**The PhonBank Project: Data and Software-assisted Methods   
for the Study of Phonology and Phonological Development**

Yvan Rose

Memorial University of Newfoundland

Brian MacWhinney

Carnegie Mellon University

**Abstract**

In this chapter, we provide a description of the tools available through the PhonBank initiative within CHILDES for corpus-based research on phonological development as well as for data sharing. We begin by situating CHILDES and PhonBank within the larger TalkBank database project, and discussing issues affecting both the development of phonological corpora and the central question of data sharing. We then focus more specifically on Phon, the software program developed within PhonBank to support the needs specific to phonological research. This description highlights the main corpus building and data compilation functions supported by Phon. As part of this discussion, we also touch on methodological issues relevant to software-assisted approaches to phonological development, and to phonology, more generally.

**Keywords**

Database; Software; Phonology; Phonetics; Acquisition; Development; IPA; Transcription; Analysis

**Academic biographies**

**Rose:** Dr. Yvan Rose is currently Associate Professor of Linguistics at Memorial University and co-director of the PhonBank Project within CHILDES. He received his Ph.D. in Linguistics from McGill University. His research concentrates on the nature of phonological representations and of their acquisition by young children. He investigates these questions through software-assisted methods, implemented in the Phon program for the analysis of transcript data on phonology and phonological development.

**MacWhinney:** Dr. Brian MacWhinney, Professor of Psychology, Computational Linguistics, and Modern Languages at Carnegie Mellon University, has developed a model of first and second language acquisition and processing called the Competition Model which he has also applied to aphasia and fMRI studies of children with focal lesions. He has also developed the CHILDES Project for the computational study of child language transcript data, the TalkBank system for the study of conversational interactions, AphasiaBank for the study of aphasia, and the E-Prime and PsyScope programs.

**Acknowledgements**

We would like to thank Jacques Durand, Ulrike Gut and Gjert Kristoffersen for inviting us to participate in the Handbook, a much needed resource for the field of corpus phonology. Special thanks to all the scholars, computer scientists and students who have contributed to the PhonBank initiative, in particular Franklin Chen and Leonid Spektor at Carnegie Mellon University as well as Gregory Hedlund at Memorial University of Newfoundland. This project is currently funded by the National Institute of Health. It has also received financial support from the Canada Foundation for Innovation, the Social Sciences and Humanities Research Council of Canada, the Office of the Vice-President, Research and the Faculty of Arts at Memorial University of Newfoundland, as well as through a Petro-Canada Young Innovator Award. Our efforts not would have been as fruitful without the tremendous feedback and support we received from the research community, for which we are grateful.

# Introduction

This chapter reviews recent work on the construction of a computerized database of recordings and transcripts documenting phonological development. This database, called PhonBank, is one of ten subcomponents of a larger database of spoken language corpora called TalkBank.[[1]](#footnote-1) Other interest areas in TalkBank include AphasiaBank, BilingBank, CABank, CHILDES, ClassBank, DementiaBank, GestureBank, Tutoring, and TBIBank. All of the TalkBank corpora have in common the fact that they use the CHAT data transcription format, which enables their smooth and accurate analysis with the CLAN programs (Computerized Language ANalysis). The PhonBank corpus is unique in that it can be analyzed both with the CLAN programs and also with an additional program, called Phon, designed specifically for phonological analysis.

These various databases are central to current research questions in various areas of language and language acquisition. Many of these questions call for multi-dimensional analyses, for example on relationships that exist between ‘smaller’ phonological domains and ‘larger’ units at the levels of the syllable, word, phrase or utterance, to name a few (e.g. Rose, Part II of the Handbook). While phonological aspects of the data are best analyzed within Phon, the researcher must take advantage of the tools built within CLAN, which supports the more general needs of TalkBank. In order to facilitate these multi-dimensional analyses, we are maintaining data compatibility between PhonBank and other TalkBank databases, with a particular focus on the relations between the CHILDES and PhonBank, as they both focus on areas related to language acquisition and related disorders. We describe data conversion in section 5 of this chapter. This is preceded by discussions phonological corpus building and data compilation within Phon, presented in sections 3 and 4, respectively. We begin with a more detailed description of the goals and current state of PhonBank, in the next section.

