Best Practices in the TalkBank Framework

Brian MacWhinney, Yvan Rose, Leonid Spektor, and Franklin Chen
Carnegie Mellon University, Psychology
5000 Forbes Ave. Pittsburgh, PA 15213
E-mail: macw@cmu.edu

Abstract

TalkBank is an interdisciplinary research project funded by the National Institutes of Health and the National Science Foundation. The goal of the project is to support data sharing and direct, community-wide access to naturalistic recordings and transcripts of spoken communication. TalkBank has developed consistent practices for data sharing, metadata creation, transcription methods, transcription standards, interoperability, automatic annotation, and dissemination. The database includes corpora from a wide variety of linguistic fields all governed by a comprehensive XML Schema. For each component research subfield, TalkBank must provide special purpose annotations and tools as a subset of the overall system. Together, these various TalkBank standards can serve as guides to further improvements in the use of speech corpora for linguistic research.

Keywords: corpora, speech, conversation analysis, phonology, syntax, transcription, metadata

1. Best Practices

The goal of this workshop is to examine best practices for configuring speech corpora for linguistic research. This would seem to be a fairly well defined goal. Ideally, one could formulate a single set of best practices that would apply across the board. However, when we consider specific corpora, systems, groups, issues, and constraints, the characterization of "best practices" becomes more complicated. Take the CallFriend corpus, as an example. The Linguistic Data Consortium (LDC) created this phone call corpus for the purposes of developing automatic speech recognition (ASR) systems. Thanks to the generosity of LDC, segments of CallFriend have been made available to the TalkBank system for transcription and further linguistic analysis. We have transcribed these calls in the CHAT editor, using Conversation Analysis standards and linked them on the utterance level to the audio media. The best practices in this case depend heavily on the particular shape of the corpus and the uses to which it will be put. These are phone calls with good stereo separation, but there are often noises on the phone line. This seems to violate best practices in speech technology, but it is quite adequate for the purposes of Conversation Analysis. On the other hand, the demographic information associated with each call is inadequate for standard sociolinguistic or sociophonetic analysis. Also, LDC provided no transcriptions for these calls, so the issue of best practices in transcription rests totally outside of the realm of the initial data collection.

When we consider best practices across a wide collection of corpora, the problem becomes further magnified. Ini particular, for each of the 386 corpora in the TalkBank database, collected under a myriad of different conditions with differing goals, we could conduct an analysis of best practices, usually with quite different results. This suggests that we should view best practices not as a single framework, but as a Swiss Army knife that presents the user with a variety of tools, each suited for a given type of linguistic analysis.

The TalkBank system is an atttempt to provide just this type of Swiss Army knife. For researchers studying child phonology, it offers the PhonBank system (Rose & MacWhinney, in press). For morphosyntactic analysis, it provides taggers (MacWhinney, 2008) and parsers (Sagae, Davis, Lavie, MacWhinney, & Wintner, 2010). For Conversation Analysts, it provides Jeffersonian coding (Jefferson, 1984) and formats for gestural analysis (MacWhinney, Fromm, Forbes, & Holland, 2011). Some of the blades of the knife can be used for many purposes; others are more specialized. In this report, we will explain how each blade has been adapted to the task at hand. In some cases, the blades offered by TalkBank are not the best available and we need to then explain how data in the TalkBank format can then be exported to other programs. In other areas, such as metadata coding, TalkBank has essentially off-loaded the issue of best practices to other systems.

2. Background

TalkBank (http://talkbank.org) is an interdisciplinary research project funded by the National Institutes of Health and the National Science Foundation. The goal of the project is to support data sharing and direct, community-wide access to naturalistic recordings and transcripts of spoken communication. TalkBank extends the model for data sharing and analysis first developed in the context of the CHILDES project (MacWhinney, 2000). Although CHILDES is the most established of these datasets, other systems, such as PhonBank, AphasiaBank, CABank, BilingBank, and SLABank have also achieved general recognition and acceptance within the relevant research communities.

CHILDES contains 68 million words of child-adult conversation across 26 languages; the other segments of TalkBank include 63 million words of adult-adult conversation with the bulk in English. Although many earlier child language corpora were not contributed along with their media, the current default format for both CHILDES and TalkBank assumes that transcripts will be linked to either audio or video on the level of the utterance. This means that all new TalkBank corpora are, in effect, speech corpora. To the degree that the methods of speech technology can be applied to naturalistic conversational data of the type collected in TalkBank, the merger of speech technology with linguistic analysis envisioned in this workshop has already taken place in the TalkBank framework.

