12

TalkBank for SLA

Brian MacWhinney

Introduction

Written language learner corpora are fairly easy to obtain. In comparison, corpora providing data on learners' spoken language usage are much less available. A major exception is the TalkBank system, which provides online multimedia data for 14 types of spoken language data. Of these 14 data banks, the three that are most directly relevant to studies of second language learning are SLABank at https://slabank.talkbank.org, which includes data from second language learners, BilingBank at https://biling.talkbank.org, which includes data from bilinguals, and the BiLing segment of CHILDES at https://childes.talkbank.org which includes data from children learning two or more languages. In addition, the methods being developed within the context of the FluencyBank project (https://fluency.talkbank.org) are important for analysis of the tradeoffs and interactions between complexity, accuracy, lexis, and fluency (CALF) (Wen & Ahmadian, in press) in L2 written and oral productions. The materials of greatest relevance to learner corpus research (LCR) and second language acquisition (SLA) are the tagged transcripts and recordings that are all freely downloadable from and immediately browsable at https://slabank.talkbank.org. This chapter reviews these resources and the framework they provide for standardization and integration in and between LCR and SLA. Although our focus here is on SLABank and BilingBank, it is important to understand the overall shape of TalkBank and how all 14 components work together to share web resources, programs, standards, and funding. Moreover, we can learn from the successes in the other banks how best to advance the growth and usage of SLABank and BilingBank.

Let us consider the relative advantages and disadvantages of written vs. spoken language data for the study of SLA. In comparison with spoken data, data from learners' written samples are much easier to collect in large quantities across languages and groups of learners. For example, the EFCamDat database (Geertzen, Alexopoulou, & Korhonen, 2013) contains over 83 million words from a million assignments written by 174,000 learners at various levels and speaking a wide variety of first languages (L1s). It is possible to apply automatic methods to tag this corpus for parts of speech and grammatical relations, although errors must be tagged by hand. This tagging then further facilitates a wide variety of methods for LCR analysis. Another advantage of written corpora is that it is easy to ensure full anonymization, as long as the writers do not insert identifying material in their essays.

In contrast, obtaining, transcribing, and analyzing spoken language data is much more difficult. Often, audio or video recordings must be made in person one at a time, and full consent must

be obtained for use of either anonymized or non-anonymized data. Accurate transcription of spoken language data can take up to 20 hours for one hour of recording. To analyze fluency patterns, the transcripts must also be linked to the audio at least on the level of the utterance and preferably on the level of individual words. Together, these processes of elicitation, recording, transcription, and linkage involve a much greater investment of time and effort than for written data.

However, there are also important advantages to studying spoken language data. In many classrooms, learners and instructors view the attainment of oral fluency as the most important goal in L2 learning. From the viewpoint of SLA theory, oral production is "a better window into implicit knowledge" (Myles, 2015, p. 314). From the viewpoint of functional and cognitive linguistics, oral performance is understood as the central determinant of language structure and function (MacWhinney, 2014). From the viewpoint of psycholinguistics and neurolinguistics, spoken language usage provides the most direct measure of language processing and functioning (Kemmerer, 2015; Magnuson, Dixon, Tanenhaus, & Aslin, 2007). From the viewpoint of Conversation Analysis (Goodwin, 2003), spoken interactions are fundamental to understanding social interactions and practices.

Furthermore, spoken language data has access to a variety of communicative channels that are not present in written language. These include a wide variety of variations in speed, loudness, breathiness, and other vocal qualities that signal emotional and attitudinal dimensions in the spoken signal. Spoken utterances are often interlaced with "thetical" (Kaltenboeck & Heine, 2015) comments such as "you know" or "I would say" that provide metacommentary on the basic sentential content. Face-to-face interactions are also accompanied by a wide variety of gestures (Kendon, 1982), facial expressions (Ekman & Friesen, 1978), and proxemics (Hall, 1966) that encode emotional, relational, perspectival, and attitudinal information that is not expressed in written communication. Because of this greater complexity, some researchers working with learners' spoken language data find it best to focus their attention on small segments of production in which the interaction between all of these channels and constraints can be examined closely (Eskildsen, 2012). However, it is then necessary to examine the generality of such patterns, and this can only be done in the context of larger, openly available, spoken language corpora.

Components of TalkBank

TalkBank provides data collected during spoken language interactions. The TalkBank system (https://talkbank.org) is the world's largest open access integrated repository for spoken language data, providing language corpora and resources in 14 areas to support researchers in Psychology, Linguistics, Education, Computer Science, and Speech Pathology. The National Institutes of Health (NIH) and the National Science Foundation (NSF) have provided support for the construction of five central components of TalkBank:

1. AphasiaBank at https://aphasia.talkbank.org for the study of language in aphasia in six languages,
2. CHILDES at https://childes.talkbank.org for the study of child language development in 42 languages from infancy to age 6,
3. FluencyBank at https://fluency.talkbank.org for the study of language fluency and disfluency in stuttering, aphasia, second language learning, and normal processing,
4. HomeBank at https://homebank.talkbank.org for the study through automatic speech recognition of untranscribed daylong recordings in the home and elsewhere, and
5. PhonBank at https://phonbank.talkbank.org for the analysis of children's phonological development in 18 languages.

The construction of BilingBank and SLABank have not yet received grant support.

Table 12.1 TalkBank Usage

	CHILDES	*SLABank*	*AphasiaBank*	*PhonBank*	*FluencyBank*	*HomeBank*	*TalkBank*
Corpus Age	30	12	10	7	1	2	14
Words (millions)	59	4.7	1.8	0.8	0.5	audio	47
Linked Media (TB)	2.8	0.15	0.4	0.7	0.3	3.5	1.1
Languages	41	13	6	18	4	2	22
Publications	7000+	67	256	480	5	7	320
Users	2950	50	390	182	50	18	930
Web hits (millions)	5.0	0.1	0.5	0.1	0.1	0.4	1.7

Table 12.1 summarizes the size, usage, and age of these databases. In that table, the contents of the other nine areas in TalkBank (ASDBank, BilingBank, CABank, ClassBank, DementiaBank, RHDBank, SamtaleBank, SLABank, and TBIBank) are given in the column labelled "TalkBank". As this table shows, the CHILDES database has supported far more publications than the other databases, but this is largely a result of the fact that so much data has been available for so many years to such a large academic community. After CHILDES, the next oldest databases are AphasiaBank and PhonBank which are supporting more and more publications over time. In comparison, there are far fewer publications arising from BilingBank and SLABank.