# PhonBank

PhonBank and Phon are recent outgrowths of earlier work on the Child Language Data Exchange System (CHILDES) Project. Work on the CHILDES Project began in 1984 with support from the MacArthur Foundation for a meeting of 16 child language researchers who agreed on a set of standards for data sharing. The co-organizers of CHILDES were Catherine Snow and Brian MacWhinney. Since 1987, support for the system has come from the National Institutes of Health, with supplemental support from National Science Foundation. In 1984, the primary emphasis was on the shift from handwritten notes to computerized files that could be subjected to automatic searching for developmental patterns. It was not until 1995 that we began to link files to digitized audio. Prior to this, in 1993, we made an initial attempt to extend the CHILDES system to work on phonology. However, the lack of a consistent method for encoding International Phonetic Alphabet (IPA) characters on the computer at that time was a stumbling block. In recent years, the advent of Unicode, XML, improved digital audio and affordable file storage have removed the remaining technical stumbling blocks to the establishment of PhonBank.

The overall goals of PhonBank are fully compatible with those of TalkBank, as well as with new developments in corpus phonology, as represented by other chapters in this book. The basic idea is that, by constructing a large database of accurately transcribed data on phonological development, we can test alternative theories of phonology and phonological development. For example, we can examine the role of universals in phonological markedness versus influences from distributional patterns in the input. We can then use these baseline patterns from normal development to understand the range of variation in learning and ways in which children with language disabilities diverge from developmental norms. The basic tools of PhonBank can also be used to study phonological systems outside of early child language. The same tools can be applied to the acquisition of second languages, learning and mastery of dialects, and phonological effects in aphasia and other communicative disorders.

PhonBank corpora differ from other child language corpora in three regards. First, the children’s productions in PhonBank corpora are all fully transcribed in IPA. Second, most of the PhonBank corpora have been collected from very young children, most of them documented between the ages of 10 and 24 months. Third, few of the PhonBank corpora include transcriptions of the adult input. This is because these corpora focus on an early period of learning when the match of the child’s forms to the adult input is not very close. However, corpora such as the Paris and Lyon French corpora, the Providence English corpus, or the MCF corpus of Portuguese-Swedish bilingualism illustrate how it is possible to collect and transcribe corpora that fully record all adult-child interactions, while still providing accurate records of children’s phonological productions.

## PhonBank: the state of the art

Currently, the PhonBank database includes the corpora of data on child phonological development listed in (1).

PhonBank database: current state

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Language** | **Corpus ID** | **Size (MB)** | **# Participants** | **Audio** | **Adult** |
| Dutch | Zink | 4.4 | 4 | ✓ | - |
| Dutch | CLPF | 3.5 | 12 | ✓[[2]](#footnote-2) | - |
| English | Stanford | 1.3[[3]](#footnote-3) | 5 | - | - |
| English | Smith | 0.5 | 1 | - | - |
| English | Davis | 13.5 | 21 | ✓ | - |
| French | Stanford | 1,3 | 6 | - | - |
| French (Québec) | GoadRose | 2.3 | 2 | ✓ | - |
| French | Paris | 12.4 | 3 | ✓ | ✓ |
| French | Kern | 5.6 | 4 | ✓ | - |
| German | TAKI | 3.9 | 5 | ✓ | - |
| German | Stuttgart | 6.3 | 6 | ✓ | - |
| Japanese | Stanford | 1,3 | 5 | - | - |
| Japanese | Ota | 1.5 | 3 | ✓ | - |
| Portuguese | MCF | 0.9 | 2 | ✓ | ✓ |
| Romanian | Kern | 3.7 | 4 | ✓ | - |
| Swedish | Stanford | 1,3 | 4 | - | - |
| Tunisian Arabic | Kern | 3.0 | 4 | ✓ | - |

This table lists the language of each corpus, the corpus identifiers/contributors, the size of the corpus in megabytes, the number of participants studied, whether or not the corpus is linked to audio data, and whether or not the adult input is transcribed. All of these corpora can be viewed in either CHAT or Phon format, with the majority of them available in both formats. Many more corpora are currently being processed for cross-format compatibility, including the Providence English corpus (55MB, 6 children) and the Lyon French corpus (25.6 MB, 4 children).

