PhonBank

Behind the Scenes

Carla Peddle

PhonBank: Behind the Scenes

Outline

- Sneak peak into what goes on behind the scenes of PhonBank
- Accomplishments we have made
- Challenges we face; and
- Improvements for the future

Phon & PhonBank: Behind the Scenes

Phon and PhonBank are already being used in the field of language acquisition

Before the software and data are released to the field there is a lot of work behind-the-scenes:

- developing software
- testing software
- preparing data for PhonBank

Phon & PhonBank: Behind the scenes

1. Work related to Phon development:
   feature set (identify all characters in the field)
   dictionaries (English, French, Catalan …)

2. Testing the application:
   segmentation
   multiple-blind transcription
   syllabification & alignment
   inventory functions

3. Manual:
   writing & editing
   implementing changes with Phon updates

4. Big work → preparing PhonBank Projects

Phon & PhonBank: Behind the scenes

Phon is designed to handle the entire workflow associated with new child language data (from segmentation to searching)

The main goal of PhonBank is to acquire existing child language data and share them with the field

Optimally data should:

- be well transcribed
- have clear media recordings

Phon & PhonBank: Behind the scenes

Many team members play different roles to make Phon & PhonBank efficient

Yvan & Greg	mostly Phon
Brian	mostly PhonBank
Carla	on middle grounds between Phon and PhonBank

PhonBank: Behind the scenes

My work happens at MUN

while Yvan travels

- promote Phon to researchers; and
- recruit new research contributors to PhonBank

PhonBank: Behind the scenes

New research contributions create more work:

1. Specific research questions

    ↳ changes to the application

        ↳ more testing

2. Nearly all new data are formatted to comply with the exacting standards of PhonBank xml

PhonBank: Behind the scenes

With large influxes of work, we hire student research assistants

Most of the PhonBank work is basic but demands:

- patience
- diligence; and
- attention to detail

Looking at the bigger picture

VERY REWARDING!

PhonBank: Behind the scenes

Project Name	Original Format
Dutch-CLPF	ChildPhon
Dutch-Zink	LIPP
English-Davis	LIPP
English-Inkelas	Excel
English-Stanford	CHILDES
French-Kern	LIPP
French-Stanford	CHILDES
German-Stuttgart	WaveSurfer
German-TAKI	WaveSurfer
Japanese-Ota	Excel
Japanese-Stanford	CHILDES
QcFrench-GoadRose	ChildPhon
Romanian-Kern	LIPP
Swedish-Stanford	CHILDES
Tunisian-Kern	LIPP

Since each project is unique and the original formatting of the projects differ, there is a distinct set of steps involved with preparing each project for PhonBank

PhonBank: Behind the scenes

Ultimate Goal: have compatible CHAT and Phon files for all of the PhonBank projects

1. Convert all data to the CHAT format
2. Subject data to full quality control through CHAT2XML verification system
3. Align any audio to transcript at the utterance level:
   - accurate playback
   - acoustic analysis
4. Import projects into Phon

PhonBank: Behind the scenes

Most original transcript formats are not compatible with PhonBank:

- ❑ LIPP
- ❑ ChildPhon
- ❑ Excel
- ❑ WaveSurfer

In most of these cases, Brian is the first to work on converting non-CHAT data into the CHAT format

PhonBank: LIPP

LIPP projects:

- Dutch-Zink
- English-Davis
- French- Kern
- Romanian-Kern
- Tunisian-Kern

Brian converts LIPP files to CHAT files

freb01.lipp → freb01.cha

Brian makes CHAT files available to the MUN team

PhonBank: LIPP

Once the MUN team receives CHAT and media files:

- ensure one-to-one correspondence
- rename one or both set of files

All files for a session have:

- same file name
- different file-type extension

freb01.cha	→	freb01.cha
emma 001 23-06-01.mpg	→	freb01.mpg

PhonBank: LIPP

Large media files are not always manageable within PhonBank

Convert large media files:

1. Open large .mpg media files in the MPEGStreamclip application
2. Export to .mp4 format

freb01.mpg

↘

MPEGStreamclip

↘

freb01.mp4

PhonBank: LIPP

Using the CLAN application

Linking: the painstaking process of listening to endless hours of media, most often of screaming children, in order to make associations between portions of a media file and corresponding utterances in a transcript

Identify start and end time values for small portions of media for utterance playback

PhonBank: LIPP

Import linked CLAN transcripts into Phon:

CHAT2XML

- exports CHAT data to an XML file
- identifies issues preventing the creation of a matching file

XML2Phon

- imports new XML files into Phon

PhonBank: ChildPhon

ChildPhon projects:

- Dutch-CLPF
- QcFrench-GoadRose

Two unrelated applications coincidentally called ChildPhon:

1. Levelt & Fikkert used 4th Dimension based software
2. Goad & Rose used FileMakerPro based software

Yvan converts the ChildPhon projects into Comma Separated Value (CSV) files

PhonBank: ChildPhon

Dutch-CLPF has sets of media clips for each session

- One-to-one correspondence between number media clips and the number of records per session
- Merge media clips by session
- Export the time values at junctures using Amadeus Pro
- Use the juncture values as start & end times for media clips
- Enter start & end time into the CSV files

PhonBank: ChildPhon

The next step is to prepare the CSV files for import into Phon

- Uniform column headers across the project
- Properly formatted content cells
- Replace ASCII characters with the Unicode equivalents

Greg imports CSV data into Phon

PhonBank: Excel

Excel projects:

- English-Inkelas
- Japanese-Ota

Brian is the first to work on converting projects into CHAT files

The MUN team uses the CLAN application to link Japanese-Ota CHAT files to the corresponding media files as with the original LIPP projects (Kern, Davis, etc.)

PhonBank: Excel

The English-Inkelas project came to the MUN team as one large CHAT file with data for several recording sessions

1. Split CHAT file by date into 200 smaller session files

2. Check CHAT files against the original Excel file

Import both of the projects into Phon using CHAT2XML and XML2Phon

PhonBank: WaveSurfer

WaveSurfer projects:

- German-Stuttgart
- German-TAKI

Brian converts the WaveSurfer files into the CHAT format

The CHAT files go to the MUN team

Import projects into Phon using CHAT2XML and XML2Phon

PhonBank: Existing CHILDES projects

Existing CHILDES projects:

- English-Stanford
- French-Stanford
- Japanese-Stanford
- Swedish-Stanford

Brian makes existing CHAT files available to the MUN team

Import projects into Phon using CHAT2XML and XML2Phon

PhonBank: Additional Work

We have also worked on other projects which are not yet available from the PhonBank directory:

1. MCF – Portugese-Swedish-English trilingual data
2. Chiat – English clinical data on velar fronting
3. English-Smith – diary study data without media

PhonBank

Once all project files have been imported into Phon we upgrade the projects with:

- □ Addition of generic IPA Target forms
- □ Correction of rogue characters
- □ Adjustment of media linkage
- □ Verification of syllabification and alignment data for the IPA Target and Actual

PhonBank

After a series of spot checks between the original project files and the Phon files, they are ready for:

✓ Automated searching

✓ Tracking individual queries

✓ Exporting data sets

✓ Reporting; and

✓ Sharing via PhonBank

PhonBank: Accomplishments

Most of my work over the last four years:

- Linking
- Training student RAs to link; and
- Supervising student “linkers”

MUN team has linked more than 1000 sessions, most with media files more than one hour in duration

For each hour of media we spend more than three hours linking

Literally thousands of hours of linking

PhonBank: Accomplishments

15 PhonBank projects ready with the release of Phon 1.4

Encompass:

- ◆ 8 languages
- ◆ 87 participants
- ◆ Nearly 2000 recording sessions

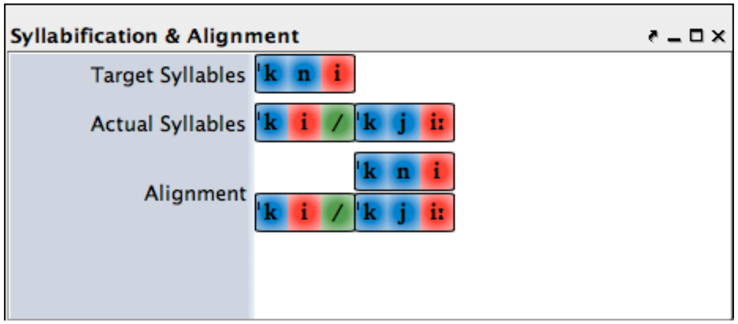
Projects are available for download or browsing on the PhonBank portion of the CHILDES database

http://childes.psy.cmu.edu/

PhonBank: Challenges

Data Formatting

- Several researchers and data formats creates a challenge for making projects comparable
- Character compatibility issues arise between old and new versions of the projects
- Rogue characters cause problems in the transcripts



Syllabification & Alignment

Target Syllables ˈk n i

Actual Syllables ˈk i / ˈk j iː

Alignment ˈk n i
ˈk i / ˈk j iː

PhonBank: Challenges

Media Issues

- Laughing, crying or overlapping participants' speech makes it difficult to hear, segment, transcribe and link

Overlap: MCF-ksm

Distance of Research Contributors

- Difficult to exchange materials
- Time difference hinders communication
- Data may be worked on by several people at once

PhonBank: Potential improvements

- ✧ Standardized transcription conventions for all converted corpora
  - ✧ Any changes must maintain the spirit of original corpus

- ✧ Corpus versioning, to assist further data annotation without overwriting each other's work

PhonBank: Behind the scenes

Thank you very much!

Questions?
Comments?