Core Issues and Topics

The TalkBank system is grounded on six basic principles: maximally open data-sharing, a standardized transcription format, format-compatible software, interoperability, responsivity to research group needs, and compliance with international standards.

Principle 1: Maximally Open Data-sharing

In the physical sciences, the process of data-sharing is taken as a given. Lamentably, data-sharing for spoken language data has not yet been adopted as the norm. This failure to share research results – much of it supported by public funds – represents a huge loss to science. Researchers often cite privacy concerns as reasons for not sharing data on spoken interactions. However, as illustrated at https://talkbank.org/share/irb/options.html, there are many ways in which data can be made available to other researchers, while still preserving anonymity. It is important for researchers to consider these options and to have the correct IRB approvals and consent forms in place from the beginning of a new project. Fortunately, data that have already been collected can still be included in SLABank, as long as they are properly anonymized.

In many areas of biomedical and psychological science, evidence has accumulated regarding the non-reproducibility of widely-cited findings. SLA researchers have also become increasingly concerned with the need for methods to evaluate replicability (Marsden & Plonsky, 2018). To support this effort, scientists have encouraged the development of Open Science based on open data-sharing. In their *Manifesto for Reproducible Science* Munafò et al. (2017) note that "once it is accepted as the norm, we doubt that data-sharing will ever go out of fashion." TalkBank is grounded on this same basic principle. To achieve replicability in the area of LCR (Marsden, Morgan-Short, Thompson, & Abugaber, 2018), there must be not only full data-sharing, but also consistent and transparent methods for transcription and computational analysis. Ideally, it should be possible for any researcher to verify the accuracy of a published finding based on corpus analysis of openly available data by running a series of commands that are fully described in the relevant publication.

Until publications provide this level of consistency and transparency, it is difficult to replicate published findings. This is not to say that researchers would knowingly publish misleading results. However, replication of results by a second research team provides a way of checking for errors and for illuminating areas of substantive disagreement about codes and categories of analysis.

Open-access data-sharing means that researchers should be able to obtain the complete content of a given corpus. All TalkBank corpora are accessible in this way. In contrast, there are many other corpus sites that provide only a limited form of query interface to their corpus. For example, they may only support key word in context (KWIC) searches or searches through a single CQL command. Such sites do not allow downloading of the corpus or the related media. Without providing full access to corpora, it is difficult for researchers to examine all possible patterns in corpora. In some cases, these restrictions are said to be necessary to preserve anonymity. However, anonymity can easily be preserved through other methods, such as avoidance of personal identifiers in filenames, replacement of last names with *Lastname*, replacement of identifying place names with *Placename*, and silencing of segments of the audio that align with these replacements.

Principle 2: A Standardized Transcription Format

Individual research projects are typically designed to sample from specific language contexts. The hope is that, by comparing the results of projects from a wide variety of contexts, one can formulate general principles of language learning and usage. However, if each project develops and uses its own idiosyncratic methods for transcription and analysis, comparison across projects can become difficult. To address this problem, some subfields have developed transcription standards, but these are often not compatible with those used in nearby fields. In order to provide maximum harmonization across these formats, TalkBank has created an inclusive transcription standard, called CHAT, that recognizes all the features required by the many different disciplines studying spoken language, as well as written language. The features and codes available in this system are documented in the CHAT manual which can be downloaded from https://talkbank.org/manuals/chat.pdf. CHAT can also be automatically converted to XML format through use of the Chatter program (https://talkbank.org/software/chatter.html) in accord with the schema available at https://talkbank.org/software/xsddoc/index.html. Although the overall system is complex, most aspects of CHAT are only necessary for special purposes.

CHAT provides several facilities of specific importance for studies of L2 learning.

1. **Linkage to audio**. CHAT allows for tagging in milliseconds of the beginning and ending points of words, phrases, and utterances. To facilitate reading of the transcript, these time marks can be hidden or displayed through an option in the CLAN program.
2. **Support for Conversation Analysis (CA) analysis**. CA (Schegloff, 2007) uses a wide variety of markings for intonational and conversational patterns. Originally, these markings were entered using typewriter alignment supplemented with handwritten symbols. To represent these markings in computer format, CHAT uses Unicode characters as summarized at https://ca.talkbank.org/codes.html .
3. **Phonological encoding**. CHAT supports a system for phonological analysis with IPA characters. This system is encoded in CHAT XML which is also compatible with the Phon program (https://phon.ca) for complete analysis of phonological patterns and development. Because Phon incorporates the complete codebase for Praat, all Praat analyses can be conducted inside Phon, providing further compatibility with CHAT.
4. **Error marking**. CHAT also includes a well-developed system for marking semantic, syntactic, morphological, phonological, and pragmatic errors. This system has been used most extensively for analyzing aphasia speech, but it is also relevant for the analysis of second language productions.

Principle 3: Format-compatible Software

The program used for analysis of TalkBank data is called CLAN. Because all TalkBank data are in a single, consistent format, commands in CLAN are able to maximize the use of detailed features of this format. Moreover, because everything is in a consistent format which fully matches the requirements of the software, users only need to learn one program to analyze all of the data in TalkBank, rather than having to learn a different program for each corpus that they study. The use of the CLAN program will be discussed in the section on *Main Research Methods and Tools*.

Principle 4: Interoperability

Although CLAN has been designed to cover a wide range of analysis methods and research goals, there are other software packages that are better adapted to certain specific tasks. For example, although CLAN includes most of the facilities of the popular AntConc program, the interface for AntConc is more easily controlled. As a result, for work with written corpora, users may wish to export CHAT data to AntConc. To facilitate analysis of TalkBank data in AntConc, CLAN has a program called CHAT2TXT that can remove the various CHAT features unique to spoken language, such as retracing, pauses, etc. for export to AntConc.

