sections of this chapter, the process that eventually led to the existence of the online Uzbek language data archive described here was initiated in the context of a constellation of circumstances and language-specific factors that is not likely to hold across a particularly broad range of investigative situations. It is nevertheless the case that the lessons learned in this process may be useful to readers who are considering engaging in such a project, using data that they have collected via fieldwork, especially if—like me—they have little to no training in language documentation and tend to work with fieldwork-collected language data in service, primarily, of syntactic analysis.

2 Uzbek

Uzbek is an under-studied Turkic language spoken by about eighteen million people, primarily in Uzbekistan. Because of the history of Soviet occupation and the geography of Uzbekistan, very few native speakers of Uzbek are monolingual: Russian, Tajik, and languages of neighboring Turkic-speaking regions (Kazakh, Kyrgyz, Turkmen) are common second and third languages, with English as an ever more prominent contender, especially among younger generations. I collect basic information about the speakers I work with, including their region of origin and their range of languages used; to the extent possible, I try to work with consultants whose primary language is Uzbek, rather than Russian. All of this is especially relevant in light of the fact that in the course of my work on this language (since about 2009), one of my most rewarding and challenging discoveries has been that there is a great deal of not only phonological and lexical, but also morphological and syntactic variation among speakers. As far as I am able to tell, this variation is conditioned primarily by generation and region. The fact of variation across age groups is not surprising: with independence from the Soviet Union in 1991 came a great number of changes, among them a resurgence of national pride and pride in the Uzbek language, which was reflected in numerous educational policy decisions.

Young people's education stopped being centered on the acquisition of Russian as a professional language and as a lingua franca; in the majority of cases, Uzbek became the primary language of schooling. I suspect, but cannot know for sure, that this brought with it numerous other small shifts in the syntactic patterns of young people's language. The regional variation is perhaps even less surprising. Uzbekistan is bordered by five countries, three of them with Turkic national languages; Uzbek speakers along borders with these languages tend to adopt certain features of the neighboring language. This is quite directly observable phonologically, for example, from the fact that standard Uzbek has no vowel harmony, but regional variants close to the border with Kazakhstan, Kyrgyzstan, and Turkmenistan have vowel harmony. Regional variation in Uzbek, and its role in Uzbek dialect classification, is discussed primarily on the basis of observable differences in lexical choice and phonetic variation, as in, for example, the work of Turaeva (2015). Although morphosyntactic variation is not—to my knowledge—documented in the descriptive literature on Uzbek, it is nevertheless the case that investigation of the language led me to find a situation where judgments, upon first impression, did not always systematically align across native speaker consultants. Further investigation showed that this lack of alignment in judgments was not random: it was limited to certain syntactic and morphosyntactic domains and displayed certain patterns of covariation that called for a deeper explanation.

A second circumstance that required some thought about the organizational principles I would apply to structuring collected data is that Uzbek is an underinvestigated language. When I began work on the language in 2009, although there were some high-quality descriptive grammars (Bodrogligeti 2003; Kononov 1960; Sjoberg 1963), there was no sustained, theoretically driven work on the structural properties of Uzbek, as far as I was aware. While the existence of descriptive resources provided me with an invaluable point of departure for my work, they were in some cases outdated, they provided no negative evidence, and they also provided no insight into the morphosyntactic variation I ultimately encountered. Finally, it is unsurprising that my theoretical perspective often led me to look for information about the structure of Uzbek that was not encompassed in descriptive grammars. Taken together, all of this led me to think that the areas of contemporary Uzbek grammar that I was interested in investigating required more systematic data collection—for example, in the form of well-formedness surveys of entire morphological paradigms, rather than just the forms of immediate interest to me—so that I could both get a holistic sense of the relevant patterns and get a handle on the variation, where I was finding it.[2] The resulting data sets involved patterns with a lot of moving parts, in terms of the grammatical features that were at play. It became apparent relatively quickly that I needed a way to sort and tag my data so that I could visualize how these factors were interacting. As I will discuss, besides the many obvious benefits this kind of organization brought with it, this approach led to significant analytical findings that most likely would not have emerged otherwise.

