Data Analysis in Phon: Where are we now and where should we go?

Greg Hedlund[1] and Todd Wareham[2]

Departments of [1]Linguistics and [2]Computer Science
Memorial University of Newfoundland

July 30, 2010

Joint work with:

- Jason Gedge (University of Alberta)
- Yvan Rose (MUN)

Introduction: Dream a Little Dream with Me . . .

- Previous work on “automatic” features in Phon has largely focused on basic (pre-)processing of input data, *e.g.*, syllabification, alignment.
- Focus here on current and potential Phon data analysis capabilities.
    - Phrase in terms of pattern matching and derivation.
    - Emphasis in this talk is on capabilities, not algorithms – let’s dream about what would be useful, and not censor ourselves with what we’ve seen before or what we think might be doable.
    - Don’t worry about how these capabilities will be implemented with respect to Phon – again, let’s focus on capabilities.

Organization of this Talk

1. Data Analyses in Phon
2. Pattern-Based Analysis: A General Framework
3. Potential Data Analyses using Phon
4. Conclusions

Data Analyses in Phon: Overview

- Fundamental unit of data storage is a session; sessions can be grouped into longitudinal time-series.
  - A session consists of information about time, place, and participants and one or more tiers of speech-data for each participant.
  - A session time-series consists of one or more sessions involving a common group of speakers that are ordered in time.
- Three phases to data analysis:
  1. Create Query (specify pattern)
  2. Create Search Results (match pattern)
  3. Create Reports (report match results)

Data Analyses in Phon: Creating Queries

What types of patterns do we need to look for?

- **Basic text searching:** Find an instance of a particular string or regular expression.
- **Aligned groups:** Find string patterns across tiers which have been aligned with groups created in Orthography.
- **Aligned phones:** Find instances of various processes, *e.g.*, match, epenthesis / deletion, substitution, metathesis, harmony.
- **Word / syllable types:** Find instances of morphological patterns, *e.g.*, stress patterns, CV(G) sequences.
- **Attributes:** Find instances by entity properties, *e.g.*, session date, participant name / age, language spoken, etc.

Data Analyses in Phon: Creating Queries (Cont'd)

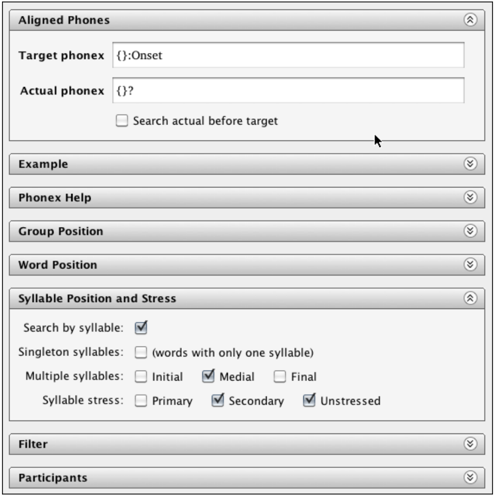
Seven basic types of queries are provided in the application:

- **Text Searching** (`Data Tiers.js`)
- **Aligned Groups** (`Aligned Groups.js`)
- **Word / Syllable Types** (`CV Sequences.js`, `Word Shapes.js`)
- **Aligned Phones** (`Aligned Phones.js`, `Metathesis.js`, `Harmony.js`)

Data Analyses in Phon: Creating Queries (Cont'd)

Each query form has options particular to its function, as well as options for specifying:

- Syllable / Word / Group position (time-domain within utterance).
- Syllable stress.
- Speaker name and age.
- Custom patterns based on user-defined data tier.



Aligned Phones

Target phonex: {}:Onset

Actual phonex: {}?

□ Search actual before target

Example

Phonex Help

Group Position

Word Position

Syllable Position and Stress

Search by syllable: ☑

Singleton syllables: □ (words with only one syllable)

Multiple syllables: □ Initial ☑ Medial □ Final

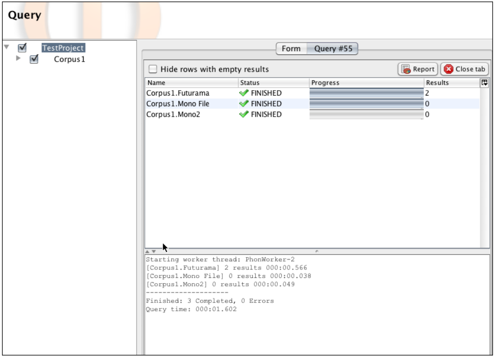
Syllable stress: □ Primary ☑ Secondary ☑ Unstressed

Filter

Participants

Data Analyses in Phon: Creating Search Results

- Queries are executed on one or more selected sessions.
- Search results are stored on disk in a relational database.
- Some queries may print additional information or error messages in the displayed console.



Name	Status	Progress	Results
Corpus1.Futurama	✓ FINISHED		2
Corpus1.Mono File	✓ FINISHED		0
Corpus1.Mono2	✓ FINISHED		0

```
Starting worker thread: PhonWorker-2
[Corpus1.Futurama] 2 results 000:00.566
[Corpus1.Mono File] 0 results 000:00.038
[Corpus1.Mono2] 0 results 000:00.049
--------------------
Finished: 3 Completed, 0 Errors
Query time: 000:01.602
```

Data Analyses in Phon: Creating Reports

- Viewing results within the application
  - Results are highlighted as they are selected, allowing review.
  - Allows deletion of individual results; especially useful for searches that may return false positives, *e.g.*, metathesis, harmony.
- Exporting results in printable format (`pdf, html, odt, xls`)
  - Report is broken into configurable sections providing inventories, result lists, comments, and summaries.
  - Provides more useful information than CSV export (below) and is extendible, *e.g.*, add new report sections..
- Exporting results in format usable by other applications (`CSV`)
  - Can select what columns are exported and their ordering;
  - Can only export matched values, – at present, no export of inventory counts or derive data (though this may change in future).