To support this type of interoperability, TalkBank has developed a series of programs to convert from CHAT to other formats. These programs can automatically convert both to and from the formats required for Praat (http://praat.org), Phon (http://phon.ca), ELAN (tla.mpi.nl/tools/elan), CoNLL-U (universaldependencies.org/format.html), SRT (for video captions), LIPP, ANVIL (http://anvil-software.org), EXMARaLDA (http://exmaralda.org), SALT (http://saltsoftware.com), LENA (http://lenafoundation.org), and Transcriber (http://trans.sourceforge.net) formats. Users may wish to do further analysis or coding in some of these programs. In that case, in order to guarantee convertibility back to CHAT, they must be careful not to alter codes in CHAT format that mark aspects not recognized by the other programs. There are no cases in which information created in the other programs cannot be represented in CHAT, because CHAT is a superset of the information represented in these other programs.

Principle 5: Responsivity to Research Community Needs

TalkBank seeks to be maximally responsive to the needs of individual researchers and their research communities. We attempt to implement all features that are suggested by users in terms of software features, data coverage, documentation, and user support. We provide this support in six ways:

1. **Corpus Pages**: We have configured separate web servers for each of the 14 TalkBank communities, each within the talkbank.org domain. Each web site provides an index to the available corpora. For example, the index at https://slabank.talkbank.org lists the 37 available corpora from second language learners, organized by target language. Clicking on any one of these links, such as the one for "English-Qatar" brings up a page with a description of the corpus, photos, and contact information for the contributors, articles for citation along with their DOI numbers, a link for downloading the media, and a link for downloading the transcripts. The corpora vary widely in terms of the amount of description provided. For example, the English-Qatar corpus page only tells us that these are interviews with Qatari adult learners of English. Other corpora, such as the one described at https://slabank.talkbank.org/access/French/PAROLE.html provide more extensive documentation.
2. **TalkBank Browser**: Each corpus page also includes a link to a facility called the TalkBank browser that allows users to playback linked multimedia corpora and read the corresponding

segment of the transcript directly in their web browser. Users can choose to have continuous playback or playback of specific sections or utterances.
3. **Instructional Pages**: For AphasiaBank, TBIBank, and RHDBank there are segments of the websites that use linked video with professional commentary to teach students about the nature of these language disabilities. In the future, we hope to add such pages to SLABank.
4. **Tutorials:** We have created screencast tutorials for learning to use the database and programs at https://talkbank.org/screencasts/. These are hosted both at our own servers and through YouTube for better distribution in certain areas of the globe.
5. **Mailing lists:** For each TalkBank area, we maintain a user-oriented mailing list at https://groups.google.com. The BilingBank group is used for discussions of both SLA and bilingualism.
6. **Presentations and Workshops**: We also conduct several presentations and workshops each year at international conferences.

The guiding principle underlying all these user support methods is that we seek to be maximally responsive to the needs of researchers and research groups, as well as instructors and clinicians. We try to fulfill all requests for new corpora, new methods, new protocols, and new computational resources. In this way, we are able to maximize the participation of research groups in TalkBank.

Principle 6: Compliance with International Standards

The sixth basic TalkBank principle is a commitment to international standards for database and language technology. Toward this end, TalkBank has joined the European CLARIN (Common Language Resources and Technology Infrastructure) Federation (https://clarin.eu). CLARIN is an association of computational linguistic communities in 21 European countries, supported by the European Union and the governments of the individual countries. CMU (Carnegie Mellon University) TalkBank is currently the only member of CLARIN outside of Europe. Much like TalkBank, CLARIN seeks to provide uniform computational methods for accessing and processing language data. Toward this end, CLARIN centers have implemented standards for publishing corpus metadata using the CMDI (Component MetaData Infrastructure) format with the Handle server (https://handle.net) and OAI-PMH (Open Archives Initiative - Protocol for Metadata Harvesting) software. Based on these metadata, CLARIN has constructed a Virtual Linguistic Observatory (VLO) (https://vlo.clarin.eu) for locating linguistic resources, and nearly a third of the corpora in that system derive from TalkBank. CLARIN also promotes participation in the Core Trust Seal program for accreditation of data centers, and TalkBank has received this approval as noted in the extensive documentation at https://www.coretrustseal.org/wp-content/uploads/2017/10/TalkBank.pdf. The Core Trust Seal program emphasizes the adoption of international standards in areas such as ease of data access, protection of confidentiality, organizational infrastructure, data integrity, data storage, data curation, and data preservation. In accord with recent emphases on reproducibility of experimental (Munafò et al., 2017) and computational analyses (Donoho, 2010), TalkBank maintains incremental GIT repositories at https://git.talkbank.org for all of its datasets. Using this resource, researchers interested in replicating earlier analyses can obtain copies of segments of the database from any particular date.

Compliance with international standards represents an important step toward securing sustainability. Because the TalkBank programs and data are publicly available, open-sourced, and deployed to machines in the CMU Cloud Computing facility with systematic maintenance scripts, it is easy for other sites to mirror, archive, and extend the system. For this basic level of survivability, the Carnegie Mellon University Library has assured long-term maintenance support. NIH and NSF have provided financial support for the development of TalkBank since 1984. However,

provision of this support is subject to ongoing progress in each database. To achieve this, we are working to pass on control of TalkBank to the next generation of researchers. We also are working to establish agreements for mutual support with two related database systems: the Linguistic Data Consortium (https://www.ldc.upenn.edu) and CLARIN (https://clarin.eu).

Main Research Methods

CLAN, written by Leonid Spektor, is a program specifically designed for the analysis of TalkBank data in CHAT. Since its beginning in 1987 as a series of DOS commands, CLAN has expanded to address the needs of researchers in an increasingly wide variety of fields, based on advances in Computational Linguistics (Le Franc et al., 2018; Lubetich & Sagae, 2014; Sagae, Davis, Lavie, MacWhinney, & Wintner, 2010) and the growing capabilities of computers and the web. The CLAN editor uses Unicode UTF-8 which supports data entry in all the world's languages. Because of this, CLAN commands work for all languages. However, part-of-speech taggers are only available for Cantonese, Chinese, Danish, Dutch, English, French, German, Hebrew, Italian, Japanese, and Spanish. Currently, CLAN includes 95 commands. We can divide these commands into 5 types.