3 Data collection and organization: Uzbek clefts

The project I describe here is the first of an ongoing series; every project in this series involves examination of distinct components of Uzbek clause structure and morphosyntax, but they share some common themes, one of which is the observation that certain sub-components of the relevant patterns involve a significant degree of variation in judgment. In all cases, including the project discussed in detail here, the basic methodology for treating data collection is quite simple: in the typical case, I prepare examples I want to ask native speaker consultants about in a raw text file, editing that file as I go through the session with every individual speaker. I keep a separate file with information about speakers' area of origin, educational history, approximate age, and gender. Elicitation sessions inside each text file are labeled with keywords, the consultant's name, and the date.

I generally do not record elicitation sessions. Audio recording can be very helpful, and it is often the norm in the kind of data collection I am describing here. However, in certain socio-cultural contexts it is also inappropriate, and can hinder the process of building trusting relationships between the language investigator and the language consultant. In the case of Uzbek, most of the people I work with are old enough to have grown up in a Soviet context, or else their parents did. Audio recording in this context gives rise to an automatic distrust of the investigator, even if the materials being recorded are not sensitive. Secondarily, many of the Uzbek consultants I work with have known me for the course of almost

a decade, and we have formed strong friendship ties. Unfettered by audio recordings, our elicitation sessions can toggle smoothly from discussions of our lives—academic and work pursuits; children's growing pains; discussions of Uzbek culture, abroad and at home—back to Uzbek example words and sentences. This would never be possible in a context where consultants knew that they were being recorded, and I believe that the freedom offered by not having to restrict the discussion narrowly to the Uzbek language at any given moment is an important one for building long-term relationships with consultants.

Recording language users' reactions to the materials I prepare for evaluation in a text file is usually sufficient when I am working with data and judgments Linzen and Oseki (2018:3) call "Class II"—these are "judgments that illustrate uncontroversial facts about the grammar of the language." Each of the cases I discuss required further investigation, in large part because they involved variation in judgments that illustrate more subtle contrasts that are crucial to theory-building (Linzen and Oseki label these "Class III").

3.1 Uzbek sluicing-like constructions and cleft strategies

The investigation I discuss here took place around 2009–2011, during which time I worked with Uzbek language consultants in Uzbekistan, Russia, and the United States. Much of the discussion that follows is drawn from Gribanova (2013), which contains far more detail than I can provide here. A portion of the data that was collected for this investigation is now archived, through the Stanford Digital Repository, at https://purl.stanford.edu/qs579kq8188—more detail about this aspect of the project is provided in section 4.

Like all Turkic languages, Uzbek uses the WH-*in-situ* strategy to form content questions (1). Uzbek also makes use of a construction (as in (2)) that looks on the surface like *sluicing*: ellipsis of some roughly clause-sized material, stranding a WH-phrase. For the sake of neutrality, I adopt the term *sluicing-like construction* (SLC) to describe the phenomenon in Uzbek.

(2) Siz kim-ga-dir pul ber-a-siz, lekin
You some-DAT-one money give-PRS-2SG but
kim(-ga)-lig-i-ni bil-ma-y-man.
who(-DAT)-COMP-3SG.POSS-ACC know-NEG-PRS-1SG
"You give money to someone, but I don't know who."

This combination of possibilities is surprising, from a typological perspective, for prominent theories of the nature of ellipsis. One of the best-defended ideas in this domain is that in constituent ellipsis, the elided constituent is unpronounced but nevertheless fully articulated and present as far as the syntax is concerned (Merchant 2001; Ross 1969). Ellipsis of the type in (1) is therefore typically taken to be the result of WH-movement—the operation that fronts WH-phrases to the left edge of clauses in languages like English—and an ellipsis operation that targets a clause-sized constituent. According to such an approach, languages without an overt WH-movement operation (among them, Uzbek) should not countenance genuine sluicing. The major thrust of Gribanova (2013) is that Uzbek SLCs such as (2) should be derived not from a clause featuring WH-movement—a

(1) a. Umida universitet-da O'zbek til-ni o'qi-y-di.
Umida university-LOC Uzbek language-ACC learn-PRS-3SG
'Umida learns Uzbek at the university.'

b. Kim universitet-da O'zbek til-ni o'qi-y-di?
who
'Who learns Uzbek at the university?'

