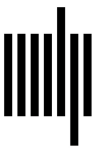
This is a section of doi:10.7551/mitpress/12200.001.0001

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 MIT Press

33 Managing Oral and Written Data from an ESL Corpus from Canadian Secondary School Students in a Compulsory, School-Based ESL Program

Philippa Bell, Laura Collins, and Emma Marsden

1 Introduction

Compulsory schooling throughout the world frequently includes the studying of a/some language(s) other than the language of instruction. In some contexts, students may be learning academic content through the new language, in conditions that provide a substantial number of hours of exposure to the language (e.g., French immersion in English-speaking Canada; English Content Language and Integrated Learning in mainland Europe). More common, however, is for the study of second language (L2)/foreign language in school settings to occur in a gradual fashion—students receive a relatively small number of hours of instruction per week over many years (Collins & Muñoz 2016). Due to the small number of hours available for learning, and thus limited input, structuring lessons to optimize learning is essential. This remains a challenging objective, though, in the absence of a clear sense of what students are able to learn and how this learning develops throughout the years of limited exposure to the target language through compulsory schooling. The data detailed here come from a corpus created to address this research objective.

In this chapter, we provide a detailed account of the creation of oral (approximately 44,000 words) and written (approximately 31,000 words) corpora and associated metadata from 230 first-language (L1) French students with no other languages aged twelve to seventeen years studying in the Quebec regular English as a Second Language (ESL) compulsory program. Although Canada is officially a bilingual French-English country, French is the only official language of Quebec. For many students, especially outside the Montreal area, the only opportunity to interact in the language, and thus learn it, comes from their school program. Any additional exposure afforded by multimedia and social interaction is done of their own individual initiative, and thus, this is similar to many foreign-language contexts throughout the world. The following account is divided into five sections documenting the necessary first steps prior to data collection (section 2), data collection (section 3), data processing and metadata tagging to create the corpora (section 4), data storage (section 5), and data sharing (section 6).

2 Organizing data collection

The organization of data collection is a key step as it is at this juncture that decisions are taken that impact all subsequent steps. For example, decisions on types and the number of tests to include affect data collection and processing in terms of their duration, personnel requirements, and cost. In addition to methodological decisions, certain practical considerations also need to be considered, which also affect subsequent steps. It is for this reason that what follows details our organizational process comprehensively. We begin by briefly summarizing our piloting, which we then refer back to frequently, given that it affected all steps in our organizational process. The remaining topics in this section describe further aspects of our procedure—obtaining ethics approval, accessing participants, securing informed consent, identifying equipment needs, and training research assistants. It ends with a description of the eight measures we prepared for data collection.

2.1 Piloting

With the exception of obtaining ethics approval, all decisions were finalized following extensive piloting, which is an essential part of data planning in the data life cycle (Mattern, chapter 5, this volume). It takes time and grant money, but its importance cannot be overstated for

creating valid data sets. Our piloting, which is detailed in Bell, Collins, and Marsden (2020), was conducted in five classrooms with 145 students. Data were collected from primary school students aged ten and eleven in grades five and six (the final two years of primary) and secondary school students aged fourteen and fifteen in the final three years of secondary. Piloting informed a range of procedural and analysis decisions detailed in the coming sections. It also allowed us to make the final selection of the elicitation tasks in terms of (1) the suitability of the topics for students of different ages, (2) the effectiveness for eliciting students' full linguistic repertoire regardless of proficiency level, and (3) the wording and language (French L1 or English L2) needed for clear instructions. It further guided our decisions on the most appropriate grade levels to target: the limited output elicited from the grade five and six primary students due to their very low proficiency in English demonstrated to us that our corpora should focus on secondary school students only.