Except for the Smith and Stanford archival corpora and a subset of the CLPF corpus, these corpora have the transcribed utterances directly linked to audio. This linkage allows users to replay individual segments and to export the media clips for phonetic analysis through programs such as Praat. It also makes it possible to replay the transcripts online through the transcript browser.[[4]](#footnote-4)

## PhonBank: further goals

Although, as mentioned earlier, the technical stumbling blocks of the 1990s have largely been overcome, the rapid development of PhonBank is still impeded by four barriers: 1) incomplete commitment to data sharing, 2) difficulties with transcription, 3) poor interoperability, and 4) impoverished analytic tools. The PhonBank project is committed to removing each of these four barriers, as discussed in the next sections.

### Data sharing

The biggest barrier to progress in the study of phonological development is the inadequate level of data sharing in comparison with other areas of child language study. In the last three years, this situation has improved markedly with the newly shared corpora listed in (1), all coming online since 2008. In order to share existing data sets, researchers have to make sure that they have secured ethics approval for using their data and they must transfer the relevant files to the PhonBank project, using the guidelines provided on the TalkBank website.[[5]](#footnote-5) Once the data are transferred, members of the PhonBank research team must convert all audio files (digital or analog) to .wav and .mp3 formats, and reformat all transcripts for CHAT compatibility. To facilitate this process, we have developed converters from LIPP, SALT, Praat, WaveSurfer, ELAN, Anvil, and Transcriber formats to CHAT (which can then be further converted into Phon, as described in section 5). However, even with these converters, the output often requires further fine-tuning to serve the particular needs of the researcher.

For the study of second language (L2) phonological development, there are currently no shared data sets. As a result, the study of L2 phonological development currently suffers from many of the symptoms that were affecting the field of L1 development prior to the inception of PhonBank. Indeed, it is at the moment extremely difficult to readily access L2 phonological data from any public source. As Rose (2010) points out, the problem is not that the data do not exist; they do, and are promoted on the websites of various research groups or as part of the scientific literature. The public release of these corpora would open the possibility of new and exciting research, enabling, for example, systematic comparisons between L1 and L2 phonological development, or across various populations of learners, from children being raised with two or more mother tongues, to early or later bilingual development. Researchers in L2 phonology would also receive significant technological support for their corpus-based studies, as well as a ready infrastructure for sharing their data. Bringing these important improvements to the current state of affairs. Whichever the field of empirical or theoretical interest, everyone involved has something to gain through data sharing.

### Transcription

As discussed in Rose (Part II of the Handbook), the representation of speech data through phonetic transcription is a double-edge sword. On the one hand, this means of representing speech data poses a number of methodological issues, many of which may introduce biases in the data. On the other hand, transcription can easily be read and processed by both humans and computer systems. In Phon, we tackle transcription-related issues from two specific angles: reliability and time. To evaluate and minimize problems with reliability, Phon offers support for the multiple-blind approach to phonetic transcription, which is supplemented by an interface for transcription validation. To our knowledge, Phon is the first application to offer such exhaustive methodological support for phonetic data transcription. Beyond reliability, another major barrier to transcription-based research has always been the time needed to transcribe new corpora. To address this problem, we have incorporated various methods that speed up the process of transcription. These include a built-in IPA map as well as dictionaries that provide generic IPA forms representing adult pronunciations for the words orthographically entered by the user in charge of data transcription. We illustrate our implementation of these systems in section 3 below.

### Interoperability

Until recently, one of the major barriers to the construction of a shared corpus of phonological development was the fact that users had prepared their corpora using a variety of programs that could not export their data in any common format. However, as noted in section 2.2.1, we have now developed programs that convert between the CHAT and Phon formats as well as seven other formats. These conversion programs are useful not only for adding corpora to the database, but also for users who wish to avail themselves of analysis schemes only supported in particular programs. For example, in order to properly analyze the development of syllable structure, CHAT data should be exported to Phon where they can be further studied and then, reformatted back to CHAT for archiving. Similarly, multimedia data in the form of individual utterances can be exported from Phon for instrumental analysis within other software packages.

### Data Compilation

The CLAN programs offer extremely good support for investigations of units such as morphemes, words and even syntactic constructions (including conversational and interactional encoding). However, they provide relatively little support for phonological research. To correct this problem, we incorporated a powerful system for phonological compilations into the Phon program. We discuss this system in more detail in the next section, after we provide an overview of Phon’s most central functionality.