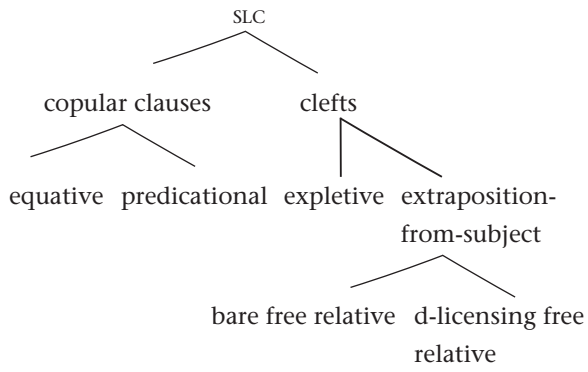
c. Umida qayer-da O'zbek til-ni o'qi-y-di?
where-LOC
'Where does Umida learn Uzbek?'

d. Umida universitet-da qaysi til-ni o'qi-y-di?
which language-ACC
'Which language does Umida learn at the university?'

strategy that would be incompatible with the organizing syntactic principles of the language—but rather from a clause that is underlyingly either copular or a cleft.[3] The challenge for such a proposal is to show that the various properties of SLCs follow from the properties of copular and cleft constructions; such a strategy therefore requires an independent investigation of copular and cleft clauses in the language.

Investigation of the copular and cleft structures of Uzbek (illustrated schematically in (3)) turned out to be a challenge, in part because there are two cleft strategies that share some features and can look similar. The idea that cleft constructions may actually correspond to two different structural possibilities in a single language is very well established (Jespersen 1927, 1937; Pinkham & Hankamer 1975), but has been difficult to show for languages like English (Gundel 1977).

(3)



The two cleft strategies of interest to us in Uzbek are shown in (4) and (5); they differ in (i) whether the copula agrees, (ii) whether the pivot shows case connectivity, (iii) whether the pivot can be an adjunct, (iv) whether extraposition from subject position (of the cleft clause or free relative) is optional or obligatory, and (v) whether the clause extraposed from subject position is a free relative or a cleft clause.

(4) Extraposition from subject (EFS)

a. Men ko'r-gan, siz e-di-ngiz.
I see-PST.PTCP you COP-PST-2SG
"Who I saw was you."

b. U siz e-di-ngiz, men ko'r-gan.
3SG you COP-PST-2SG I see-PST.PTCP
"It was you, who I saw."

As noted in (5), the EXPL strategy is not available to all speakers whom I consulted, and within (5), there is variation especially as to whether an accusative nominal can serve as the pivot of the cleft—a pattern I will return to next.

3.2 Organizing the data and identifying patterns

On the surface, the cleft data I was eliciting was difficult for me to make sense of by just looking at individual examples of grammatical and ungrammatical possibilities. There were two sources for this difficulty: first, the fact that there are five properties of these clefts, listed in (i)–(v) in section 3.1, which ultimately point to two underlying structures; and second, the fact that a subpart of the overall patterns exhibits variation.

Tackling the first problem is what initially led me to take a more systematic approach to data organization and management than the basic methodology that I outlined at the beginning of section 3. While there are probably far more technologically sophisticated approaches I could have taken, for my purposes it sufficed to list all the cleft structures I had elicited in an Excel file. Each sentence was annotated with information about speaker judgment(s) and was tagged for a subset of the various pieces of information in (i)–(v).[4] I ended up doing this for an extensive set of sentences, and by using the "sort data" (by tags) function in Excel, I was able to understand how clusters of properties were patterning together. This revealed, for example, that the EFS strategy

(5) Expletive (EXPL)

a. Siz-ga e-di, men pul ber-gan-im. [*subject to variation*]
you.DAT COP-PST.3SG I money give-PST.PTCP-1SG.POSS
"It was to you that I gave money."

b. Siz-ni e-di, men ko'r-gan-im. [*subject to variation*]
you-ACC COP-PST.3SG I saw-PST.PTCP-1SG.POSS
"It was you that I saw."

always required a non-case-marked nominal in its pivot and that this same cleft strategy required agreement on the copula. This is distinct from the EXPL cleft, where no copular agreement is ever attested and a broader range of constituents is permitted in the pivot position. Table 44.1 summarizes the entire range of properties.

Configuring the data set in this way also revealed gaps in my exploration—it turned out that there were pairings of properties I had implicitly assumed would or would not co-occur, but had not actually tested in my initial elicitations. Once I had put all the data I had into a spreadsheet, I was able to see what kinds of clusterings or pairings of properties I did not yet have data for and to follow up on those pairings. After a few rounds of this, I was able to visualize the data and to conclude that I was looking at two patterns that were only superficially similar. The next step was to connect these clusters to structural analyses that could provide a principled explanation as to why the various properties in table 44.1 pattern the way they do.