2.2 Ethics approval

This section refers to the necessary administrative steps for ethics approval rather than questions relating to ethical use of participants' data (see Holton, Leonard, & Pulsifer, chapter 4, this volume for information on ethical use of data). To conduct research with humans in Canada, it is necessary to obtain ethics approval, which is accorded by the university at which the principal investigator (main grant holder) works. Obtaining approval requires the completion of an extensive form that ensures the research does not contravene any laws and that the participants will not be negatively affected by the research. Completion of the form requires a thorough understanding of how data collection will be conducted, even though pilot testing may change some of the original plans. It is standard practice to provide drafts or samples of all documents that will be used with the participants including consent forms, which are closely evaluated. In Quebec, students over sixteen are allowed to sign alongside a parent so we had one consent form for students aged fifteen and younger, and one for students sixteen and older. It is important to bear in mind that obtaining approval can take some time (approximately two months during term time in our context, longer if the process is undertaken during the summer break).

2.3 Access to participants (collection sites)

Accessing participants to contribute data can be time-consuming and extremely challenging within school settings. Schools in Quebec are governed by school boards that may have their own ethics approval process that must be completed before gaining access to schools and classes. Administrative steps to gaining access vary by school, school board, age of participants, to name a few, and as such, the protocol to be followed varies considerably. We gained access through approaching ESL pedagogical advisors who work with ESL teachers at different school boards. Initially, we wrote an information letter detailing the goals of the research and specific information on what students and their teachers would need to do. The advisors sent this document either to teachers they believed would be interested or to all teachers with whom they worked. The teachers then contacted the researchers directly.

The time interval from sending out the letters to pedagogical advisors to starting data collection in a school varied from two weeks to two months, with shorter delays usually being associated with more experienced teachers as they knew whom to ask for permission, if needed, and were organized in terms of distributing and collecting consent forms. This process also included a lot of communication with several teachers who ultimately could not participate for different reasons. We needed to control for L1 background, for example, and some classes had too many non-native speakers of French for our purposes. Some teachers realized they could not accommodate our needs. This last point is important to bear in mind for this type of data collection as our time needs (approximately one hour with all students and one to two hours in a setting where a small number of students could be seen individually for approximately ten minutes) were difficult to meet in contexts where teachers only worked with a class for as little as sixty minutes per week.

2.4 Informed consent

As soon as data collection was organized with a teacher, consent forms were sent to the school. In some situations, it may be necessary for the research team to present the forms to the students in person. When working with children under seventeen years of age in Quebec, consent needs to be provided by their parents/guardians,

although those sixteen years and older can also provide contingent consent. As such, for these participants, we provided two versions of the consent form—one which only asked for the parents'/guardians' signature and one which included a space for the students to sign. Obtaining consent could take over a week if teachers only saw their students once a week. To increase the likelihood of parents providing consent, we offered to send them information on the results of the research project two years after its completion, which approximately two-thirds requested. We also divided consent into two sections—participation in the current project and consent for data to be used for future research projects. This was done to ensure we had explicit consent to use the data for other projects, as well as to provide parents with an option if they were uncomfortable not knowing how the data would be used in the future. In this study, no participant/parent opted out of data being used in the future. In total, consent was not provided by 81 students/parents of 354 that received the forms.

2.5 Data collection equipment

Successful oral and written data collection requires appropriate recording equipment, and piloting was an excellent means of understanding true equipment needs. When collecting oral data, audio recorders are the main piece of equipment. Piloting allowed us to ascertain that our audio devices were suitable, and that lapel microphones were undesirable as they led to lost data when students played with them. Nor did they notably improve the facility with which we could transcribe the oral data. Our audio devices were purchased on Amazon.ca at a cost of either C$26.99 or C$29.99. The two models were both EVISTR mini voice recorders (L157 and L169) with 8 GB of storage equivalent to 560 hours of taped audio. Two models were purchased due to an insufficient quantity of one model being available. The audio files were saved in .wav format.

