Each corpus in TalkBank must be accompanied by a 0metadata.cdc file that includes these fields: corpus title, language, contributors, media format, country, format of study year of contribution, and DOI. This file is created by TalkBank workers when the corpus is added to the database. Other aspects of metadata are placed into the homepage for the corpus in the following sections. The principles for creating metadata apply to all the corpora in TalkBank including clinical corpora, PhonBank, and others.

Acknowledgments. Each corpus has a statement that asks the user to cite some particular reference when using the corpus. For example, researchers using the Adam, Eve, and Sarah corpora from Roger Brown and his colleagues are asked to cite Brown (1973).

Restrictions. Contributors can set restrictions on the use of their data. For example, researchers may ask that they be sent copies of articles that make use of their data. Most researchers have chosen not to set limitations on the use of their data, apart from the general guidelines for use of TalkBank data stated in the Ground Rules.

Warnings. This documentation file should also warn other researchers about limitations on the use of the data. For example, if an investigator paid no attention to correct transcription of speech errors, this could be noted.

Project Description. There should be detailed information on the history of the project. How was funding obtained? What were the goals of the project? How was data collected? What was the sampling procedure? How was transcription done? What was ignored in transcription? Were transcribers trained? Was reliability checked? Was coding done? What codes were used? Was the material computerized? How?

Codes. If there are project-specific codes, these should be described.

Demographic data. Where possible, extensive demographic, dialectological, and psychometric data should be provided for each informant. There should be information on topics such as age, gender, siblings, schooling, social class, occupation, previous residences, religion, interests, friends, and so forth. Information on where the parents grew up and the various residences of the family is particularly important in attempting to understand sociolinguistic issues regarding language change, regionalism, and dialect. Without detailed information about specific dialect features, it is difficult to know whether these markers are being used throughout the language or just in certain regions.

For clinical corpora, demographic data may include test scores, medical diagnosis, history of disability (stroke, tumor, trauma, etc.), acute status, chronic period, therapy, etc.

Situational descriptions. The readme file should include descriptions of the contexts of the recordings, such as experimental room or the layout of the child's home and bedroom or the nature of the activities being recorded. Additional specific situational information can be included in the @Situation and @Comment fields in each file.