Dealing with the second challenge required me to home in on the empirical domains where the variation was found. Once I had gone through the initial stages of organizing the data it became clear that not all speakers would accept EXPL clefts at all. I designed a further survey that contained primarily the types of strings where I knew there was some variation. This survey demonstrated that those speakers who did accept the EXPL cleft exhibited further variation as to whether they accepted accusative-marked nominal pivots—a pattern that is also found in Japanese (Kizu 1997) and Turkish (Merchant 1998) clefts. A direct benefit for my analysis of adding these data to my data set was that the variation in acceptability of accusative-marked pivots turned out to be directly correlated to whether speakers would accept an accusative-marked WH-remnant in an SLC configuration. This co-variation between cleft and SLC behavior ended up being one of the strongest and most satisfying pieces of evidence in Gribanova (2013) for the analytical claim that SLCs were in fact derived from cleft and copular constructions in Uzbek. Figure 44.1 gives a partial snapshot of the survey data, although certain columns in the spreadsheet were collapsed for the purposes of fitting the spreadsheet on the page. The crucial result—all and only the speakers who accept accusative pivots in clefts also accept accusative remnants in the SLC—can be seen in columns A–J, where each column represents a judgment by a native Uzbek speaker.

4 Archiving and transparency

4.1 Motivations for archiving data from syntax field research on Uzbek

The process of putting the evidence I had collected together and annotating that data led me to consider the question of whether to make the data publicly accessible. Important considerations specific to archiving methodology, archiving ethics, and intellectual property are discussed extensively by Andreassen (chapter 7, this volume); Holton, Leonard, and Pulsifer (chapter 4, this volume); and Collister (chapter 9, this volume), respectively. My aim here is to try to articulate why I was driven to archive my own data and the considerations involved in doing so. These considerations are especially relevant, given that there are many reasons such a task may be difficult to undertake. Preparing data

Table 44.1
The distinctive properties of EXPL and EFS clefts

Extraposition	Expletive
The copula agrees with the most accessible nominal	The copula bears default agreement
No case connectivity on the pivot	Case connectivity on the pivot
The pivot is an argument	The pivot is an argument or an adjunct
Subject position contains a free relative or a third-person pronoun	No overt element fills the subject position; null expletives only
Extraposition of the free relative from subject position is optional	Extraposition of the cleft clause from object position is obligatory
The free relative is a genuine headless relative clause (one of two types)	The cleft clause is not a traditional relative clause