# Corpus elaboration within Phon

As mentioned above, the development of Phon aims at the specific needs of acquisitionists working in the realm of phonological productions. The development of this application began in 2003 through a joint honours thesis by Gregory Hedlund and Philip O’Brien, under the supervision of Rod Byrne and Yvan Rose, in close collaboration with Todd Wareham (Hedlund & O’Brien 2004). This work offered the original blueprint for the application. Additional interdisciplinary research work laid the ground for later integration of data annotation algorithms in the program (e.g. Maddocks 2005; Gedge, Hedlund, Rose & Wareham 2007). The development of Phon has also been heavily influenced by central needs of PhonBank, as per the foundational principles listed in (2).

Phon: design and implementation criteria

* 1. Reliability
  2. Theoretical flexibility/neutrality
  3. Simplicity
  4. Compatibility
  5. Extensibility
  6. Availability

The scientific goals behind PhonBank-supported projects make reliability an absolute must. In this regard, we endeavour to produce the most accurate system possible, with the help of the research community, several members of which contribute to software testing and also provide feedback on usability. Analytical flexibility and neutrality are also fundamental from a scientific standpoint, as we want PhonBank data and analysis tools to be useable by a maximal number of researchers, from all theoretical orientations. In this respect, we also welcome suggestions for the incorporation of new functionality. This scientific orientation is also matched by our more general development philosophy, summarized by the last three criteria in (2), according to which Phon is fully compatible with Mac OS X, Windows and Linux operating systems that support Java 1.6, with its functionality designed to be as extensible as possible. Finally, Phon is freely available as open-source software and, as such, naturally lends itself to future development.

Keeping with the view that Phon should maximally accommodate all theoretical approaches to phonology and phonological development, we have developed its features in close collaboration with several scholars, research associates and graduate students engaged in the PhonBank research consortium. In the next section, we highlight the main functions of the application. Many of Phon’s key components have been described in previous publications (e.g. Rose, Hedlund, Byrne, Wareham & MacWhinney 2007, Rose 2008). Whenever appropriate, we supplement these descriptions by updates on refinements introduced in later versions of the application. In order to best illustrate Phon’s functionality, our discussion loosely follows the general workflow it supports. We assume as a starting point that the user has in hand a set of recorded data in digital format. Of course, corpora can be transcribed in Phon without the need for recorded media. In this case, the general workflow remains the same, except for the media-related operations described in section 3.1.2, as they would be irrelevant in this context.

We present the general interface of Phon in (3). It consists of a series of view panels, each dedicated to specific functions within the application. For example, in this screen shot, we can see the Media & Segmentation panel to the left, alongside the Record Data, Syllabification & Alignment and Waveform views. The user can configure the location and proportion of each panel. For example, in (3), the Record Data view is superposed over two additional panels, namely Session Information and Tier Management. Similarly, the Media & Segmentation is laid over the Record List and IPA Lookup panels. Each of these panels can be brought to the front by clicking on their respective tabs within the application.

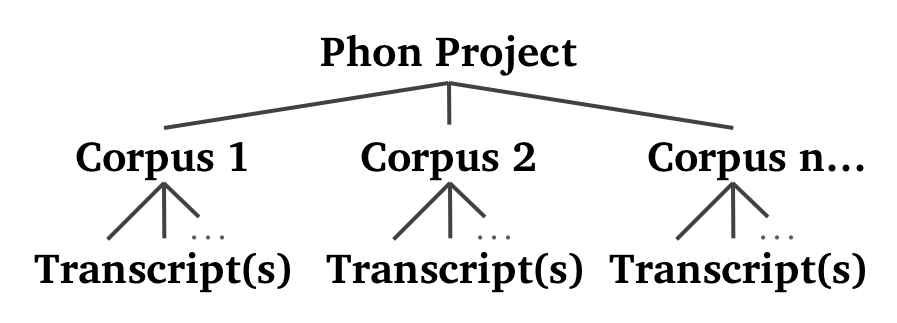
A description...Phon’s general graphical user interface

More functions and related panels are available from the application’s View menu. We discuss a number of these functions in the sections that follow.

## Corpus organization and transcription

### Project management

The elaboration of a corpus of transcribed phonological data often requires the combination of a number of data transcripts, be they from a single speaker over a period of time or from multiple speakers. Irrespective of the data gathering protocol, Phon offers functions to create and manage sets of data transcripts, following the general corpus structure in (4), whereby a project contains one or many corpora, each of which contains a set of data transcripts.



