Data citation formats
Draft proposal

Stanley Dubinsky
University of South Carolina
dubinsky@sc.edu

Developing Standards for Data Citation & Attribution in Linguistics
Workshop 2 | Austin, TX | April 8-10, 2016

Asking the questions …

For the various types of data you use or produce what are:

- The term(s) by which you refer to it
- The nature of the data (e.g. recorded speech, transcribed sentences, sound samples, stimuli, …)
- The units of analysis (e.g. conversations, words, sounds, sentences, gestures, etc.)
- The digital form of storage (e.g. pdfs, mp3 files, jpegs, txt files, spreadsheets, etc.)
- The size (in megabytes or gigabytes) of a typical file
- The most typical manner of storage (CDs, hard drives, cloud, tape, index cards, printed or typed pages)

Thanking those who answered the call

- **Gene Buckley**
- **Kathryn Campbell-Kibler**
- **Donna Christian**
- **Elaine Chun**
- **Jennifer S. Cole**
- **Megan Crowhurst**
- **Ellen Kaisse**
- **Chris Kennedy**
- **Ken Latta**
- **John Lawler**
- **Diane Lillo-Martin,**
- **Pieter Muysken**
- **Alyson Reed**
- **Ian Roberts**
- **Joe Salmons**
- **Bridget Samuels**
- **George Walkden**
- **Natasha Warner**

Aggregating the answers

- 1. The term(s) by which you refer to it

data; data analysis files; stimuli; experimental data, perception data; corpus, texts, archival data; corpus if it's spontaneous speech, acoustic recordings, sound files

- 2. The nature of the data (e.g. recorded speech, transcribed sentences, sound samples, stimuli, …)

audio files, transcriptions, demographic information, extracted acoustic measures, recorded speech (can include label files, transcriptions, acoustic measurements), data analysis files (including responses and statistics), internet videos, internet written comments, transcripts (recorded interactions/interviews, collected videos), sound samples, questionnaires, translation tasks, acceptability judgments, manuscript text (part-of-speech tagged, lemmatized and parsed), elicitation field notes (including target language material, glosses, fieldworker comments, date elicited, original/microfilm page numbers), experimental subjects' responses (including reaction times, and typed or mouse-click responses)

Aggregating the answers

- 3. The units of analysis (e.g. conversations, words, sounds, sentences, gestures, etc.)

linguistic units: audio "snippets" (word, phrase, or sentence); sounds, words, gestures, nonce words, words lists; phrases, sentences, conversations, bits of conversation, conversational turns, comments, genres; longer texts (stories, letters); responses to sounds, words, or sentences (e.g. lexical decision, word spotting, phonetic identification)

sociocultural units: stances, alignments, footing, social identity dimensions, miscellaneous sociocultural factors

- 4. The digital form of storage (e.g. pdfs, mp3 files, jpegs, txt files, spreadsheets, etc.)

audio: .wav, .aiff, .mp3, .mp4

text: .txt, .rtf, .edat (EPrime),.pdf, .FLEx (Field Language Explorer), .psd (UTF-8 text according to Penn conventions), .doc, .docx, .trs, TextGrid

other: .xls, .xlsx, .htm, .html, .xml, .csv

Aggregating the answers

- 5. The size (in megabytes or gigabytes) of a typical file

Varies widely; no typical size; varies insanely; wav files (50 KB – 2 GB); mp4 files (50-100 MB); xlsx spreadsheets (1-20 MB); docx files (less than 1MB)

- 6. The most typical manner of storage (CDs, hard drives, cloud, tape, index cards, printed or typed pages)

University servers, cloud, computer hard drives, USB drives, CDs, linguistic archives, paper records

Some citation resources

- Child Language Data Exchange System (CHILDES): http://childes.psy.cmu.edu/

  CHILDES citation policy: http://talkbank.org/share/rules.html

- Corpus NGT (an open access online corpus of movies with annotations of Sign Language of the Netherlands – abbreviated as SLN or NGT): http://www.ru.nl/corpusngtuk/

  Citation policy for the NGT corpus:
  http://www.ru.nl/corpusngtuk/using_the_corpus/creative_commons/