This workshop has specified a set of 12 themes for analysis of best practices. These are:

1. speech corpus designs and corpus stratification schemes
2. metadata descriptions of speakers and communications
3. legal issues in creating, using and publishing speech corpora for linguistic research
4. transcription and annotation tools for authentic speech data
5. use of automatic methods for tagging, annotating authentic speech data
6. transcription conventions in conversation analysis, dialectology, sociolinguistics, pragmatics and discourse analysis
7. corpus management systems for speech corpora
8. workflows and processing chains for speech corpora in linguistic research
9. data models and data formats for transcription and annotation data
10. standardization issues for speech corpora in linguistic research
11. dissemination platforms for speech corpora
12. integration of speech corpora from linguistic research into digital infrastructures

TalkBank addresses issues 2, 3, 4, 5, 6, 7, 9, and 11. Issues 1, 8, 10, and 12 lie outside the scope of TalkBank and are left either to individual researchers or the wider scientific community. In the next sections, we will outline TalkBank approaches to the eight best practices issues it has addressed.

3. Metadata

TalkBank has addressed the Metadata issue by subscribing to both the OLAC and IMDI formats. For each of the 386 corpora in TalkBank, we create a single text file in a consistent format that provides information relevant to all files in the corpus. The OLAC program, which is built into the CLAN programs, compiles this information across the database into a single file for harvesting by OLAC. For IMDI, we also include headers in the individual files that provide further file-specific metadata. Rather than using the ARBIL program, we use the IMDI program in CLAN to combine this information into files that can be included in IMDI. In addition, all of the transcripts and media of the complete CHILDES and TalkBank databases are included in IMDI and freely available through that system.

We are working to specify further detailed best practice specifications for metadata in the area of sociolinguistics. Toward that end, we have contributed to recent workshops organized by Chris Cieri and Malcah Yaeger-Dror at NWAV and LSA, designed to improve best practices in the coding of sociolinguistic metadata and to stimulate data-sharing in that field. To facilitate the declaration of sociolinguistic and other corpora, we have provided a page at http://talkbank.org/metamaker that allows researchers to describe the shape and availability of corpora that are not yet included in any major database. This information is then transmitted to OLAC.

4. Legal Issues and Data Sharing

The 386 corpora in TalkBank have all cleared IRB review, and nearly all are available for open access and downloading. In the process of establishing this level of open access, we have acquired decades of experience with IRB and legal issues. The results of this experience are encoded in a set of principles for data-sharing, IRB guidelines, suggested informed consent forms, alternative levels of access or password protection, and methods for anonymizing data, all available from http://talkbank.org/share

In practice, the only corpora that require password access are those from participants with clinical disabilities. For the other corpora, we are careful to replace last names with the capitalized English word "Lastname" and addresses with the word "Address". For some corpora, such as the Danish SamtaleBank corpus, we have also replaced the last names and addresses in the audio files with silence.

Often researchers claim that their data cannot be shared, because access has not been approved by their IRB. In practice, we have found that this is seldom the case. IRBs will nearly always approve data sharing with anonymization and password protection. In reality, researchers use IRB restrictions as a way of avoiding opening their data to other investigators, because they believe that other researchers can achieve a competitive advantage. In this sense, the reference to legal and IRB issues is frequently used to divert discussion of the underlying problem of competitive advantage in academics. We believe that best solution to this problem is for granting agencies to require data sharing as a condition for further funding.

5. Transcription Tools

Apart from its scope, coverage, multilinguality, and size, there is another core feature that characterizes TalkBank. This is the fact that all of the data in the system are formatted in accord with a single consistent standard called CHAT that is bidirectionally convertible to

TalkBank XML (http://talkbank.org/xsddoc). Over the years, CHAT has been crafted as a superset of its component transcription standards. For example, it supports at the same time standard Jeffersonian Conversation Analysis (CA) coding, the linguistically-oriented transcription methods of child language, phonological coding methods through IPA, disfluency analysis methods for speech errors and stuttering, and new methods for gesture coding in nested dependent files. Each of these transcription standards is implemented as a subcomponent of the overall TalkBank CHAT standard and individual transcripts can declare to which set of conventions they adhere. This approach allows us to provide all the codes that are needed for each subdiscipline without requiring any of them to make use of all the codes for their own special corpora.