1. analysis commands,
2. profiling commands
3. format conversion commands,
4. utility commands, and
5. morphosyntactic tagging commands.

The first four command groups are described in separate chapters in the CLAN manual which is freely downloadable from https://talkbank.org/manuals/clan.pdf. The fifth group of commands is described in the MOR manual, which is downloadable from https://talkbank.org/manuals/mor.pdf. In the next five sections, we discuss the commands in each of these 5 categories. All CLAN commands can be run recursively on a directory structure, potentially analyzing every file in the database. In addition, all of the commands can send output to files or the screen.

Analysis Commands

CLAN includes 23 commands for the basic corpus analysis functions of string searching and index tabulation. Of these, the most frequently used are FREQ, KWAL, and COMBO. FREQ produces tabulations of word, lemma, or word group frequencies. It is highly customizable with over 30 switches that allow for producing output with reverse concordance, exclusion of repetitions, output in context, spreadsheet output, etc. KWAL (key word in line) outputs each matched utterance along with its line number. As in many of the other commands, KWAL and FREQ output can be triple-clicked to go back to the exact place in the original file where a string has matched. KWAL can also output a certain number of lines of context before and after the utterance in which the desired match was located. COMBO extends the capacity of KWAL to include regular expression (RegEx) matching ability.

CLAN's tabulation commands compute indices such as MLU (mean length of utterance), MLT (mean length of turn), MAXWD (the N longest words in a file), and WDLEN (word lengths in histograms). COOCCUR computes n-grams (bigrams, trigrams, etc.); vocD computes lexical diversity (Malvern, Richards, Chipere, & Purán, 2004); and TIMEDUR computes total time and pause time. There are also four analysis programs (CHAINS, CHIP, DIST, and KEYMAP) that track speech acts, sequences, and overlaps between speakers across a discourse.

Format Conversion Commands

To maximize interoperability, CLAN includes 18 format conversion programs. Of these, 7 convert from CHAT format to other formats, including ANVIL, CA, CONLL, ELAN, Praat, SRT (for subtitles), and EXMaRALDA, and there are 11 programs that convert from other formats to CHAT, including ANVIL, CONLL, ELAN, LAB, LENA, LIPP, Praat, RTF, SALT, SRT, and Text.

Profile Computation

CLAN also includes 9 programs that compute profiles. These profiles output the results of a "canned" or "packaged" group of analysis commands such as FREQ, MLU, and vocD. Profiling commands can be run in two modes. In "summary" mode, they simply output the results for each of the analyses without making comparisons against a larger database. In summary mode, one can run the profiling command on any number of transcripts in a single pass. The results will then go to a large Excel spreadsheet in which each row represents the output for one of the many input files.

In the second or "comparison" mode, the focus is on comparing the current transcript with a larger database. In such cases, the question being asked is whether the language from the current participant can best be compared against some other reference group. For example, in the case of the KIDEVAL profiling command, the question is whether the current child is well matched to others in their age group or whether they may be ahead or behind that comparison group. To evaluate this, KIDEVAL provides output on 28 standardized language measures, along with the tabulation of the use of the 14 most common inflectional morphemes in English. For each of these measures and morpheme counts, KIDEVAL can compare the current transcript with age-matched transcripts from the CHILDES database, providing a standard deviation that indicates how closely the current participant matches the larger group. For example, a transcript from an English-speaking child aged 3;6 (three years and six months) would be compared with 350 transcripts from other English-speaking children of that age to evaluate whether the child's language was consistently within the range of the comparison group. If the child had scores that fell more than one standard deviation below those of the comparison group, then the child could be considered to be a candidate for speech therapy.

For people with aphasia, the EVAL program compares a person with groups such as controls, Broca's, Wernicke's, or conduction aphasics. Other packaged profiling commands compute C-NNLA (Thompson et al., 1995), C-QPA (Rochon, Saffran, Berndt, & Schwartz, 2000), DSS (Lee, 1974), IPSyn (Scarborough, 1990), FluCalc (Bernstein Ratner & MacWhinney, 2018), MORTABLE, and SUGAR (Pavelko & Owens, 2017). Of the 9 profiling commands, the one that seems most ideal for use with L2 data is FluCalc, because it focuses directly on the measure of language fluency, accuracy, and complexity, as discussed in the section on complexity, accuracy, lexis, and fluency (CALF) below.

For second language data, we can currently only run the profiling commands in corpus-internal summary mode. For example, within the BELC corpus of L2 English in Barcelona, we can compare the results for a given 12-year-old in the *interviews* folder with the mean and standard deviation of the other 44 12-year-olds who participated in this activity. We can even compare this adolescent with the mean and standard deviation of the 26 10-year-olds or the 9 14-year-olds who took this task. However, because we do not have data from other contexts or languages, our comparison cannot extend beyond this corpus. However, if corpus creators could make consistent use of a particular picture book or video retelling task (i.e. *Frog where are you? Loch Ness Monster, Modern Times,* etc.), then it would be possible to compare samples from a given age group in Barcelona to those from the same age group in a similar learning context with other L1s (i.e. German L1 in Hamburg or Mandarin L1 in Beijing).

Utility Commands

CLAN includes 33 commands that are used to improve file consistency and format. These are quite useful during corpus development. However, once a corpus is fully curated, tagged, and checked, there is seldom need to rely on these programs.