For collecting written data, no specialized equipment was needed. Some researchers may wish to collect typed texts to reduce transcription time (Gilquin 2015), but this was not feasible in our context. As we needed to collect written data from all students in a class simultaneously due to the limited number of ESL class hours, this would have meant the provision of up to thirty-four computers at one time as schools cannot guarantee access to computers/laboratories. This was beyond our budget and would not have been a useful long-term purchase. In addition, as children are still more accustomed to handwriting than typing in class, we did not want children's typing skills to become a confound in our data, as the nature of their productions (such as deviations from the norm) could either be attributable to typing skills or a true reflection of their written language.

One of the proficiency measures, the elicited imitation (Ortega 2003), required sentences to be read aloud, which were then repeated and recorded by each student. To this end, we needed a laptop and speakers. Piloting allowed us to ascertain correct volume levels and optimal speaker placement within a classroom. Despite our efforts to ensure that our intended procedure would be appropriate, the majority of these data were ultimately unanalyzable due to noise issues—our piloting classroom for this test only was an extremely quiet grade five primary classroom, which in the end did not adequately reflect the noise challenges in secondary school classrooms.

Piloting also demonstrated the importance of having clear information on when and where each data collection would occur to ensure that all equipment (audio recorders, speakers, laptop, paper copies of tests, and so on) was in the right place at the right time. As different research assistants collected data on different days, this information, which we presented in an Excel sheet, needed to be shared with all assistants and updated regularly. We achieved this by using Dropbox. In this document, we included explicit instructions on where equipment would be taken after one data collection and how this equipment would be transferred to another research team for the next data collection.

2.6 Training research assistants

To ensure all tasks were completed in the allotted time and to ensure all students received the same instructions and had the same interactions with the different assistants, extensive training was required. Without such training, there is a risk of data collection lacking consistent standards (poor reliability) across collection sites and times. Training also minimizes the likelihood of tasks not being completed or being administered incorrectly.

The need to adjust and expand training became evident during pilot testing. For the oral data collection,

students worked individually with an assistant. Piloting demonstrated that the interactions between students and different assistants (seven in total) varied greatly. Some assistants provided much more help than others, which meant, for example, certain language forms may have been primed (McDonough & Trofimovich 2008) by the assistants for some students, but not for others.

Following pilot testing, we created a checklist and a script for both oral and written data collection. The checklist allowed assistants to verify they had all the materials and equipment needed for data collection. We also used the oral checklist for assistants to write down the number of each audio file alongside each student's name. This had the added benefit of reminding assistants to stop and start the file after each oral measure and between students. Use of these protocol documents ensured that instructions and order of tests were identical across all participants, which helped to ensure that any influence that undertaking one test had on undertaking another was experienced equally across all participants.

2.7 Final data collection materials

Four tasks were used to collect data for the corpus. Four other measures were used to provide further information on the participants (metadata). The process we undertook before finalizing our data collection materials, including task selection, has been fully documented in a previous publication (Bell, Collins, & Marsden 2020) whose goals included methodological transparency to reduce research biases and to improve overall data collection (Gawne & Styles, chapter 2, this volume; Marsden 2019). Here, we present an overview of the different measures.

2.7.1 Written argumentative task Students were given twenty minutes to respond to one of two questions, which were both yes/no questions. Students chose the question to which they responded and they were asked to provide three reasons to justify their response.

1. Should students be allowed to use their cell phones in school?
2. Do aliens exist?

2.7.2 Written narrative task Students were given a time limit of twenty minutes to write a story based on an image of two police officers at the house of a young boy and his mother. The instructions asked them to look at the image, to imagine what has happened, what is happening now, and what will happen. This measure was adapted from a measure used in published research with Quebec ESL students aged ten to twelve (Collins et al. 1999; Collins & White 2011).

2.7.3 Oral argumentative task Students were given a handout containing images and the following sentences, each of which appeared next to the relevant image.

> Alex is studying for his math exam tomorrow. Emma cannot study because her parents are out having dinner. She is babysitting her baby brother. Emma copies her friend Alex's answers. Alex sees her.