Phon data structure

In longitudinal studies, transcripts typically correspond to data recording sessions. In other types of studies, for example cross-sectional studies, a transcript can be used for each participant, and a corpus can correspond to a particular population or age group.

Each transcript file contains information relative to the context of the study (e.g. participants, ages, languages) as well as a reference to an audio/video file in the case of multimedia corpora. This information, contained in the file header, is followed by a list of records containing transcribed data and related annotations. Each record corresponds to a user-determined speech unit, typically a word, phrase or utterance, depending on research needs.

### Media linkage and segmentation

After the creation of a Phon project, the next step consists of associating a transcript to a media file. Once the file is located on the drive and loaded into PhonMedia for playback, the user is in a position to begin record segmentation. This task consists of the identification of the speech samples that are relevant for research. Similar to a comparable system in CLAN, Phon does not edit media files; it associates user-identified time intervals with data records in the transcript. In typical workflows, media segmentation precedes all transcription-related tasks. In contrast to this, experimental protocols, e.g. based on picture naming tasks, may benefit from a variant of the media segmentation function in Phon, which consists of linking media to an already-annotated session transcript (a similar function is available in CLAN).

The segmented speech segments can be played back directly from the graphical user interface (GUI). Whenever needed, the user can also fine-tune the segments start and/or end time values, a task made easier with the incorporation of waveform visualization, as per the illustration in (3) above.

### Phonetic transcription and validation

As mentioned in section 2.2.2, Phon is, to our knowledge, the first application to offer a fully-integrated system for multiple-blind IPA transcription and validation. Multiple-blind transcription is in essence identical to the double-blind protocol. It consists of the IPA transcription of the relevant speech segments by two or more data transcribers working independently of one another, and with no access to each other’s work. Transcribers can also perform instrumental measurements whenever necessary. In order to assist this task, the segmented speech samples can be exported as individual clips for visualization in speech analysis software programs (e.g. Praat).[[6]](#footnote-6)

The use of blind transcriptions also implies the need for transcription validation, in order to determine which of the blind transcriptions will ultimately be used in data compilations. One common practice under this protocol consists of performing an inter-transcriber reliability test. It is not always clear, however, how the discrepancies detected between the transcribers may affect data analysis. This is especially problematic since there exists no standard practice in the field. One solution to this problem consists of improving transcriptions methods, for example through a consensus-based approach to phonetic transcript validation. We developed an interface within Phon which facilitates record-by-record comparisons of the blind transcriptions by a team of two (or more) data ‘validators’. In most cases, the validators simultaneously listen to the record’s media segment and, together, select which of the blind transcriptions best represents the speech sample. Whenever necessary, the selected transcription can be further adjusted according to the details noticed by the transcript validators. While this method is relatively onerous both in time and human resources, its combined steps help maximizing transcription reliability for research purposes. Of course, regardless of the amount of care put into it, and in spite of its crucial role in creating readable transcripts of spoken forms, the symbolic representation of speech sounds remains a methodological compromise.

The use of blind transcription and associated validation systems is optional. Depending on research needs, the user can decide whether to use these functions. If the user decides not to use blind transcriptions, the transcriptions are entered directly into the transcript. Similarly, the decision to use password protection for blind transcripts, which may be unnecessary in many situations, is left to the user. Note as well that only validated blind or directly-entered transcriptions can be used for further annotation and data compilation; non-validated blind transcriptions are saved within the transcript but cannot be accessed by other Phon functions.[[7]](#footnote-7)

Aside from the mode of data entry used, Phon provides additional functionality to streamline the inherently time-consuming process of transcription. For example, a built-in IPA map is provided. Also included is an IPA Lookup function, which provides model IPA transcriptions based on the words identified in the Orthography tier. The screen shot in (5) offers a glimpse at the overall transcription facility in Phon, with a partially transcribed record, the built-in dictionary function underneath it, and a section of the IPA map overlaying in the lower right corner.

A description...