- Linguistics Data Consortium (LDC): http://www.ldc.upenn.edu/

  Citation and naming in the LDC:
  https://www.ldc.upenn.edu/data-management/citing
  https://www.ldc.upenn.edu/data-management/providing/filenames-metadata

- World Atlas of Language Structures (WALS): http://wals.info/

  WALS Online (citation etiquette):
  http://listserv.linguistlist.org/pipermail/lingtyp/2008-November/002428.html

Other archives

- **Gene Buckley**

"… when I refer in my own notes to my fieldwork, I have a format that includes the date and speaker, plus the time stamp of the recording in some cases. (The audio files are typically divided into morning and afternoon, so that is included also.) The snippet referred to is a word, phrase, or sentence. My internal references are quite abbreviated. For example, the material I elicited in my fieldwork trip in July 2013 is in a FieldWorks Language Explorer 'text' called 'F 13-7'. I have one such 'text' for each of my six recent trips. The individual sentences have references like '2013-07-15am (AS)' where AS is the speaker's initials, and it was the morning of July 15, 2013. That matches the name of the audio file, 2013-07-15am.wav. Where I noted the time, it will be '2014-05-26pm - 1:14 (AS)', i.e. at 1 hour and 14 minutes on the recording made the afternoon of May 26, 2014. In other cases, when I was working in Praat, I have the number of seconds instead: '2012-09-13pm - 1309s (AS)'."

- **Ken Latta**

Lüpke, Frederike. 2009. Data collection methods for field-based language documentation. In Peter K. Austin (ed.), *Language Documentation and Description* 6:53-100. London: SOAS. http://www.elpublishing.org/docs/1/06/ldd06_04.pdf

Other archives

- **Bridget Samuels**

"I have been involved in developing the data repository "hub" for a consortium of developmental biologists focusing on the craniofacial region."

FaceBase - Comprehensive data and resources for craniofacial researchers: http://facebase.org

- **George Walkden**

Corpus of Historical Low German (CHLG): http://www.chlg.ac.uk/helipad/index.html

Desiderata for general citation

- Who? Author, aggregator (aggr.), editor (ed.), data generators (gen.)
- When? Time of collection (e.g. rec. or coll.), archival/publication date
- What? Title of collection, or descriptive label
- Category? Experimental stimuli, notes, sentences, recorded speech
- Enhancements? Parsed, tagged, transcribed, labeled for properties (e.g. experimental conditions)
- Length? Number of (i) pages, (ii) minutes:seconds, or (iii) items.
- Format? Application and file extension used to create it.
- Part of? Published article or book, named research project or archive
- DOI or URL?

Desiderata for in-text datum citation

- Page number?

  Internal/external (e.g., p. 17[45], pp. 17-19[45-50])

- Item number?

  E.g. Stimulus 45b, Sound file A-57.

- Location in sound/video file?

  E.g. Start/4:55, Run/4:55-5:07

Examples

- Bibliographic citation

Springstein, Bruce. 2013[Coll:14-VI-2005]. Informal Balinese conversations [Gen: Dewa Made Beratha and Anak Agung Bagus]. Transcribed typed text in Sociolinguistic parameters of Balinese discourse [NSF Grant 1261948], Field notes volume 5. 134[268] pp. Adobe Acrobat/.pdf [scanned and OCR]. URL:http://people.cas.sc.edu/smith/NSFgrant1261948/fieldnotes5.pdf DOI:10.1002/0470841559.vol5

- In-text citation

Springstein 2013:71[142]/lines 23-25

Examples

- Bibliographic citation

Warwick, Dionne. 1995[Rec:30-IX-1989]. Elicited telephonic greetings in Jamaican call center [Gen: anonymous subjects]. Recorded sound files associated with “Analysis of free-form answers in Jamaican English Creole”, *CALICO Journal* 12(2/3): 59-87. Length: 15:38. Sony Sound Forge/.wav Stable URL: http://www.jstor.org/stable/24152780

- In-text citation

Warwick 1995[Recorded data]:03:28-04:01