Students were asked to provide oral responses to each of the following questions:

If you were Emma, would you have cheated? Why?

If you were Alex, what would you do? Why?

No time limit was provided.

2.7.4 Oral narrative task Students were given eleven images from the book *Frog, Where Are You?* (Meyer 1969), which has been used as an instrument in other studies (for examples and some elicited data, see https://www.iris-database.org/iris/app/home/search?query=frog). Students were asked to recount the story. No time limit was provided.

2.7.5 Student questionnaire To help interpret the data, we obtained biographical information from each participant in a questionnaire containing seventeen questions that could be completed in under ten minutes. In our context and in line with our research objectives, it was important to collect data from francophone students who did not have any additional languages other than L2 English learned in the Quebec compulsory education program. The questionnaire included questions on home languages, schools where they had previously studied, study outside Canada, and exposure to all languages. It is also here that one can ask students to evaluate their proficiency in all their languages. As has been argued elsewhere, proficiency in the language of the corpus should be measured as objectively as possible (Bell & Payant 2020; Thewissen 2013), but self-evaluation for other languages is probably sufficiently fit-for-purpose and can be collected quickly and may help explain counterintuitive results about use of the target L2 (Sinclair 2005).

The questionnaire responses identified twenty-one students who had too much experience with English

to be included in the study—English-speaking parents or close family or previous study at an English-medium school or in a special program in which exposure to English is far more frequent than in the regular, compulsory program. For those students who identified a parent or family member as speaking English, but who also wrote they only spoke French, we assessed their written and oral texts to see whether the student appeared to be more proficient in English than would be expected.

Eighteen students identified a language other than English or French as being the main language used at home. These students also identified their proficiency in this language as being the same or better than their French proficiency. Their data were thus excluded.

We chose to write our questionnaire in French to ensure that students, regardless of English proficiency, could understand. Nevertheless, it is still important to pilot the questionnaire to ensure the questions are written clearly and that they generate the intended information for the research project. In our prepilot questionnaire, we referred to different Quebec ESL programs (e.g., Intensive English, core, enriched), but many students were not familiar with these terms alone. We thus provided more detailed information to ensure we gathered the required information.

2.7.6 Proficiency measures Two measures of proficiency (an elicited imitation and a yes/no vocabulary measure) were used to ensure that the oral and written texts constituting the corpora could be classified based on objective measures of proficiency. We chose to include two measures, as our preferred measure (the elicited imitation) requires students to use audio recorders, which increases the likelihood of lost data. Furthermore, as the proficiency measures had to be collected from all students simultaneously, we were concerned that the elicited imitation would be difficult to analyze due to background noise. We thus included the yes/no test, which is quick and easy to administer, and has been widely used as a general proficiency measure (Harsch & Hartig 2016).

2.7.6.1 Proficiency measure 1: Elicited imitation An elicited imitation test was selected as a general proficiency measure. This test asks participants to repeat aural sentences aloud after a time delay to encourage reconstruction of meaning, not direct imitation. The test was adapted from an extant measure that can be found on IRIS, a digital repository for second language researchers (https://www.iris-database.org/iris/app/home/detail?id=york%3a852670&ref=search; Ortega et al. 2002) and has been normed across eight languages.

2.7.6.2 Proficiency measure 2: Yes/no test The yes/no test (Meara & Buxton 1987; see Collins & White 2011, https://www.iris-database.org/iris/app/home/detail?id=york%3a934278&ref=search for a version of this test) was chosen to provide an independent means of assessing each student's proficiency (Hasko 2013). In this vocabulary test, students have to say whether they know or do not know a word. The test includes real words (two-thirds) and pseudo-words (one-third). The target words came from Meara's X-Lex test (2005), which were drawn from the 5,000 most frequent words of English in five bands reflecting 1,000 words each. Thirty words (20 real; 10 pseudo) from each of the five 1,000-word levels were included for a total of 150 words for students to judge (Harsch & Hartig 2016). The test can be marked in a number of ways. We followed Cobb's example (personal e-mail correspondence, August 11, 2017) used on Compleat Lexical Tutor (https://www.lextutor.ca). For each real word that is identified as known, the participant receives one point. For each pseudo-word identified as known, the participant loses two points.