Transcription within Phon: a snapshot

While these facilities offer several advantages for phonetic transcription, the system does not solve all issues involved in phonetic transcriptions. For example, the IPA Lookup function offers IPA forms that are inherently limited in their detail. As such, they must be carefully double-checked to avoid data misrepresentations. Such pitfalls include dialect-particular details about vowel or consonant realization and phonological patterns such as vowel reduction, segmental contraction or sandhi effects, all of which may ultimately affect the representation of a word or phrase (e.g. “I am not” versus “I ain’t”). If used with the required care, dictionaries of IPA forms can markedly reduce the amount of time needed to attain new phonological transcriptions, without affecting the level of detail required in data analyses.

### Word grouping

It is common knowledge (at least among phonologists) that phonological patterning tends to be confined within syntactic or morphological domains. Given this, systematic research in phonology often requires the subdivision of transcribed utterances into smaller domains such as phrases, clitic groups or individual words. Using the word grouping function, the user can break the transcribed utterance into sub-domains, represented in the GUI through unobtrusive grey bracketing, as illustrated in (6).

Example word groups[[8]](#footnote-8)

A description...

The alignment of word group across all relevant also effectively provides a system for ‘vertical’ utterance splicing across target and produced forms within each utterance.[[9]](#footnote-9)

## Automatized data annotation systems

As discussed in Rose (Part II of the Handbook), given combinatorial properties of phonological subsystems (e.g. phones can be described as feature sets; strings of phones are organized in predictable syllable groupings), several aspects of data annotations can be obtained through automatized means of data coding. In the subsections that follow, we describe the main systems implemented in Phon, all of which aim at optimizing the workflow, without neglecting central requirements such as reliability and theoretical neutrality.

### Descriptive phonological features

Each IPA symbol or diacritic entered in IPA Target or Actual transcriptions is automatically associated with a set of phonological features. Symbol-to-feature associations are performed internally, through a set of descriptive features built into Phon. This feature set is as descriptively neutral as possible, in order not to impose any bias on the data compilations.

The user can also specify a custom feature set for the extraction of tightly defined data compilations. In addition, a full feature set editor is planned in future versions, in order for the user be able to add or remove features for a given symbol, or add symbols to the set. The latter function will be useful for the transcription of sounds that have no clear correspondence with the IPA, something particularly useful in the context of acquisition studies. Researchers using this system will be able to add a cover symbol such as “F” in the transcript, and associate it to a feature set such as {Labial, Continuant},[[10]](#footnote-10) irrespective of other potential phone characteristics such as its voicing or exact labial articulation.

### Syllabification

Phon also automatically labels each word transcribed in the IPA Target and Actual tiers for syllable-level information (e.g. onset, nucleus, coda). These annotations are obtained through deterministic algorithms, in a way similar to traditional theories of syllabification (e.g. Selkirk 1982). Support is provided for a number of languages (e.g. Catalan, Dutch, English, French, Portuguese, Spanish, …); additional syllabification algorithms can be added upon request. Each time the algorithm produces an unwanted outcome, for example in the case of distorted learners’ productions which do not match the types of strings expected by the algorithm, the user can modify the annotation directly through the GUI. Syllable-level annotations are useful in data compilations which make reference to phone location within the string, as well as more general data characterizations, for example about syllable types (e.g. CV versus CVC), as we illustrate in section 4.1.2.

### Phone alignment between target and actual forms

The two automatized systems of data annotation described above can be used in virtually all types of corpus-based research in phonology. More particular to acquisition studies is the frequent need to compile information about phonological accuracy, which requires systematic comparisons between the learner’s produced forms and corresponding target (expected) forms. For example, the investigation of segmental substitutions (e.g. the production of [w] for target [r]) requires phone-by-phone comparisons between target and actual forms. In order to support this need, Phon performs automatic alignment of IPA Target and Actual phones. This function, illustrated in (7), relies on a ‘best-guess’ dynamic programming algorithm (Maddocks 2005).

Phone alignment

* 1. Transcriptions b. Aligned phones

A description...A description...

The alignment is confined within word groups, which are, as mentioned above, considered to be independent domains of analysis. Phone alignment transcends syllable boundaries within words and word groups. As we can see in (7b), the actual syllable [bæ] straddles the first two syllables in the target form. However, phone alignments cannot straddle word group boundaries. Similar to all automatized functions built into Phon, the alignments produced by the algorithm can be modified at will by the researcher. For example, the [b] in the actual form can easily be realigned with target [n] through Phon’s GUI.