Data Analyses in Phon: Over the Rainbow

- Many neat questions are currently hard to answer, *e.g.*,
  - Does speaker X have phone-acquisition order Y?
  - Do the (majority of) speakers in $\mathcal{X}$ have phone-acquisition order Y?
  - Does speaker X have the same phone-acquisition order as the speakers in $\mathcal{X}$?
  - Is the acquisition of phone a correlated with accurate production of syllable-form b in the speakers in $\mathcal{X}$?

  - What is the phone-acquisition order of speaker X?
  - What is the (consensus) phone-acquisition order of the (majority of) speakers in $\mathcal{X}$?
  - What are the subpopulations of the speakers in $\mathcal{X}$ with respect to phone-acquisition order?
  - What aspects of syllable-structure are correlated with the acquisition of phone b in the speakers in $\mathcal{X}$?

. . . Can we do better? . . .

Pattern-Based Analysis: A General Framework

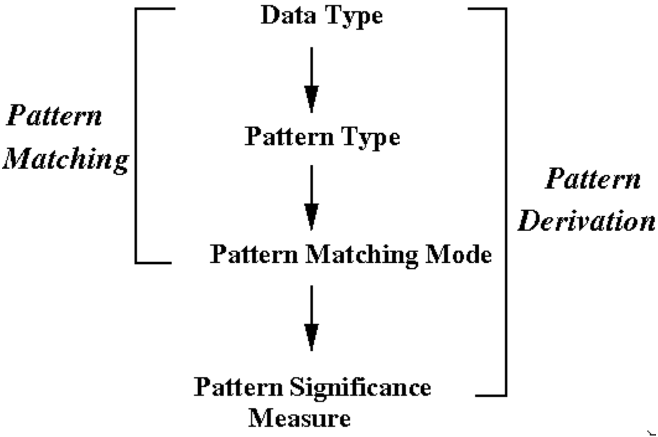
- Pattern matching vs. pattern derivation:

  Pattern Matching: Get occurrences of pattern P in text T.
  Pattern Derivation: Get set of significant patterns $\mathcal{P}$ that occur in set of texts $\mathcal{T}$.

- How is this relevant to linguists?

Pattern	$\Leftrightarrow$	linguistic hypothesis
Pattern matching	$\Leftrightarrow$	verifying specified hypothesis against specified data
Pattern derivation	$\Leftrightarrow$	determining hypotheses that are well-supported by specified data

Pattern-Based Analysis: A General Framework (Cont'd)



Potential Data Analyses using Phon: Data Types

- In Phon, data currently stored as sessions and session time-series; can also group these into corpora.
- Could also store and operate on data that summarize individual sessions or groups of sessions , *e.g.*,
  - Set of distinct items in a session (produced phones, word-form CV-types)
  - One or more frequencies

  Such summarized sessions may in turn be ordered to make summary session time-series.
- Could also transform time-dimension, *e.g.*, absolute → MLU.

Q1: What are linguistically useful types / summaries of Phon data?

Potential Data Analyses using Phon: Pattern Types

- In Phon, a pattern is currently a segment (possibly across several aligned tiers) in an individual session; using a regular expression, can look for any of a set of segments encoded by that expression.
  - Pattern also includes attributes (speaker name / age-range, etc) that regulate / further restrict instances of segment-match.

  Such patterns are time-series over tiers in individual sessions.
- Could also specify richer types of patterns, *e.g.*,
  - Time-series over (possibly summarized) session time-series (acquisition-order of attempted consonant clusters, frequencies over time of accurately-produced syllable types)
  - Correlations (two or more segments that always co-occur within an individual session or across sessions).

Q2: What are linguistically useful types of patterns?

Potential Data Analyses using Phon: Pattern Matching Modes

- Specify match of pattern P and text T by function $match(P, T)$ which returns rating of similarity of P and T; may also return alignment of corresponding elements in P and T.
- Matches can be exact or approximate.
- In Phon, patterns are currently only matched exactly.
- Many flavors of approximate matching, *e.g.*, approximate match of corresponding-element values, altered temporal spacing and/or ordering of corresponding elements. Moreover, when deriving patterns relative to a set of texts, patterns may also occur exactly (in all texts) or approximately (in some proportion of the texts, with some frequency in each text).

Q3: What are linguistically useful pattern matching modes?

Potential Data Analyses using Phon: Measures of Pattern Significance

- When deriving patterns, there are typically many patterns that are common to a group of texts; select relevant patterns using some measure of significance, *e.g.*,
  - Length / complexity of pattern
  - (Minimum / maximum) degree of pattern match
  - Proportion of texts exhibiting pattern
  - Strength of correlation (for correlation-patterns)

Q4: What are linguistically useful measures of pattern significance?

Potential Data Analyses using Phon: Meta-Pattern Analyses

- Could use pattern-matching function $match()$ to assess degree of similarity of pairs of sessions or session time-series.
- Many potential uses for such similarities, *e.g.*,
  - Partition group into collection of (possibly overlapping) subgroups
  - Classify new individual into appropriate subgroup
- Partitioning may expose previously unrealized substructure in speaker populations; wrt speech therapy, classification may allow diagnosis of individuals as well as prognoses and suggestions for appropriate therapy.

. . . ??? . . .

Conclusions

- There are many possibilities for pattern-based data analyses in Phon, especially with respect to previously-unsupported types of patterns and session time-series – what would you as linguists find useful?
- Your task in this as linguists is to dream – let computer scientists figure out how to make your dreams a reality.