parse	gloss	A	B	C	D	E	F	G	H	I	J	clause type	[remnant/pivot] [case/category]
Men ol-ib kel-gan-im Farhod-dan e-di.	1SG take-CV come-PST.PTCP-1SG.POSS Farhod-ABL COP-PST					ok	ok	ok		ok		copular clause	ablative
Men ol-ib kel-gan-im Toshkent-dan e-di	1SG take-CV come-PST.PTCP-1SG.POSS Tashkent-ABL COP-PST					ok	ok	ok		ok		copular clause	ablative
U kim-dan-dir pul ol-di, lekin kim-dan-lig-i-ni bil-ma-y-man.	3SG who-ABL-one money take-PST but who-ABL-NMLZ-3SG-ACC know-NEG-PRS-1SG	ok	ok	ok	ok	ok	ok	ok	ok	ok		SLC	ablative
U kim-dan-dir pul ol-di, lekin kim-dan ekan-lig-i-ni bil-ma-y-man.	3SG who-ABL-one money take-PST but who-ABL EVID-NMLZ-3SG-POSS-ACC.POSS-ACC know-NEG-PRS-1SG	ok	ok	ok	ok	ok	*	ok	ok	ok		SLC	ablative
U kim-dan-dir pul ol-di, lekin u kim-dan ekan-lig-i-ni bil-ma-y-man.	3SG who-ABL-one money take-PST but 3SG who-ABL EVID-NMLZ-3SG-POSS-ACC.POSS-ACC know-NEG-PRS-1SG	*	*	*	ok	*	*	ok	ok	ok		SLC	ablative
Siz-ni e-di, men ko'r-gan-im.	2SG-ACC COP-PST 1SG see-PST.PTCP-1SG.POSS	**ok**	*	*	*	*	*	*	**ok**	**ok**	**ok**	**cleft**	**accusative**
U siz-ni edi, men ko'r-gan-im.	3SG 2SG-ACC COP-PST 1SG see-PST.PTCP-1SG.POSS	*	*	*	*	*	*	*	*	*		**cleft**	**accusative**
Men kim-ni-dir ko'r-d-im, lekin kim-ni-lig-i-ni bil-ma-y-man.	1SG who-ACC-one see-PST-1SG but who-ACC-NMLZ-3SG.POSS-ACC know-NEG-PRS-1SG	**ok**	*	*	*	*	*	*	**ok**	**ok**	**ok**	**SLC**	**accusative**
Men kim-ni-dir ko'r-d-im, lekin kim-ni ekan-lig-i-ni bil-ma-y-man.	1SG who-ACC-one see-PST-1SG but who-ACC EVID-NMLZ-3SG-POSS-ACC.POSS-ACC know-NEG-PRS-1SG	?							*			**SLC**	**accusative**
U chiroy-li e-di, men ko'r-gan-im	3SG beauti-ful COP-PST 1SG see-PST.PTCP-1SG.POSS	*		*	*	ok	*	*	*	*		cleft	AP
Men ko'r-gan-im chiroy-li e-di.	1SG see-PST.PTCP-1SG.POSS beauti-ful COP-PST	*		ok	ok	ok	*	ok	ok	ok		copular clause	AP
Men ko'r-gan qiz chiroy-li e-di.	1SG see-PST.PTCP girl beauti-ful COP-PST	ok		ok	ok	ok	ok	ok	ok	ok		copular clause	AP
U siz e-di-ngiz, men ko'r-gan-im.	3SG 2SG COP-PST-2SG 1SG see-PST.PTCP-1SG.POSS	ok	ok	ok	ok	*	ok	ok	*	ok		cleft	bare
Men ko'r-gan-im, siz e-di-ngiz.	1SG see-PST.PTCP-1SG.POSS 2SG COP-PST-2SG	ok	ok	ok	ok	ok	ok	ok	ok	ok		copular clause	bare
Men ko'r-gan-im siz e-di.	1SG see-PST.PTCP-1SG.POSS 2SG COP-PST	*		*	*	*	*	*	*	*		copular clause	bare
Men kim-ni-dir ko'r-d-im, lekin u kim ekan-lig-i-ni bil-ma-y-man.	1SG who-ACC-one see-PST-1SG but 3SG who EVID-NMLZ-3SG-POSS-ACC.POSS-ACC know-NEG-PRS-1SG	ok	ok	ok	ok	ok	ok	ok	ok	ok		SLC	bare
U kim-ga-dir pul ber-di, lekin u kim ekan-lig-i-ni bil-ma-y-man.	3SG who-DAT-one money give-PST but 3SG who EVID-NMLZ-3SG-POSS-ACC.POSS-ACC know-NEG-PRS-1SG	ok	ok	ok	ok	ok	ok	ok	ok	ok		SLC	bare
Men sovg'a qil-gan-im, siz-ga e-di.	1SG present do-PST.PTCP-1SG.POSS 2SG-DAT COP-PST					ok	*	ok		ok		copular clause	dative
U kim-ga-dir pul ber-di, lekin kim-ga-lig-i-ni bil-ma-y-man.	3SG who-DAT-one money give-PST but who-DAT-NMLZ-3SG.POSS-ACC know-NEG-PRS-1SG	ok	ok	ok	ok	ok	ok	ok	ok	ok		SLC	dative
U kim-ga-dir pul ber-di, lekin u kim-ga ekan-lig-i-ni bil-ma-y-man.	3SG who-DAT-one money give-PST but 3SG who-DAT EVID-NMLZ-3SG-POSS-ACC.POSS-ACC know-NEG-PRS-1SG	*	*	*	ok	*	*	ok	ok	ok		SLC	dative
U kim-ga-dir pul ber-di, lekin kim-ga ekan-lig-i-ni bil-ma-y-man.	3SG who-DAT-one money give-PST but who-DAT EVID-NMLZ-3SG-POSS-ACC.POSS-ACC know-NEG-PRS-1SG	ok	ok	ok	ok	ok	*	ok	ok	ok		SLC	dative
Men sovg'a qil-gan-im, siz uchun e-di.	1SG present do-PST.PTCP-1SG.POSS 2SG for COP-PST					ok	*	ok		ok		copular clause	PP

Figure 44.1
Variation in the acceptability of accusative pivots.

for archiving requires resources, in the form of time and money. And as Berez-Kroeker et al. (2019) and Alperin, Schimanski, La, Niles, and McKiernan (chapter 13, this volume) have pointed out, a major de-motivating factor is that our field on the whole does not have clear guidelines for attributing credit for data stewardship for the purposes of tenure and promotion.

