Research in the 21st Century: Big and Precise Data

- **Big Data** computing offers new and exciting opportunities for research
- Potential to reveal behavioural patterns, trends, and associations in language acquisition and use
- **Precise Data** remains of primary importance for several types of linguistic research
  - Advances in our field also come from increasingly detailed data descriptions and assessments
- **Combining Big and Precise Data**
  - Collaborate to build and share large databases
  - Invent new specialized tools for data analysis

---

CHILDESEhttps://childes.talkbank.org

- **Child Language Data Exchange System**
  - Where everything started!
    - 1983 - Brian MacWhinney & Catherine Snow
- Largest repository of child language data worldwide
- **CLAN** program for lexical and morpho-syntactic analysis

Missing from CHILDESE: support for phonetics and phonology...

---

PhonBank (& Phon)https://phonbank.talkbank.org

- **PhonBank**
  - Initially: component of CHILDESE in the areas of phonological development and disorders
  - Now welcoming all areas of corpus-based research on phonetics and phonology
- **Phon**
  - Software for building and analyzing phonological corpora
  - Transcript-based phonological analysis
  - Acoustic analysis
  - Clinical analysis
Contributions (2008-2018)

- 50/58 corpora include audio/video
- Together, these corpora document:
  ≈ 7500 recorded sessions
  ≈ 1125 children at different ages and with various language backgrounds

In the pipeline: Child L2, German (typical, disordered), Italian, Icelandic, Portuguese (typical, Down), Turkish, Vietnamese, …

Got data? Bring it on!!

Contributions (2008-2018)

- Data corpora document:
  • Typical language development (monolingual, bilingual, L2 acquisition)
  • Speech-language disorders
  • For a few corpora: TextGrid annotations toward acoustic analysis
- Theses and publications
  • As of January 2016 we have inventoried over 600 scientific references based on PhonBank data

Much more work needs to be done…
  • More languages need to be documented
  • More speakers and speaker populations are needed for most of the languages already documented
  • Clinical and L2+ acquisition data

Getting the most out of big and precise Data

Problem: General data don’t tell us much more than what we already know from research or practice

Solution: We need open access to powerful yet usable systems for data analysis
Development philosophy

- Tension between functionality and usability
  - We need to measure different types of data in uniform ways
  - How can we make software that is both maximally powerful and decently usable?
- Some solutions
  - Integrate as many possibilities for data coding and analysis as possible within a single logical framework
  - Data annotations structured on different ‘levels’
  - Relational analyses informed data annotation structure
- Empower the user
  - Full workflow support
  - Intuitive user interface
  - Scripting!

Scripting languages differ from system programming languages in that they are designed to "glue" applications together


Phon: overview

- Database system for phonological/phonetic corpus building and analysis
  - Transcription-based (phones, features, positions)
  - Acoustic (using Praat libraries)
- Phon can be used in virtually all types of corpus-based research on phonetics and phonology
  - e.g. L1, L2+ acquisition, loanwords, socio-phonetics, fieldwork, ...

Data preparation: from raw to multifaceted, structured data

- Transcriptions and annotations structured and referenced to maximize data output
  - Interface to help data transcription (e.g. IPA map; IPA dictionaries; …)
- Hierarchical data structure
- Automatic algorithms (syllabification, alignment)
- Agile corpus building (Voormann & Gut 2008)
- Preliminary datasets analyzable with minimal efforts
- Additional annotations added only when needed

Phon automatically structures and annotates transcript data

Position within phrase/utterance
- Initial, medial, final

Phonological features
- Labial, nasal, stop, ...

Syllable- and word-level info
- (syllable position, stress, tone, …)

Phone-by-phone alignment between target (model) and actual forms

Word-by-word alignment across annotation tiers
- Words on one tier relate to encodings (e.g. grammatical categories) on other tiers
Make queries through fillable forms and minimal scripting

- Look for: /s/ non-word-final codas
  [scripting “s:c”]
  [check buttons]
- Report data from: Interview participant
  [dropdown menu]
- Add: Orthographic and IPA data to each result
  [list tiers]
- Restrict query to: Nouns
  [enter tier name]
  [enter filter criterion]

Manage query results through the Report Builder

- Interface to ‘compose’ a data report
- Combine individual reports (e.g. aggregate, table, listing) within a single, global report
- Example report:
  - Duration and Formant measurements (based on speech acoustics)
  - Accuracy measure on production of vowels (based on IPA transcriptions)
  - Composed reports can be saved for later use, sharing with students, collaborators, ...

Access standard analyses

- Standard analyses (e.g. PCC, PMLU, frequency counts; phonological processes) have traditionally been cumbersome to compute
- Limits on the number of tokens that can be humanly analyzed
- Limits on reliability
- It's just no fun...
- Phon offers full solutions through ready-to-use combinations of specialized algorithms and purpose-specific data reports

The Analysis Composer

- Compose a multifaceted analysis from various queries and built-in analyses
- Composed analyses can be saved for later uses
- Replicability: Analyze more data with consistency
- Share with:
  - students
  - colleagues
  - ...
Getting data and analysis results out of Phon

• Three ways to save reports from queries and analyses
  • Print-ready (HTML format)
  • Excel spreadsheet/workbook formats
  • CSV (useful for importation into R, SPSS, LibreOffice, …)
• TextGrid data generated / processed in Phon
  • 100% Praat-compatible
  • Can be used outside of Phon
• Phon data transcripts can be exported as text (CSV) or CLAN (CHAT) formats

From here...

• Funding until the end of 2021
• Planned support for:
  • Automatic diarization (time stamping and speaker identification) using ASR technologies
  • Forced alignment for TextGrids
  • More data mining methods
  • Integration of graphing functions for data reports
• Continued efforts to improve usability
• Expanded documentation and resources for all users

Future potential

• Adding more languages and documenting more areas of phonological and phonetic development
  • L2, L3, Trilingual, Syndromes, …
• Further expansions in other areas of research on phonetics and phonology
  • Sociophonetics (Hernandez & Rose 2019)
  • Forensics (Rose 2017)
  • Language Documentation (Rose & Brittain, to appear)
• Connecting to other Banks
  • FluencyBank, AphasiaBank, HomeBank (among others) are all natural fits for phonology

Thank you!

With very special thanks to our very amazing Guru

Questions, feedback:
yrose@mun.ca
Useful websites:
• Phon - https://www.phon.ca
• PhonBank - https://phonbank.talkbank.org/