As we can see from the brief descriptions of Phon’s automatized data annotation systems, our design centres around the need to optimize as many aspects of data preparation as possible, while keeping with the central requirement that these systems do not introduce unwelcome biases in the corpus. Systematic data verifications by the researcher thus remain a must. After completion of all transcription, annotation and related verification tasks, the data are ready for compilation. Phon supports powerful yet flexible methods for data compilation, which we introduce next.

## Data compilation

The user can perform compilations on virtually all aspects of the transcripts, including orthographic or phonetic transcriptions as well as more specific information relative to phonological features, syllabification or target-actual phone alignment. Phon also provides support for the detection of patterns that span across strings of consonants and vowels, which can be used in compilations of data on consonant and vowel harmony as well as consonant metathesis. Finally, all data compilations can incorporate session-level information such as the participants’ names (identifiers), ages or age ranges.

The Phon data compilation system relies on a plug-in architecture: Each search interface is an individual scripted program that combines the components required for the query itself with a set of instructions specifying the relationships between these components. In this respect, scripts can be used to adorn compilations with metadata further specifying the data sets returned by the queries. For example, in searches for the actual realization of complex onset (e.g. the [bɹ] cluster in ‘bran’ [bɹæn]), the application can return metadata such as “-P2” to indicate that the second position of the cluster is deleted in a production such as “ban” [bæn]. Such meta-annotations are useful for post-hoc data treatment in spreadsheet or statistical software packages.

From a more technical perspective, the scripts are based on the JavaScript programming language. Built-in or user-defined javascripts govern the operations performed by Phon’s search engine. This system supports searches based on text strings (e.g. orthographic or IPA transcriptions) and regular expressions. In order to support needs specific to phonological research, the system also incorporates support for phonological expressions, through Phonex, a tailor-made pattern matching system. Phonex allows the searching of phones or metaphones (e.g. stress markers, syllable boundary markers) at all levels of data annotation described in section 3. A Phonex matcher is, in a nutshell, a statement written in a language which Phon can understand. For instance, using a matcher such as “b: Onset” the researcher can search for all instances of [b] in onset positions. To run a similar query on all labial consonants (e.g. [b, p, f, v, …]), no exhaustive listing of all the potential consonants is needed; the desired class of phones can be specified through the Phonex matcher “{Consonant, Labial}: Onset”. The researcher can thus generate various types of data compilations, each of them offering its specific angle on data compilation. A number of these methods are described in the next section.

# Data query and reporting: a few illustrations

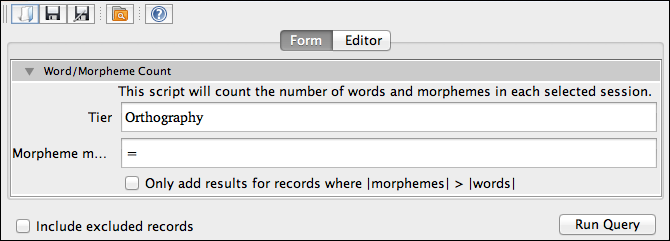
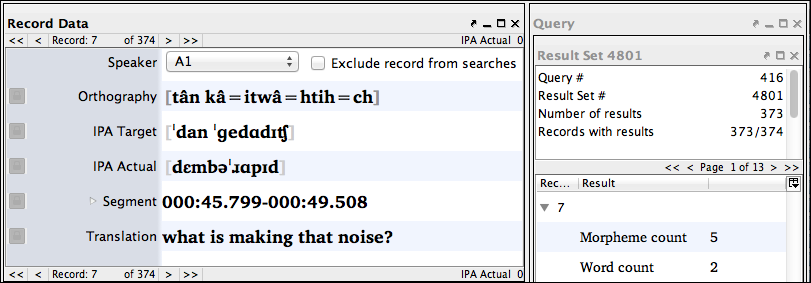
We begin our description of the Phon data compilation system with a number of query types, each of which focuses on specific aspects of phonological corpora. We then provide a brief overview of how the data returned by these queries can be exported for reporting or further processing outside of Phon.