2.7.7 Teacher questionnaire The teacher questionnaire, containing eleven open-ended questions written in English, was included to understand the teachers' teaching philosophy in terms of types of activities used in the classroom, use of French, school board teaching requirements, and beliefs on the teaching of grammar and vocabulary. It took approximately fifteen minutes to complete and was sent via e-mail to the teachers and returned at their convenience. This questionnaire was deemed important to help interpret, if applicable, any differences found between classes.

3 Data collection

The extensive pilot testing and intensive planning allowed for the data collection to be completed with only one major problem—the elicited imitation data were largely unanalyzable. In the following paragraph, we provide detailed information about minor problems that were encountered. Two key elements to facilitate data collection of the type we were conducting are (1) having sufficient

numbers of research assistants to manage the flow of tasks and students (particularly during the individual oral testing sessions), which should include an assistant whose role is to ensure student movement between classrooms and assigning students to assistants waiting for another student, and (2) budgeting sufficient amount of time to set up the equipment and materials and organize the configuration of the oral testing rooms. Given the restrictions of the students' timetables, as little as a five-minute delay in testing could mean one measure would not be given to a set of students, which in turn either leads to rescheduling (often impossible) or lost data.

Other minor problems encountered in data collection in classroom contexts merit highlighting. Public address systems are common in Quebec schools, so interruptions during in-class testing had to be managed. This was of importance during the administration of the elicited imitation as all students were closely listening to audio and then repeating. The research assistants were told to be next to the laptop so that if the address system was used, the audio recording could be stopped immediately. On occasion, the regular teacher was replaced by a substitute who was not aware of our research so it was important that all research assistants felt comfortable providing a brief summary of the research and its goals. Some teachers did not realize that they would have to teach during the oral data collection class, which demonstrated the need to better prepare teachers for the procedures in any future data collections. As we were speaking individually with students, teachers needed to teach their normal class while letting four or five students out at a time to meet with the research assistants. Finally, it is useful to let other teachers and administrative staff be aware of the testing to reduce the likelihood of interruptions, which happened during oral testing when curious teachers poked their heads round the door of the classroom in which we were working to see what was happening.

4 Data processing

After data collection, the data must be processed to create the corpus (for more information on transforming data, see Han, chapter 6, this volume). Concretely, this process includes transcription, verification, and annotation, although it is also at this time that format standards and file naming (key practices for responsible and consistent data management; see Mattern, chapter 5, this volume) will occur (discussed in section 5 on data storage). Transcription requires the written and oral texts be rendered into a chosen format based on a set of norms. We chose to transcribe using the Codes for the Human Analysis of Transcripts (CHAT; (MacWhinney 2000), which is the transcription system of the TalkBank system (MacWhinney 2007). CHAT requires the integration of texts and tags (both in-line tags and independent tagging lines) rather than the separation of the original text and any annotation (XML format). These two differing approaches affect transcription with corpus linguists recommending separation (Sinclair 2005), while researchers using TalkBank must use the integrated approach. Our choice to use CHAT was based on recommendations in the field of second-language acquisition (Myles 2005; MacWhinney 2017a), the first author's knowledge of this transcription system, and the possibilities for automatic analyses using natural language processing steps within TalkBank.