Morphosyntactic Tagging

CLAN provides part-of-speech taggers and grammatical dependency taggers for Cantonese, Chinese, Danish, Dutch, English, French, German, Hebrew, Italian, Japanese, and Spanish. These taggers rely on lexicons designed specifically to deal with spoken language. Morphological tagging is done through a set of rules determining patterns of allomorphy and affix sequencing. Once all possible taggings are generated, they are disambiguated automatically by a trainable program called POST (Parisse & Le Normand, 2000). After that, the MEGRASP program (Sagae, Davis, Lavie, MacWhinney, & Wintner, 2007) creates a syntactic analysis in terms of grammatical dependency relations. Here is an example of the output created by the running of these programs on a single utterance in written segment of the BELC corpus:

```
*TEX:   my name is David .
%mor:   det:poss|my n|name cop|be&3S n:prop|David .
%gra:   1|2|MOD 2|3|SUBJ 3|0|ROOT 4|3|PRED 5|3|PUNCT
*TEX:   my animal favorite is the dog .
%mor:   det:poss|my n|animal n|favorite cop|be&3S det:art|the n|dog .
%gra:   1|3|MOD 2|3|MOD 3|4|SUBJ 4|0|ROOT 5|6|DET 6|4|PRED 7|4|PUNCT
*TEX:   my eyes is browns .
%mor:   det:poss|my n|eye-PL cop|be&3S v|brown-3S .
%gra:   1|2|MOD 2|3|SUBJ 3|0|ROOT 4|3|PRED 5|3|PUNCT
*TEX:   my from is Barcelona Spain .
%mor:   det:poss|my prep|from cop|be&3S n:prop|Barcelona n:prop|Spain .
%gra:   1|0|INCROOT 2|1|JCT 3|2|POBJ 4|5|NAME 5|3|PRED 6|1|PUNCT
```

Although the automatic tagging is accurate, there are learner error patterns here which require further coding. In the second utterance, *favorite* is tagged as a noun, rather than an adjective. This error likely arises from the fact that Spanish places the adjective after the noun. So, there should be the error code [* wo] here to indicate the wrong word order. Also, in the third utterance, there is a violation of subject-verb agreement which should be coded as [* m:vsg:a] for a morphological agreement error using a singular verb. Although tagging is automatic, these error codes must be added by hand.

Once a corpus has been tagged for part-of-speech on the %mor line and grammatical relations on the %gra line, it is possible to conduct a wide variety of further automatic analyses, often using the KWAL and COMBO commands. The profiling commands described earlier make extensive use of the %mor and %gra tiers to track the use of morphological and syntactic forms by learners.

It is also possible to use combinations of monolingual taggers with bilingual corpora. To do this, transcribers need to place precode marks such as [- eng] on utterances that contain switches from one language to another. Here is an example from the hogan2.cha file in the Eppler German-English corpus in BilingBank:

```
*FRI:   [- eng] police (.) want to see someone upstairs.
*SOP:   [- eng] the drug factory.
*SOP:   wir haben eine drug@s factory@s oben.
```

Because the primary language of this transcript is German, utterances in German are unmarked, whereas utterances in English use the [-eng] code. However, when an English word is used within a largely German utterance, as in the case of the third utterance with *drug factory*, then the code-switched words are marked with @s. Once a transcript has been marked in this way, it is then possible to run the two sets of morphosyntactic taggers automatically to tag the complete corpus.

Database Searches

In order to promote fuller and more direct access to the entire database, including both CHILDES and the 13 other TalkBank components, we have recently developed a database search system called TalkBankDB at https://talkbank.org/DB. TalkBankDB relies on a PostgreSQL database in JSON format. The front end web interface of TalkBankDB is written in standard HTML, CSS, and JavaScript to ensure cross-browser support. Care is taken so the JavaScript code is clearly commented and maintainable, following the popular "web component" design pattern common in many large-scale web apps. Searches are composed by successive additions of criteria. For L2 corpora, users first select SLABank and then can either choose a particular corpus or search by target language. Once the search criteria are configured, there is a button to conduct the search, and the browser will then return all files that match the criteria. At this point, it is possible to either run further analyses on the web or else download the matched set of files for further analysis in R, Python, or a statistical package, using the **Save** button. It is also possible to click on the matched items in the left-hand column to directly open up the relevant file(s).

CA Analysis

Both CHAT and CLAN provide full support for transcription and analysis in classic Jeffersonian format (Hepburn & Bolden, 2013; MacWhinney & Wagner, 2010) for Conversation Analysis (CA). Although this format is very different from the basic CHAT format used elsewhere in the databases, both formats can be expressed fully in the CHAT XML. Some important features of CLAN's support for CA analysis include the ability to mark beginnings and ends of overlaps in a way that allows CLAN's INDENT command to automatically re-align overlaps after editing has changed the spacing. Another feature is the introduction of a set of 26 special Unicode characters (see https://ca.talkbank.org/codes.html) for features of CA markup such as tone movement, final contour, speeding up and slowing down, creaky voice, and so on.

CA analysis can provide an interesting second window on corpora. Corpus linguistic analysis can reveal pervasive patterns across learners, genre, and development. CA analysis, on the other hand, is better designed to study the detailed ways in which learners interact with other conversational participants. By looking closely at small segments of interactions, sometimes involving as few as a dozen utterances, we can come to understand how learners deal in real time with conversational issues such as repair, sequencing, misunderstanding, word retrieval, face presentation, and topic continuation. A particular strength of TalkBank in general and SLABank in particular is the fact that many of the transcripts are directly linked on the utterance level to the corresponding segment of the audio or video. Based on methods like this, one can study the use of forms both in the general case and in specific interactional contexts.

Protocol Development

There is currently no standard set of methods for the elicitation of spoken second language corpus data. However, our experience with the construction of the AphasiaBank database indicates that SLA and LCR could benefit from the introduction of a common set of data elicitation methods

in the shape of a recognized data collection protocol. The reason for this is that data collected using the same set of methods are much more comparable than data collected in diverse and often incomparable ways. Because the bulk of the data in AphasiaBank were collected using this protocol, researchers have been able to use the database in over 300 additional publications.

The details of the AphasiaBank protocol can be viewed at https://aphasia.talkbank.org/protocol/. The protocol includes a series of standardized tests, a conversation about the stroke that led to aphasia, descriptions of several cartoon picture series, a retelling of the Cinderella story, and description of the simple procedure of making a peanut butter and jelly sandwich. It takes about 60 minutes to administer this protocol, and the whole interaction is recorded through video for subsequent transcription in CHAT using CLAN. This protocol has now been used in English with 420 persons with aphasia and 286 control participants. A smaller number of protocols have been collected in other languages. Because of the consistency of data administration, transcription, and analysis of this corpus, researchers have been able to publish nearly 300 papers based on this dataset.