Despite these difficulties, once I had organized the data I had collected for myself, I was quite motivated to make it accessible; there were clear benefits both to my own work and to the community of native Uzbek speakers with whom I worked. The central benefit to my own work and analytical process was knowing that I could be open and transparent about areas where judgments were subject to variation, and I was able to let go of the pressure to present an idealized snapshot of the language. Putting all of the elicited materials online has enabled me to embrace the actual linguistic situation as I have found it; this has fed back into my analytical claims in a tremendously beneficial way, as I discussed in section 3. On the logistical side of things, there is also a clear sense in which working on an under-investigated language requires papers to be on the longer side, and this can be a source of tension or difficulty in the publication process. It is almost always the case that not all the data collected in an investigation can be provided in any reasonable-length manuscript. It has been beneficial for me to be able to refer, in submitted papers, to the entire range of data archived online. This has enabled me to feel I am being responsible to the full range of facts I have collected, and to the language itself, without necessitating that the entirely of the collected data be in the materials I submit for publication.

It is important to acknowledge that open archiving of data collected via fieldwork is not always the appropriate choice for every community; for a useful discussion, see Holton, Leonard, and Pulsifer (chapter 4, this volume). Such decisions will hinge at least in part on the specific language, type of data, and the language politics associated with the community of speakers. In the case of the Uzbek-speaking community, both in Uzbekistan and in the diaspora, the speakers I worked with felt proud that their language was receiving a scholarly kind of attention.[5] After decades of linguistic oppression in the Soviet context, many Uzbek speakers are happy to see their language thriving in the context of Uzbek national language policy, and the fact that someone studies their language has generally been, in my experience, a source of excitement.

A final and important motivation for making public the results of fieldwork-based data collection in syntactic investigations is that this can help to address many of the perennial empirical challenges associated with judgment-based studies. Sprouse, Schütze, and Almeida (2013) demonstrated experimentally that English-based judgments collected using a relatively small sample can be quite reliable. While this was shown to be true for English, the situation for judgments collected about certain grammaticality contrasts in other majority languages—Hebrew and Japanese—have been shown to be far less reliable (Linzen & Oseki 2018). One reason this might be the case—as suggested by both Phillips (2009) and Linzen and Oseki—is that English-based judgments are exposed to scrutiny by speakers of English in the review process, providing more opportunities for questionable judgments to be flagged. This same situation does not

always hold for other languages, and it *especially* does not hold for languages with few or no native speaker linguists that have the kind of generative training required to serve as reviewers for the relevant journals. What this means for a language like Uzbek is that the core evidence on which analyses are based is subject to less scrutiny; as a consequence, there is a greater risk that contrasts or data points may be incorrect or not reliable.

The first step in remedying this situation is to be aware of it; in my case, it has also helped to take the extra step of being as transparent as possible about the data that I collect, so that interested parties—language specialists, native speaker linguists of all traditions, language consultants, and others—can access the original judgments on which I am basing my analysis and challenge those judgments, or use them to make the case for an alternative understanding of the empirical space. In at least one recent case, graduate students working on Uzbek for a field methods course developed a project looking at some overlapping areas of interest with mine and were able to draw on the data I had collected on predicate formation,[6] in combination with further independent fieldwork, to formulate their own analysis (Donovan & Nematova 2019). I see this as an early, positive sign that making data publicly accessible in this way can lead to real scientific progress.

4.2 Logistics

There are several decisions I made with respect to how the data I collected should be represented and made accessible online. I will recount those decisions here in case the same considerations arise for other scholars considering a similar move, acknowledging simultaneously that many of these decisions are specific to my subfield, the language in question, and my institution. Kung (chapter 8, this volume) provides a detailed discussion of the kinds of considerations that may be at stake in making decisions about data management and organization, whether or not that data ends up being publicly accessible.

A major factor that made this process easier for me was the existence of the Stanford Digital Repository,[7] which is run by the Stanford Library system and which is designed for managing and depositing scholarly information resources of a variety of forms. The ready availability of a digital archive with solid data management practices and institutional support made a world of difference to me. The repository makes scholarly data available through the library catalog and provides a permalink for each folder in the digital archive, so that I can link to data in publications without being worried that the link will change. Similar repositories already exist or are being developed across universities internationally; the Linguistic Society of America has a useful discussion of considerations in choosing an archive for language data and provides a list of a range of digital repositories.[8] After a bit of research and deliberation I decided that using the .csv file format would be the most accessible way to present and disseminate the data I had collected; the remaining decisions concerned what kind of information to put in each data set online.