Data processing is enormously time- and labor-intensive, which is one important argument for the field to develop a collaborative ethic and, for example, share instruments through IRIS (Marsden, Mackey, & Plonsky 2016) and share their corpora (despite valid concerns in areas such as ethical data use; see Holton, Leonard, & Pulsifer, chapter 4, this volume) and corpora representativeness for future users (Sinclair 2005). Processing oral data is particularly arduous with many transcription decisions needing to be made (Cottier, Wlodarski, & Bell 2019). However, even with written data, issues related to the interpretation of handwriting exist. It has been mentioned that written corpora are now often already word processed (Gilquin 2015), but this is likely more realistic when collecting data from adults. In our context, the only means of collecting word processed documents would have been through our providing individual laptops for all students, and it would have introduced a confound in the data.

Our transcription conventions were taken from the CHAT manual (MacWhinney 2000), which allowed us to copy the students' texts as written. However, as CHAT requires words to be spelled correctly for them to be recognized and analyzed by the Computerized Language Analysis program (MacWhinney 2017b), it was necessary to change incorrect spellings, which then required in-line annotation to ensure later analysis could be conducted transparently (i.e., it was possible to know

whether a participant's word had been spelled incorrectly in their original production). It is perhaps worth noting that CHAT was originally created for oral language transcription and for L1 participants, thus orthographic representation of (non-native) sounds was less of an issue during the development phase of CHAT.

After transcription, all texts were verified by another person to ensure the reliability of transcription. Even though this step takes longer for oral data, it is also vital for written data as it is at this juncture that mistakes can be found.

5 Data storage

As discussed by Mattern (chapter 5, this volume), responsible and consistent data management practices are vital if corpora are to be considered reliable. First, we present the file-naming system employed. Then, how all the files are stored will be discussed.

We used an Excel sheet as the master list, although other open use programs or accessible formats (e.g., .csv) could be used. In accordance with the ethics protocol we followed, this sheet is the only document in which participants are fully identified. It is only accessible to the main researcher and the head research assistant whose access will end once the sheet is finalized and any data cleaning or analysis requiring names is complete.

The file-naming system was created prior to data processing to ensure consistency across files from the data set. Data from piloting were also included in this system, which explains the inclusion of files from two years of primary school students. This allows us to include the pilot data for certain analyses, although these data cannot be made available to the wider research community as ethics approval for the piloting was restricted to use by the research team only. First, a four-digit number was employed to identify the grade level: 1000: primary grade five; 2000: primary grade six; 3000: secondary grade one; 4000: secondary grade two; 5000: secondary grade three; 6000: secondary grade four; 7000: secondary grade five.

Participants within each grade level were then identified using a number, which was integrated into the four-digit grade level number. For example, the student identified as number 1 in secondary grade one was given the code 3001, student 2 was 3002. The data within each grade level are not identified based on class (students at each grade level came from multiple classes) as the corpora focuses on individuals in terms of grade level rather than in terms of which class they were in (this would be useful/necessary, for example, if the corpora were designed to address instructional practices). The head researcher has access to this information, but there was no reason to provide this a priori to respond to the main research objectives. As little information was collected in terms of potential class differences (apart from teachers' brief self-report about their teaching philosophy in the survey), it is unlikely the corpora will be analyzed based on class in the future.

Each measure was given a code: oral argumentative (OA), oral narrative (ON), written argumentative (WA), written narrative (WN), elicited imitation (EI), student questionnaire (Q), and teacher questionnaire (TQ).

Separate files (CHAT format) for each student were created for the four tasks that formed the corpus (OA, ON, WA, WN). Each file integrated the grade level, the student identification number, and the task. For example, 3061.OA referred to the oral argumentative text from a secondary grade one student whose identification number was 61.

As CHAT files had to be verified, we added .v to verified files. Thus, the final version for analysis would be 3061.OA.v. In hindsight, it may have been more sensible to label unverified files with a longer file name, which could then be deleted after verification (e.g., 3061.OA.u where the u stands for unverified).

After the verified files had been automatically analyzed for parts of speech, the file name had .M added to it to indicate that it had been analyzed using the MOR program in the CHILDES suite of programs.