A similar approach could be used for the creation of a unified second language learning database. In fact, corpora such as FLLOC (http://flloc.soton.ac.uk), SPLLOC (http://splloc.soton.ac.uk), and LANGSNAP (http://langsnap.soton.ac.uk) that are included in SLABank have already taken this approach, by including recognized tasks such as interviews and story retelling. Other useful protocol segments include passage reading, word list reading, map tasks, translation tasks, error correction, and sentence combining. Each of these components can provide important and somewhat independent sources of information regarding learners' abilities and development.

1. **Passage reading** and **word list reading** can provide consistently comparable data across participants that can then be aligned against target forms using automatic speech recognition (ASR) technology. Diarization through automatic speech recognition (ASR) (Cristia, Ganesh, Casillas, & Ganapathy, 2018) can provide marks of the beginning and end of each word and segment. ASR analysis can also analyze deviations in learner pronunciations from the target forms. We have been using this methodology in the AphasiaBank project to study people with apraxia of speech (AoS).
2. **Interviews** provide a view of learners' abilities in a maximally natural conversation.
3. **Story retelling** has provided extremely rich data in fields such as child language and aphasia. The corpora in the European Science Foundation (ESF) database (which is included in SLABank) rely heavily on the retelling of the stories in short movie clips. In dozens of CHILDES corpora, children are asked to retell stories about a boy, his dog, and a frog. In the AphasiaBank protocol, both control participants and participants with aphasia retell the Cinderella story after being reminded of the plot by scanning through a wordless picture book. Data from these retellings can be used for the study of the control of fluency, grammatical correctness, morphosyntactic development, lexical diversity, and narrative structure.
4. **Map tasks** and similar procedural tasks can evaluate the use of language for conversational problem solving.
5. **Translation and error correction** tasks can assess the ability of learners to apply specific language patterns in areas such as syntax, lexicon, and morphology. They also provide consistent data on error patterns for specific targets.
6. **Sentence combining** tasks can evaluate learners' control of methods for constructing complex constructions.

The IRIS database (http://iris-database.org) includes a wide range of further task types. If SLA and LCR researchers could agree on a core set of these tasks that would constitute a shared protocol, then it would be possible to achieve the same level of database structure already achieved for AphasiaBank, FluencyBank, and TBIBank. The tasks in such a general protocol can also be applied in successive sessions across months or even years to study the overall course of second language learning.

Complexity, Accuracy, Lexis, and Fluency

Proposals regarding a tradeoff between fluency and accuracy (Skehan, 1998) have generated many interesting attempts to measure and characterize fluency and accuracy objectively (Wen & Ahmadian, in press). This concept of a tradeoff between these dimensions has been further elaborated by many researchers, including several in this volume, to include the dimensions of syntactic and lexical complexity. As a result, the full set of dimensions now includes complexity, accuracy, lexis, and fluency (CALF). Each of these dimensions is the subject of a chapter in this handbook. The specific measures discussed in these chapters are very close to those being computed by CLAN's FLUCALC, VOCD, MATTR, and MEGRASP programs.

Fluency. As noted by Huensch (Chapter 22, this volume), CLAN has been used to evaluate fluency in terms of pause duration, retracing, filled pauses, and disfluencies. Once transcription is complete, CLAN's FLUCALC program calculates and tabulates each of these features automatically. In addition, CLAN's interoperability with additional programs such as Praat and Phon facilitates further fluency analyses.

Accuracy. CLAN can evaluate accuracy using the TalkBank system of error analysis (Chapter 23, this volume) that has been developed based on experience in coding errors in aphasia and child language learning. The description of the TalkBank error coding system can be found in Chapter 18 of the CLAN manual (http://talkbank.org/manuals). TalkBank's system for error coding and analysis is very similar in terms of both the level of coding and the mechanisms for analysis as the system developed in Louvain (Dagneaux, Denness, & Granger, 1998; Granger, 2003) for SLA texts.

Lexis. CLAN provides three methods for computing lexical diversity (Chapter 25, this volume). The FREQ program computes the traditional TTR (type-token ratio) measure, based either on the specific words being spoken or on an analysis of lemmas, as analyzed on the tagged and analyzed morphological analysis tier (%mor). The second method uses the vocD framework (Malvern et al., 2004) which is less sensitive to sample size than TTR, and the third uses the MATTR computation (Covington & McFall, 2010), which is even less sensitive to sample size than vocD. CLAN's RARITY program examines the extent to which a speaker uses words that are less frequent in the general spoken vocabulary.

Complexity. CLAN's analysis of complexity (Chapter 24, this volume) relies on the automatic computation of a full grammatical dependency tagging (%gra) for each utterance, as displayed in the previous section on morphosyntactic tagging. Once this tagging is available, there is a specific version of the FREQ command described in the CLAN manual that can send the required data to Excel to tabulate and add complex structures and compute the ratio of complex grammatical relations over all grammatical relations.

Representative Corpora and Research

SLABank currently includes 37 corpora from second language learners, and BilingBank includes 13 corpora from bilinguals. Nearly all of these corpora are accompanied by audio, although only a few have been linked to the audio at the utterance level. In addition to these corpora from adult learners and bilinguals, the CHILDES database has 32 corpora tracing the development of childhood bilingualism. For a detailed description of the research engendered by each corpus, please consult the corpus pages located from the index at https://slabank.talkbank.org/access/ for SLABank, https://biling.talkbank.org/access/ for BilingBank, and https://childes.talkbank.org/access/Biling/ for the bilingual corpora in CHILDES.

The corpora in languages for which we have MOR taggers have been annotated for morphological structure (%mor), and some have also been automatically annotated for grammatical relation structure (%gra). The languages for which taggers are available include Cantonese, Danish, Dutch, English, French, German, Hebrew, Italian, Japanese, Mandarin, and Spanish

(MacWhinney, 2008). This means that, for SLA Bank, we cannot currently tag the corpora dealing with the learning of Czech and Hungarian.

