An understanding of developmental production patterns requires a multi-faceted analysis incorporating:
- Perceptual factors
- Articulatory effects
- Grammatical organization

We thus need:
- A strong empirical approach to child language studies (a lot of data!!!)
- A refined system for detection of phonological patterns within and across positions/contexts

The PhonBank Project

Yvan Rose – MUN
Brian MacWhinney – CMU
75+ colleagues and 6 coworkers

Two Core Principles

Problems: Data Lockup, Format Babel
Solutions: Full Access, Full Structure

Open access
- Levels of permission
- Web-based access, streaming, commentary

Full structure
- Tight definitions in XML
- Full automatic checking
- Multiple frameworks
- Interoperability
**CHILDES, TalkBank, PhonBank**

<table>
<thead>
<tr>
<th></th>
<th>CHILDES</th>
<th>TalkBank</th>
<th>PhonBank</th>
</tr>
</thead>
<tbody>
<tr>
<td>Age</td>
<td>24 years</td>
<td>8 years</td>
<td>3 years</td>
</tr>
<tr>
<td>Words</td>
<td>47 million</td>
<td>8 + 55 million</td>
<td>.5 million</td>
</tr>
<tr>
<td>Media</td>
<td>1.5 TB</td>
<td>1 TB</td>
<td>.5 TB</td>
</tr>
<tr>
<td>Languages</td>
<td>32</td>
<td>18</td>
<td>7</td>
</tr>
<tr>
<td>Publications</td>
<td>3000+</td>
<td>89</td>
<td>4</td>
</tr>
<tr>
<td>Users</td>
<td>2500</td>
<td>500</td>
<td>68</td>
</tr>
</tbody>
</table>
Technical overview

- Programmed in Java
- Main computer platforms supported
- Unicode compliant
- XML (TalkBank) data structure
- Working toward compatibility with other TalkBank-compliant applications
- 100% IPA support
- Most features integrated within a unique interface
- Open-source

Workflow supported

- Search results visualization and management
- Project management
- Corpus query
- Syllabification and alignment
- Data transcription
- Media linkage and segmentation

Other Useful Features

- Integrated support for multi-blind transcriptions
- Multiple undos
- Text search and replace
- User-defined tiers
- Data copy to clipboard (tier, record, session)
- XML and CSV data import and export
- ...
**Latest update: PhonBank**

**PhonBank additions**
- Additions of corpora
  - Babbling data (Arabic, English, French, Romanian, ...)
  - Word production data (English, French, Japanese)
  - More to come!!
  - You happen to have some data to share?
- Transcript data linkage to digitized recordings
  - Many of the corpora currently in CHAT format
  - (Conversion to Phon format later this spring)
- For details: http://childes.psy.cmu.edu/data/PhonBank/

**General functionalities**
- System to import CHILDES data (previously converted through CHAT2XML)
General functionalities

- Addition of new built-in dictionaries of pronounced forms
  - Dutch
  - German
  - Icelandic
  - Italian
- New XML file format
  - Eliminates a lot of session visualization issues
  - (Step required for later incorporation of Praat data)
- Elimination of built-in media player
  - Simpler, more flexible approach to segmentation

Revamped Session Editor

- Overhaul of the session editor
  - Performance:
    - Much faster!!!
    - Support for larger session sizes
    - Better use of screen real-estate
  - Syllabification
    - Ability to select specific algorithms for each IPA
      Target or Actual form
    - Key-based modification of syllabification
      (e.g. O=Onset; N=Nucleus; ...)

Brand New Search Engine

- Search ‘plugin’ system
  - Ability to create own search plugin on the fly without a need to reprogram the application
  - Based on Java (ECMA) scripting
- Built-in scripting editor
- Persistent search results
  - Queries and results saved in a relational database
  - Support for *large* result sets
  - Access to query history
  - (Query tagging not completed)

Data visualization and reporting

- Visualization:
  - Search results integrated with the session editor
  - Session opens as results are displayed
  - Results summaries
    - Clipboard accessible
- Reporting (we are about to tackle this issue...)
  - We will support the functionality currently existing in Phon 1.3
  - We would like your input about other report content/format
Phon 1.4: unfinished work

- Graphical User Interface overhaul
  - Editor (final touches)
  - Search results history and visualization
- Media playback and export functions:
  - Native (stable!!!) media engines
  - Also a component required for integration with Praat
  - Segment playback + adjustment
- Reporting functions (as mentioned already)

From there: Phon ↔ Praat integration

- Integration between Phon and Praat through TextGrid
  - Export record to TextGrid; extract relevant media clip
  - All acoustic analyses done from within Praat
  - System to store Praat-generated (acoustic measurement) data into Phon XML structure
  - System to extract acoustic management data relative to phonological contexts using Phonex query system
    - E.g. Get F0, Intensity and Duration data for all mid vowels in word-initial, stressed syllables
    - Data exported for further processing

Phon ↔ Praat integration: issues

- Under investigation
  - Ability to import existing TextGrid data into Phon
  - Ability to communicate with other application that are themselves compatible with TextGrid
    - This is much more tricky than it may seem
    - We will probably need input from users/other programmers to achieve this challenge
  - Let’s discuss the potential implications together...