Storage includes paper and electronic data, and many of the decisions involve following established ethics protocols. Paper copies of all the measures have been scanned to create electronic copies. The hard copies, including consent forms, are in a locked filing cabinet in the head researcher's office, as they have the participants' names on them. Electronic files are anonymous aside from one Excel sheet that acts as a master copy, which also includes the student questionnaire data and scores on the two proficiency measures. This file is only available to the head researcher and the head research assistant. Electronic files are in a variety of formats—the four transcribed oral and written texts are in .cha, the electronic copies of hard copies are in PDF, the audio files are in .wav. All electronic files are stored on two external

hard drives and in the Dropbox of the head researcher. Assistants working on transcription/verification also have access through Dropbox to the files they need.

6 Data sharing

Gawne and Styles (chapter 2, this volume) discuss the centrality of openness in terms of methodological transparency and the importance of making data accessible to other researchers aside from those initially involved. Indeed, second language researchers now have a digital repository, IRIS (https://www.iris-database.org), in which instruments, materials, and data are shared and made highly searchable due to its fine-grained, domain-specific metadata (Marsden, Mackey, & Plonsky 2016). This collaborative effort has been an important step in advancing research practice and methodology.

In terms of data sharing, specifically, it is important to also share the detailed methodological steps undertaken in creating the corpus. In other words, data sharing should only occur if sufficient information is provided for researchers to understand the data in-depth (Sinclair 2005). Researchers can often contact the original researchers with queries, but this requires the secondary researchers to pose specific questions and the original researchers to be available. In reality, many methodological decisions that could affect the data will not be considered in any single study. For example, a secondary researcher may not consider whether participants contributing to the corpora were allowed to use outside resources such as dictionaries/spell-check. However, the inclusion/exclusion of external resources during the production of texts that make up the corpora should be taken into account during data analysis and interpretation to meet certain research objectives.

The data from the corpora are not finalized so are not yet ready for sharing. As previously discussed, the corpus files are in CHAT format and thus, the TalkBank regulations regarding sharing will be used. This does mean the audio files will need to be manually cut to delete the initial interactions that identify the school and the participant by name and to add a participant identification number. Even though the addition of this number to the audio file itself could be seen as superfluous (as the name of the audio file links to the relevant transcript), it is deemed necessary in case file labels are changed inadvertently, resulting in the loss of the participant's identification number. This process will be conducted by a research assistant using an open-source program such as Audacity.

7 Conclusion

The goal of this chapter was to present pertinent information on the creation and storage of a corpus of oral and written texts from learners of ESL in compulsory schooling in Quebec, Canada. The different steps undertaken by the research team have been described to help promote understanding of corpus building and maintenance in general and with respect to contextual factors. Noteworthy for our project were the decisions and procedures that were dictated by working with intact classes in schools and child participants. Certain issues become more or less important depending on the context, and we believe our discussion highlights a number of decisions that must be taken to ensure the validity, reliability, sustainability, and usefulness of these types of data sets.

References

Bell, P., L. Collins, and E. Marsden. 2020. Building an oral and written learner corpus of a school programme: Methodological issues. In *Learner Corpus Research and Second Language Acquisition*, ed. B LeBruyn and M. Paquot, 214–242. Cambridge: Cambridge University Press.

Bell, P., and C. Payant. 2020. Designing Learner Corpora: Collection, Transcription, and Annotation. In *The Routledge Handbook of Second Language Acquisition and Corpora*, ed. N. Tracy-Ventura and M. Paquot. New York: Routledge.

Cobb, Thomas. n.d. Compleat Lexical Tutor v.8.3. https://lextutor.ca. Accessed January 29, 2019.

Collins, L., R. H. Halter, P. M. Lightbown, and N. Spada. 1999. Time and the distribution of time in L2 instruction. *TESOL Quarterly* 33 (4): 655–680. https://doi.org/10.2307/3587881.