To maximize accessibility,[9] I used the Latinate writing system for Uzbek (even though Cyrillic still continues to be used quite productively in Uzbekistan). For each token, I typically provided the linguistic utterance as an Uzbek speaker would write it, then provided a morpheme-by-morpheme parse, and a gloss. Apart from this, I provided information about judgments of any speakers that were asked about the string and any comments they might have made about it in a separate "notes" field. Separate from this, the only information provided in the body of each .csv file was the series of tags associated with each sentence; these tags, in combination with the ability to sort data by tag, were very useful to me in my own analytical work, so it made sense to leave this information in the posted files as well.

For each .csv file that contains judgments, there is an associated file, sometimes containing information about speakers, and noting any abbreviations used in the tagging or morpheme-by-morpheme glosses. Speakers were given a random letter to cross-reference their judgment in the main file with any data I reported about them in the metadata file. Information I reported, for certain studies, included gender, approximate age, and region of origin. This balance of information on the one hand, with some opacity about speakers' individual identities on the other, is what I found to be both fair and optimal from the perspective of protecting speakers' privacy, but each situation will be different.

Finally, as I mentioned in section 4.1, preparing data for online accessibility takes resources: I was using Excel files as a starting point for my own private use, and the gap between that and what was posted online is not insignificant: I needed to provide morpheme-by-morpheme glosses and definitions of abbreviations and tags, all of which is work-intensive. Early in my time on

the tenure track at Stanford, I was fortunate to have a grant from Hellman Foundation to fund my work on Uzbek. I used some of this funding to train an undergraduate research assistant, Allison Dods, and later to train a native speaker linguist, Sharifa Djurabaeva, to assist me in preparing the files for public use. Without this help it is unlikely that I would have had the time I needed to undertake the endeavor on my own. This speaks volumes, I think, about how far even a relatively small amount of seed funding can go in helping archiving efforts along.

5 Conclusion

I learned many lessons from my attempts to make data I have collected via fieldwork publicly accessible, and with each project I become more careful about how the data are structured and how thoroughly and transparently things are organized. A lesson to be taken from these experiences is that especially when it comes to under-investigated languages, some effort can go quite a long way, even if that effort is coming from someone who lacks explicit training in language documentation and archiving.

Notes

1. As Uzbek is a spoken language, I will use the term *speaker* here in place of the more inclusive *user*.

2. This approach the investigation of a linguistic system, with due attention to the whole system rather than to isolated subparts of the system, reflects my commitment to what Sandy Chung has called "whole-language description."

3. Analogous strategies had already been applied successfully, by this point, for typologically similar languages such as Japanese (Kizu 1997) and Turkish (Merchant 1998). But the alternative view—in which some sort of exceptional WH-movement is forced in languages that are typically WH-in-situ in order to give rise to a genuine sluicing construction—is well attested as well (Ince 2006, 2012; Takahashi 1994), and so the Uzbek discussion was meant to weigh in on a matter of some debate.

4. This having been my first attempt at such an approach, I was less systematic in my tagging than I probably ought to have been. Later projects, including one on predicate formation strategies in the language—http://purl.stanford.edu/bq499mh5981—used a more thorough system for tagging each token for various properties that I thought would be important.

5. I obtain oral permission from speakers for these materials to be posted online. Individual speaker names/identities are never revealed in the archived materials. Relatedly, I ask speakers whether they want to be acknowledged by name in papers I ultimately publish.

6. http://purl.stanford.edu/bq499mh5981.

7. https://library.stanford.edu/research/stanford-digital-repository.

8. https://www.linguisticsociety.org/content/finding-archive-your-endangered-language-research-data.

9. Uzbekistan has implemented a Latinate alphabet gradually since it became an independent Republic in 1991, so it seemed to me that using the Latinate system would make the data accessible both to the majority of Uzbek speakers and to the scholarly community at large.

References

Berez-Kroeker, Andrea, Lauren Gawne, Susan Smythe Kung, Barbara Kelly, Tyler Heston, Gary Holton, Peter Pulsifer, et al. 2019. Reproducible research in linguistics: A position statement on data citation and attribution in our field. *Linguistics* 56 (1): 1–18.