The benefit of this approach is that the analysis programs can operate on all corpora in a consistent way and users only need to learn the CLAN program (http://talkbank.org/software) to analyze everything in TalkBank. In this regard, the TalkBank framework differs fundamentally from that of other systems such as LDC or Lacito. These other archiving system accept corpora in a wide variety of formats and users must learn different tools and methods to process each of the alternative corpora, even within a particular topic area.

The imposition of consistent coding standards comes at a cost. Transcription in CHAT can be rigorous and demanding. For the beginner, it takes several days to learn to transcribe smoothly. In some other cases, researchers are unwilling to use CHAT at all and prefer to create corpora in their own formats. When those corpora are contributed to the database, we then write special purpose programs to reformat them. However, we can automatically convert corpora formatted in SALT, ELAN, EXMARaLDA, Transcriber, Praat, or ANVIL.

To facilitate the mechanics of transcription, the CLAN editor supports several methods of linking to the media during and after transcription. These methods include Transcriber Mode, Sound Walker Mode, Sonic Mode, and Hand Editing Mode. Transcriber Mode uses the space-bar method of the Transcriber program (http://trans.sourceforge.net). Sound Walker mode operates like the old dictation machine with an optional foot pedal. Sonic Mode relies on display of the waveform for both audio and video files. We are interested in further improvements of CHAT transcription based on presegmentation of the audio using HTK routines.

6. Automatic Annotation

TalkBank has developed systems for automatic tagging of morphology (MOR), dependency syntax (GRASP), and phonology (Phon). Based on the morphosyntactic codes produced by MOR and GRASP, the CLAN programs can automatically compute syntactic profiles for the DSS (Lee, 1974) and IPSyn (Sagae, Lavie, & MacWhinney, 2005). MOR part-of-speech taggers have been developed for 11 languages and GRASP dependency grammars for 3 languages. These systems are described in detail in another LREC paper in this volume. In the area of phonology, the Phon program requires manual IPA transcription of non-standard child forms. However, the IPA representation for standard adult forms can be inserted automatically from the orthographic transcription. In addition, Phon provides automatic segmentation of phonological forms into syllables and syllable positions.

Apart from automatic tagging, CLAN provides methods for automatic transcript analysis. For example, the MORTABLE program provides complete counts of all grammatical morphemes in a set of transcripts, based on codes in the %mor line. The EVAL program provides package analyses of overlaps, pauses, morpheme counts and so on. We are now working to supplement these methods for automatic tagging and analysis with methods that automatically align transcripts to media at the word level and then compute a variety of fluency measures. For more careful, special purpose analyses, CLAN provides 14 analytic measures such as VOCD (Malvern, Richards, Chipere, & Purán, 2004), MLU, FREQ, and many others.

7. Transcription Conventions

To provide detailed coding methods for specific subfields, the TalkBank XML format strives to integrate best practices from each of the relevant subfields into a single unified annotation format. Unlike Partitur systems such as Anvil, EXMaRALDA, or ELAN that use time marks as the fundamental encoding framework, TalkBank XML takes the spoken word as the fundamental encoding framework. This provides results that are easy to scan across the page. Overlap alignment is also well supported through special Unicode characters that mark overlap begin and end. However, the display of overlap is not as graphic and intuitive as in the Partitur format. Because CHAT can be quickly transformed into ELAN and EXMARaLDA formats, users who need to study overlap in this way can have both views available. The only problem with this solution is that editing work done in the other systems may not be importable back to CHAT, unless the user is careful to only use CHAT conventions in the other system.

Here, we will summarize the major dimensions of CHAT transcription, coding, and annotation. The basic format involves a main line that is then supplemented by a series of dependent tiers.