Collins, L., and C. Muñoz. 2016. The foreign language classroom: Current perspectives and future considerations. *The Modern Language Journal* 100 (1): 133–147. https://doi.org/10.1111/modl.12305.

Collins, L., and J. White. 2011. An intensive look at intensity and language learning. *TESOL Quarterly* 45 (1): 106–133.

Cottier, D., N. Wlodarski, and P. Bell. 2019. The impact of methodological decisions in transcribing written data on research findings. Paper presented at the American Association of Applied Linguistics Conference, Atlanta, March 9–12.

Gilquin, G. 2015. From design to collection of learner corpora. In *The Cambridge Handbook of Learner Corpus Research*, ed. S. Granger, G. Gilquin, and F. Meunier, 9–34. Cambridge: Cambridge University Press. https://doi.org/10.1017/CBO9781139649414.002.

Harsch, C., and J. Hartig. 2016. Comparing C-tests and yes/no vocabulary size tests as predictors of receptive language skills. *Language Testing* 33 (4): 555–575.

Hasko, V. 2013. Capturing the dynamics of second language development via learner corpus research: A very long engagement. *The Modern Language Journal* 97 (S1): 1–10. https://doi.org/10.1111/j.1540-4781.2012.01425.x.

MacWhinney, B. 2000. *The CHILDES Project: Tools for Analyzing Talk*. 3rd ed. Mahwah, NJ: Lawrence Erlbaum Associates.

MacWhinney, B. 2007. The TalkBank Project. In *Creating and Digitizing Language Corpora: Synchronic Databases,* vol. 1, ed. J. C. Beal, K. P. Corrigan, and H. L. Moisl, 163–180. Houndmills: Palgrave-Macmillan.

MacWhinney, B. 2017a. A shared platform for studying second language acquisition. *Language Learning* 67 (S1): 254–275.

MacWhinney, B. 2017b. Tools for analyzing talk, part 2: The CLAN Program. https://childes.talkbank.org/. PDF document downloaded May 2, 2017.

Marsden, E. 2019. Methodological transparency and its consequences for the scope and quality of research. In *Routledge Handbook of Research Methods in Applied Linguistics*, ed. J. McKinley and H. Rose, 15–28. New York: Routledge.

Marsden, E., A. Mackey, and L. Plonsky. 2016. The IRIS Repository: Advancing research practice and methodology. In *Advancing Methodology and Practice: The IRIS Repository of Instruments for Research into Second Languages,* ed. A. Mackey and E. Marsden, 1–21. New York: Routledge.

McDonough, K., and P. Trofimovich. 2008. *Using Priming Methods in Second Language Research.* New York: Routledge.

Meara, P. M. 2005. *X_Lex: The Swansea Vocabulary Levels Test.* Version 2.05. Swansea, UK: Lognostics.

Meara, P., and B. Buxton. 1987. An alternative to multiple choice vocabulary tests. *Language Testing* 4 (2): 142–154.

Meyer, M. 1969. *Frog, Where Are You?* New York: Dial Press.

Myles, F. 2005. Review article: Interlanguage corpora and second language acquisition research. *Second Language Research* 21 (4): 373–391.

Ortega, L. 2003. Syntactic complexity measures and their relationship to L2 proficiency: A research synthesis of college-level L2 writing. *Applied Linguistics* 24 (4): 492–518.

Ortega, L., N. Iwashita, J. M. Norris, and S. Rabie. 2002. An investigation of elicited imitation tasks in crosslinguistic SLA research. Paper presented at the Second Language Research Forum, Toronto, October 3–6.

Sinclair, J. 2005. Corpus and text: Basic principles. In *Developing Linguistic Corpora: A Guide to Good Practice*, ed. M. Wynne, 1–16. Oxford: Oxbow.

Thewissen, J. 2013. Capturing L2 accuracy developmental patterns: Insights from an error-tagged EFL learner corpus. *The Modern Language Journal* 97 (S1): 77–101.

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