Studies based on the current SLABank corpora have largely been conducted by the groups who developed and contributed the corpora. For example, the Vercellotti corpus of L2 English has been carefully time-aligned at the phrasal level. As a result, it was amenable to a close analysis for CALF features (Vercellotti, 2017), such as pause duration, retracing, and lexical diversity. The large Barcelona English Language Corpus (BELC) was collected for the purpose of evaluating the effects of age on the acquisition of English as a second language (Muñoz, 2006). As a result, the use of CLAN programs for that corpus focused on comparison of lexicon and syntax across different age groups. The Eppler corpus of German-English code-switching includes thorough marking of the source language for each sentence and marking of sentence-internal switches. The fact that the corpus is fully tagged for both English and German has made it possible to examine the predictions from three alternative syntactic theories for mixed determiner-noun constructions (Eppler, Luescher, Deuchar, & Theory, 2017). The CLAN programs were also used in the analysis of the FLLOC corpus on the learning of French (Marsden & David, 2008) and the SPLLOC corpus on the learning of Spanish (Domínguez & Arche, 2014), among many others.

Future Directions

To be maximally effective, SLABank needs to address three major goals. The first is to encourage researchers to share the data they have collected. Researchers may avoid data-sharing because of concerns regarding prohibitions from Institutional Review Boards or regulations such as the General Data Protection Regulation (GDPR) and the California Data Privacy Protection Act (CDPPA). However, good methods exist for addressing all of these requirements. Moreover, funding agencies and universities expect that researchers will share their data as much as possible. Another possible barrier to data-sharing is that TalkBank data must be coded in CHAT format. However, many projects already use CHAT, and for those that are using other formats, there are 11 CLAN format conversion programs, and more can be created as needed. A third barrier to data-sharing is that many researchers are unfamiliar with the use of CLAN tools for data analysis. However, for those more comfortable with tools such as AntConc or ELAN, it is easy to output CHAT data for use by these other programs.

Increased data-sharing of SLA corpora will improve our current coverage of learner groups and target languages. Currently, the most heavily represented target languages are English, Spanish, and French. Adding data from other major target languages such as Arabic, German, Mandarin, and Russian is important. It is also important to have data on the L2 acquisition of less commonly studied languages, as well as data collected from alternative language-learning contexts, including online learning, instructed learning in the classroom, study abroad, learning in bilingual homes, learning in immigrant communities, and so on. The possibility of collecting data through online methods is particularly appealing, given the large potential coverage of these methods. Such data could include oral and/or written story retelling, oral passage reading, and spoken or written translation. These data could also be linked to data from online tutorial and online experimental methods (MacWhinney, 2017).

A second major goal for the LCR and SLA communities will be the creation of a shared protocol for data elicitation. Such a protocol would greatly improve our ability to characterize patterns in second language learning across L1s, L2s, learner types, contexts, and ages. To develop this protocol, the SLABank system will need to solicit input from all segments of the community and researchers will need to pilot the method in new projects.

Work on these goals will require extensive buy-in from the LCR and SLA communities, as well as major support from funding agencies. The good news is that it is clear what needs to be done and how to proceed.

Further Reading

MacWhinney, B. (2008). Enriching CHILDES for morphosyntactic analysis. In H. Behrens (Ed.), *Trends in corpus research: Finding structure in data* (pp. 165–198). Amsterdam: John Benjamins.
This paper explains morphosyntactic and grammatical dependency tagging in CLAN.

MacWhinney, B. (2017). A shared platform for studying second language acquisition. *Language Learning*, *67*(S1), 254–275.
This paper focuses on the collection of L2 learning data through web-based exercises, but also includes a discussion of how these methods can be linked to corpus development.

MacWhinney, B. (2019). Task-based analysis and the Competition Model. In Z. Wen & M. Ahmadian (Eds.), *Researching Second Language Task Performance and Pedagogy: Essays in Honor of Peter Skehan*. (pp. 305–315). New York: John Benjamins.
This paper explores links between the Competition Model and CALF trade-off analysis, using the CLAN programs.

MacWhinney, B. (2019). Understanding spoken language through TalkBank. *Behavior Research Methods, 51*, 1919–1927.
This article explains how TalkBank methods have been used in various fields to explore psychological, clinical, and linguistic issues.

Related Topics

Chapters 22, 23, 24, and 25.

References

Bernstein Ratner, N., & MacWhinney, B. (2018). Fluency bank: A new resource for fluency research and practice. *Journal of Fluency Disorders*, *56*, 69–80. doi:10.1016/j.jfludis.2018.03.002

Covington, M. A., & McFall, J. D. (2010). Cutting the Gordian knot: The moving-average type–token ratio (MATTR). *Journal of Quantitative Linguistics*, *17*(2), 94–100. doi:10.1080/02687038.2012.693584

Cristia, A., Ganesh, S., Casillas, M., & Ganapathy, S. (2018). Talker diarization in the wild: The case of child-centered daylong audio-recordings. Paper presented at the Interspeech, Hyderabad.

Dagneaux, E., Denness, S., & Granger, S. (1998). Computer-aided error analysis. *System*, *26*(2), 163–174.

Domínguez, L., & Arche, M. J. J. L. (2014). Subject inversion in non-native Spanish. Lingua, *145*, 243–265.

Donoho, D. L. (2010). An invitation to reproducible computational research. *Biostatistics*, *11*(3), 385–388.

Ekman, P., & Friesen, W. (1978). *Facial action coding system: Investigator's guide*. Palo Alto: Consulting Psychologists Press.

Eppler, E. D., Luescher, A., Deuchar, M. J. C. L., & Theory, L. (2017). Evaluating the predictions of three syntactic frameworks for mixed determiner–noun constructions. *Corpus Linguistics and Linguistic Theory*, *13*(1), 27–63.

Eskildsen, S. W. (2012). L2 negation constructions at work. *Language Learning*, *62*(2), 335–372.

Geertzen, J., Alexopoulou, T., & Korhonen, A. (2013). *Automatic linguistic annotation of large scale L2 databases: The EF-Cambridge open language database (EFCAMDAT).* Paper presented at the proceedings of the 31st second language research forum. Cascadilla Proceedings Project, Somerville.