1. **The main line.** This line uses a combination of eye-dialect and conventional orthography to indicate the basic spoken text. A full explication of the entire CHAT coding scheme would be outside of the scope of the current chapter. The manual of conventions is available at http://childes.psy.cmu.edu/manuals. These conventions include a wide variety of CA

codes marked through special Unicode characters entered through combinations of the F1 and F2 function keys with other characters. This system is described at http://talkbank.org/CABank/codes.html and in MacWhinney and Wagner (2010)

2. **Morphological and syntactic lines.** The MOR and GRASP programs compute these two annotation lines automatically. The forms on these lines stand in a one-to-one relation with main line forms, excluding retraces and nonwords. This alignment, which is maintained in the XML, permits a wide variety of detailed morphosyntactic analyses. We also hope to use this alignment to provide methods for writing from the XML to a formatted display of interlinear aligned morphological analysis.
3. **Phonological line**. The %pho line stands in a one-to-one relation with all words on the main line, including retraces and nonwords. This line uses standard IPA coding to represent the phonological forms of words on the main line. To represent elision processes, main line forms may be grouped for correspondence to the %pho line. The Phon program developed by Yvan Rose and colleagues (Rose, Hedlund, Byrne, Wareham, & MacWhinney, 2007; Rose & MacWhinney, in press) is able to directly import and export valid TalkBank XML.
4. **Error analysis**. In earlier versions of the system, errors were coded on a separate line. However, we have found that it is more effective to word-level code errors directly on the main line, using a system specifically elaborated for aphasic speech at http://talkbank.org/AphasiaBank/errors.doc.
5. **Gesture coding**. Although programs such as ELAN and Anvil provide powerful methods for gesture coding, we have found that it is often difficult to use these programs to obtain an intuitive understanding of gesture sequences. Simply linking a series of gesture codes to the main line in TalkBank XML is similarly inadequate. To address this need, we have developed a new method of coding through nested coding files linked to particular stretches of the main line. These coding files can be nested indefinitely, but we have found that two levels of embedding are enough for current analysis needs. Examples of these gesture coding methods can be found at http://talkbank.org/CABank/gesture.zip.
6. **Special coding lines**. CLAN and TalkBank XML also support a wide variety of additional coding lines for speech act coding, analysis of written texts, situational background, and commentary. These coding tiers are not aligned only to utterances and not to individual words.

8. Dissemination Platforms

The fundamental idea underlying the construction of TalkBank is the notion of data sharing. By pooling their hard-won data together, researchers can generate increasingly accurate and powerful answers to fundamental research questions. The CHILDES and TalkBank web sites are designed to maximize the dissemination of the data, programs, and related methods. Transcript data can be downloaded in .zip format. Media can be downloaded or played back over the web through QuickTime reference movie files. The TalkBank browser allows users to view any TalkBank transcript in the browser and listen to the corresponding audio or see the corresponding video in continuous playback mode, linked on the utterance level. We also provide methods for running CLAN analyses over the web, which we are now supplementing with analyses that use the XML database as served through the Mark Logic interface.To teach the use of the system, we have produced manuals, instructional videos and powerpoint demonstrations which we use in a wide variety of workshops internationally

9. Conclusion

Together these various TalkBank facilities provide a comprehensive, interoperable set of best practices for the coding of spoken language corpora for research in linguistics, psycholinguistics, speech technology, and related disciplines. New methods and improvements to these practices are continually in development, as we expand the database to include a fuller representation of the many forms of spoken communication.

10. References

Jefferson, G. (1984). Transcript notation. In J. Atkinson & J. Heritage (Eds.), *Structures of social interaction: Studies in conversation analysis* (pp. 134-162). Cambridge: Cambridge University Press.

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., Fromm, D., Forbes, M., & Holland, A. (2011). AphasiaBank: Methods for studying discourse. *Aphasiology, 25*, 1286-1307.

MacWhinney, B., & Wagner, J. (2010). Transcribing, searching and data sharing: The CLAN software and the TalkBank data repository. *Gesprächsforschung, 2*, 1-20.

Malvern, D. D., Richards, B. J., Chipere, N., & Purán, P. (2004). *Lexical diversity and language development*. New York: Palgrave Macmillan.

Rose, Y., & MacWhinney, B. (in press). The Phon and PhonBank initiatives.

Sagae, K., Davis, E., Lavie, A., MacWhinney, B., & Wintner, S. (2010). Morphosyntactic annotation of CHILDES transcripts. *Journal of Child Language, 37*, 705-729.

Sagae, K., Lavie, A., & MacWhinney, B. (2005). Automatic measurement of syntactic development in child language *Proceedings of the 43rd Meeting of the Association for Computational Linguistics* (pp. 197-204). Ann Arbor: ACL.