Bodrogligeti, András J. E. 2003. *An Academic Reference Grammar of Modern Literary Uzbek*. LINCOM Studies in Asian Linguistics 50. Munich: LINCOM.

Bowern, Claire. 2008. *Linguistic Fieldwork: A Practical Guide*. London: Palgrave MacMillan.

Donovan, Michael, and Shakhlo Nematova. 2019. What counts as second-to-last? The case of the Uzbek question particle. Presented at the 93rd Annual Meeting of the LSA, New York, January 3–6.

Gribanova, Vera. 2013. Copular clauses, clefts, and putative sluicing in Uzbek. *Language* 89 (4): 830–882.

Gundel, Jeanette. 1977. Where do cleft sentences come from? *Language* 53 (3): 53–59.

Ince, Atakan. 2006. Pseudo-sluicing in Turkish. In *University of Maryland Working Papers in Linguistics 14*, ed. Nina Kazanina, Utako Minai, Philip J. Monahan, and Heather L. Taylor, 111–126. College Park: University of Maryland, Department of Linguistics.

Ince, Atakan. 2012. Sluicing in Turkish. In *Sluicing: Cross-linguistic Perspectives*, ed. Andrew Simpson and Jason Merchant, 248–269. Oxford: Oxford University Press.

Jespersen, Otto. 1927. *A Modern English Grammar*, vol. 3. London: Allen and Unwin.

Jespersen, Otto. 1937. *Analytic Syntax*. London: Allen and Unwin.

Kizu, Mika. 1997. Sluicing in *Wh-in-situ* languages. In *Proceedings of the Chicago Linguistic Society 33*, ed. Kora Singer, Randall Eggert, and Gregg Anderson, 231–244. Chicago: Chicago Linguistic Society.

Kononov, Andrei Nikolaevič. 1960. *Grammatika sovremennogo Uzbekskogo literaturnogo jazyka*. Moscow: Akademija Nauk SSSR, Institut Vostokovedenija.

Linzen, Tal, and Yohei Oseki. 2018. The reliability of acceptability judgments across languages. *Glossa* 3 (1): 1–25.

Merchant, Jason. 1998. "Pseudosluicing": Elliptical clefts in Japanese and English. In *ZAS Working Papers in Linguistics 10*, ed. Artemis Alexiadou, Nanna Fuhrhop, Paul Law, and Ursula Kleinhenz, 88–112. Berlin: Zentrum für Allgemeine Sprachwissenschaft.

Merchant, Jason. 2001. *The Syntax of Silence: Sluicing, Islands and the Theory of Ellipsis*. Oxford: Oxford University Press.

Phillips, Colin. 2009. Should we impeach armchair linguists? *Japanese/Korean Linguistics* 17:49–64.

Pinkham, Jessie, and Jorge Hankamer. 1975. Deep and shallow clefts. In *Papers from the Eleventh Regional Meeting of the Chicago Linguistic Society*, ed. R. E. Grossman, L. J. San, and T. J. Vance, 429–450. Chicago: Chicago Linguistic Society.

Ross, John Robert. 1969. Guess who? In *Papers from the 5th Regional Meeting of the Chicago Linguistic Society*, ed. Robert Binnick, Alice Davidson, Georgia M. Green, and Jerry L. Morgan, 252–286. Chicago: Chicago Linguistic Society.

Sjoberg, Andrée. 1963. *Uzbek Structural Grammar*. Uralic and Altaic Series 18. Bloomington: Indiana University Press.

Sprouse, Jon, Carson Schütze, and Diego Almeida. 2013. A comparison of informal and formal acceptability judgments using a random sample from *Linguistic Inquiry* 2001–2010. *Lingua* 134:219–248.

Takahashi, Daiko. 1994. Sluicing in Japanese. *Journal of East Asian Linguistics* 3:265–300.

Turaeva, Rano. 2015. Linguistic ambiguities of Uzbek and classification of Uzbek dialects. *Anthropos: International Review of Anthropology and Linguistics* 110:463–476.

Vaux, Bert, Justin Cooper, and Emily Tucker. 2007. *Linguistic Field Methods*. Eugene, OR: Wipf and Stock.

This is a section of doi:10.7551/mitpress/12200.001.0001

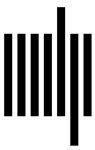
The Open Handbook of Linguistic Data Management

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 open access edition of this book was made possible by generous funding and support from the authors



The MIT Press

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