Goodwin, C. (2003). Conversational frameworks for the accomplishment of meaning in aphasia. In *Conversation and brain damage* (pp. 90–116). Oxford: Oxford University Press.

Granger, S. (2003). Error-tagged learner corpora and CALL: A promising synergy. *CALICO Journal*, *20*, 465–480.

Hall, E. (1966). *The hidden dimension*. New York: Random House.

Hepburn, A., & Bolden, G. B. (2013). The conversation analytic approach to transcription. In J. Sidnell & T. Stivers (Eds.), *The handbook of conversation analysis* (pp. 57–76). New York: Wiley.

Kaltenboeck, G., & Heine, B. (2015). Sentence grammar vs. thetical grammar: Two competing domains. In B. MacWhinney, A. Malchukov, & E. Moravcsik (Eds.), *Competing motivations in grammar and usage* (pp. 348–363). New York: Oxford University Press.

Kemmerer, D. (2015). *The cognitive neuroscience of language*. New York: Psychology Press.

Kendon, A. (1982). The study of gesture: Some observations on its history. *Recherches Sémiotiques/Semiotic Inquiry*, *2*(1), 45–62.

Lee, L. (1974). *Developmental sentence analysis*. Evanston: Northwestern University Press.
Le Franc, A., Riebling, E., Karadayi, J., Yun, W., Scaff, C., Metze, F., & Cristia, A. (2018). The ACLEW DiViMe: An easy-to-use diarization tool. Paper presented at the Interspeech 2018, Mumbai.
Lubetich, S., & Sagae, K. (2014). Data-driven measurement of child language development with simple syntactic templates. Paper presented at the COLING 2014, Dublin, Ireland.
MacWhinney, B. (2008). Enriching CHILDES for morphosyntactic analysis. In H. Behrens (Ed.), *Trends in corpus research: Finding structure in data* (pp. 165–198). Amsterdam: John Benjamins.
MacWhinney, B. (2014). Presentation. In L. Scliar-Cabral (Ed.), *O português na plataforma CHILDES* (pp. 9–20). Florianopolis: Editora Insular.
MacWhinney, B. (2017). A shared platform for studying second language acquisition. *Language Learning*, *67*(S1), 254–275.
MacWhinney, B., & Wagner, J. (2010). Transcribing, searching and data sharing: The CLAN software and the TalkBank data repository. *Gesprachsforschung*, *11*, 154–173.
Magnuson, J., Dixon, J., Tanenhaus, M., & Aslin, R. (2007). The dynamics of lexical competition during spoken word recognition. *Cognitive Science*, *31*(1), 133–156.
Malvern, D., Richards, B., Chipere, N., & Purán, P. (2004). *Lexical diversity and language development*. New York: Palgrave Macmillan.
Marsden, E., & David, A. (2008). Vocabulary use during conversation: A cross-sectional study of development from year 9 to year 13 among learners of Spanish and French. *Language Learning Journal*, *36*(2), 181–198.
Marsden, E., Morgan-Short, K., Thompson, S., & Abugaber, D. (2018). Replication in second language research: Narrative and systematic reviews and recommendations for the field. *Language Learning*, *68*(2), 321–391. doi:10.1111/lang.12286
Marsden, E., & Plonsky, L. (2018). Data, open science, and methodological reform in second language acquisition research. In A. Gudmestad & A. Edwonds (Eds.), *Critical reflections on data in second language acquisition* (pp. 219–228). Amsterdam: John Benjamins.
Munafò, M. R., Nosek, B. A., Bishop, D. V., Button, K. S., Chambers, C. D., du Sert, N. P., ... Ioannidis, J. P. (2017). A manifesto for reproducible science. *Nature Human Behaviour*, *1*, 0021.
Muñoz, C. (2006). *Age and rate of foreign language learning*. Clevedon: Multilingual Matters.
Myles, F. (2015). Second language acquisition theory and learner corpus research. In S. Granger, G. Gilquin, & F. Meunier (Eds.), *Cambridge handbook of learner corpus research* (pp. 309–331). Cambridge: Cambridge University Press.
Parisse, C., & Le Normand, M.-T. (2000). Automatic disambiguation of the morphosyntax in spoken language corpora. *Behavior Research Methods, Instruments, and Computers*, *32*(3), 468–481.
Pavelko, S., & Owens, R. (2017). Sampling utterances abd grammatical analysis revised (SUGAR): New normative values for language sample analsysi measures. *Language, Speech, and Hearing Services in Schools*, *48*(3), 197–215.
Rochon, E., Saffran, E., Berndt, R., & Schwartz, M. (2000). Quantitative analysis of aphasic sentence production: Further development and new data. *Brain and Language*, *72*(3), 193–218.
Sagae, K., Davis, E., Lavie, A., MacWhinney, B., & Wintner, S. (2007). High-accuracy annotation and parsing of CHILDES transcripts. In *Proceedings of the 45th meeting of the association for computational linguistics* (pp. 1044–1050). Prague: ACL.
Sagae, K., Davis, E., Lavie, A., MacWhinney, B., & Wintner, S. (2010). Morphosyntactic annotation of CHILDES transcripts. *Journal of Child Language*, *37*(3), 705–729. doi:10.1017/S0305000909990407
Scarborough, H. S. (1990). Index of productive syntax. *Applied Psycholinguistics*, *11*(1), 1–22. doi:10.1017/S0142716400008262
Schegloff, E. (2007). *Sequence organization in interaction: A primer in conversation analysis*. New York: Cambridge University Press.
Skehan, P. (1998). *A cognitive approach to language learning*. Oxford: Oxford University Press.
Thompson, C. K., Shapiro, L. P., Tait, M. E., Jacobs, B. J., Schneider, S. L., & Ballard, K. J. (1995). A system for the linguistic analysis of agrammatic language production. *Brain and Language*, *51*, 124–129.
Vercellotti, M. L. (2017). The development of complexity, accuracy, and fluency in second language performance: A longitudinal study. *Applied Linguistics*, *38*(1), 90–111.
Wen, Z., & Ahmadian, M. (Eds.) (in press). *Researching L2 task performance and pedagogy: In honor of Peter Skehan*. Amsterdam: John B[illegible]mins.