Longer Term Goal

- CLAN, ELAN, SFS...
  - Phonological Data
  - Acoustic Data
  - CLAN, ELAN, SFS...
Acknowledgements

Funding: Current development of Phon and PhonBank is supported by the National Institute of Health. Earlier development of Phon was funded by grants from National Science Foundation, Canada Fund for Innovation, Social Sciences and Humanities Research Council of Canada, Petro-Canada Fund for Young Innovators, and the Office of the Vice-President (Research) and the Faculty of Arts at Memorial University of Newfoundland.

Dictionaries: Built-in dictionaries of pronounced forms were obtained from several organizations:

- English: CMU Pronouncing Dictionary (Speech at CMU) http://www.speech.cs.cmu.edu/speech/
- French: Lexique Database http://www.lexique.org/
- Catalan and Spanish: Universitat Politècnica de Catalunya http://www.upc.es/

Special thanks: We owe special thanks to a wonderful group of early adopters and beta testers. These include researchers and PhD students from Universidade da Lisboa (Laetitia Almeida, Susana Correia, Teresa da Costa, Maria João Freitas); Universitat Autònoma de Barcelona (Joan Boràs Comes, Ana Estrella, Maria del Mar Vanrell, Pilar Prieto, Jill Thorson); Center for Advanced Research in Theoretical Linguistics (Helene Nordbärd Andreasen, Berit Anne Bals, Bruce Morén); Universiteit Leiden (Claartje Levelt); Radboud Universiteit Nijmegen (Paula Fikkert, Nicole Altvater-Mackensen); Ecole Normale Supérieure (Aliyah Nordgaard Andreassen, Berit Anne Bals, Bruce Morén); Universiteit Leiden (Claartje Levelt); Radboud Universiteit Nijmegen (Paula Fikkert, Nicole Altvater-Mackensen); Ecole Normale Supérieure (Aliyah Nordgaard Andreassen, Berit Anne Bals, Bruce Morén); Universiteit Leiden (Claartje Levelt); Radboud Universiteit Nijmegen (Paula Fikkert, Nicole Altvater-Mackensen); Ecole Normale Supérieure (Aliyah Nordgaard Andreassen, Berit Anne Bals, Bruce Morén); Universiteit Leiden (Claartje Levelt); Radboud Universiteit Nijmegen (Paula Fikkert, Nicole Altvater-Mackensen); Ecole Normale Supérieure (Aliyah Nordgaard Andreassen, Berit Anne Bals, Bruce Morén); Universiteit Leiden (Claartje Levelt); Radboud Universiteit Nijmegen (Paula Fikkert, Nicole Altvater-Mackensen); Ecole Normale Supérieure (Aliyah Nordgaard Andreassen, Berit Anne Bals, Bruce Morén); Universiteit Leiden (Claartje Levelt); Radboud Universiteit Nijmegen (Paula Fikkert, Nicole Altvater-Mackensen); Ecole Normale Supérieure (Aliyah Nordgaard Andreassen, Berit Anne Bals, Bruce Morén); Universiteit Leiden (Claartje Levelt); Radboud Universiteit Nijmegen (Paula Fikkert, Nicole Altvater-Mackensen); Ecole Normale Supérieure (Aliyah Nordgaard Andreassen, Berit Anne Bals, Bruce Morén); Universiteit Leiden (Claartje Levelt); Radboud Universiteit Nijmegen (Paula Fikkert, Nicole Altvater-Mackensen); Ecole Normale Supérieure (Aliyah Nordgaard Andreassen, Berit Anne Bals, Bruce Morén); Universiteit Leiden (Claartje Levelt); Radboud Universiteit Nijmegen (Paula Fikkert, Nicole Altvater-Mackensen); Ecole Normale Supérieure (Aliyah Nordgaard Andreassen, Berit Anne Bals, Bruce Morén); Universiteit Leiden (Claartje Levelt); Radboud Universiteit Nijmegen (Paula Fikkert, Nicole Altvater-Mackensen); Ecole Normale Supérieure (Aliyah Nor