## Data query

### Productivity measures

General productivity measures such as the Mean Length of Utterance (MLU; e.g. Brown 1973) often provide a basis for the interpretation of phonological patterns. Using the Phon scripting interface, we recently created a basic MLU calculation system.[[11]](#footnote-11) This system provides assessments on two different measures: the word and the morpheme. The latter is particularly useful to calculate meaningful MLUs in polysynthetic languages such as Turkish or Cree, in which words often correspond to phrases or even full sentences in ‘Western’ languages. In the examples in (8), morpheme counts require explicit data coding, to identify morpheme boundaries. Morphemes are identified with the symbol “=” in the Orthography tier, as in (8a); word boundaries are calculated based on the spaces that occur between the orthographically transcribed words. The screen shot in (8b) illustrates the visualization of one record alongside related search results, which contains 2 words and a total of 5 morphemes. Finally, the Word/Morpheme count function also generates mean values for each session selected for query, in (8c). These latter results can easily be formatted in spreadsheets or other software for statistical analysis.

Example MLU search[[12]](#footnote-12)

* 1. Search parameters
  2. Sample of per-record results visualization
  3. Session summary results

**Session: A1.A1-2006-10-18-v27**

**-------------------------------------**

**Word total: 735**

**Word avg: 1.96524064171123**

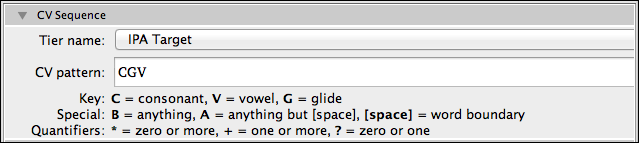
**Morpheme total: 1032**

**Morpheme avg: 2.7593582887700534**

As mentioned above, this function is, in its current shape, relatively simplistic, as it is does not perform MLU calculations based on more refined sets of criteria. Regarding phonological productivity, for example, it does not support measures such as the phonological MLU (pMLU) proposed by Ingram (2002) (see Taelman, Durieux & Gillis 2005 for further discussion of this measure).[[13]](#footnote-13)

### Queries on words and syllable types

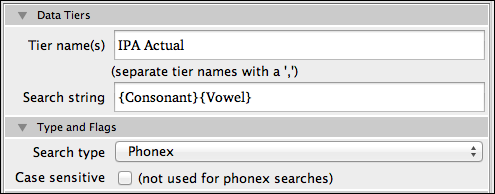
As already mentioned in section 3.2.2, Phon can extract information about syllable types. Such computations are useful to assess syllable complexity using descriptive sequences of consonants, glides and vowels (e.g. Kern & Davis 2009 on babbling development; Levelt, Schiller & Levelt 1999/2000 on the acquisition of syllable types). The interface for such searches is illustrated in (9).

General syllable shapes

As we can see from this screenshot, syllable types can be specified by the researcher through sequences of descriptive capital letters, in line with common practice in the scientific literature. Phon also provides functionality to establish inventories of word types, for example words with particular stress patterns (e.g. iambic, trochaic or amphibrach forms; e.g. Prieto 2006).

### Phone-based queries

Typically one of the most frequently used methods to assess phonological development consists of establishing inventories of phones and phone combinations attested in the child’s productions. Such inventories are relevant to a series of research questions, ranging from relatively concrete considerations about the child’s articulatory abilities (e.g. Kern & Davis 2009) to more abstract accounts of these abilities in terms of feature co-occurrence constraints (e.g. Levelt & van Oostendorp 2007; van ‘t Veer 2010). Phon supports the extraction of such inventories, which can also be compiled using criteria for syllable- and stress-related information. This enables queries on specific segmental contexts, for example between vowels and consonants in CV and VC sequences (e.g. Kern & Davis 2009; Fikkert & Levelt 2008). As the screenshot in (8) illustrates, the simple Phonex matcher “{Consonant}{Vowel}” entered in a query on the IPA Actual tier will return all CV sequences produced by the child.

Query returning CV sequences found in IPA Actual forms

While the search illustrated by this screenshot relies on the general features {Consonant} and {Vowel}, virtually all descriptive features found in the phonological literature can be used to build families of phones, as we discuss next.

### Feature-based queries

Similar to adult phonological systems, child phonologies often exhibit congruent patterns of phone distributions. For example, the phonological literature is filled with examples whereby certain classes of phones are restricted to certain positions within the word or syllable. Empirical generalizations about positional asymmetries are most easily described in terms of features. While the theoretical status of features is subject to some controversy,[[14]](#footnote-14) their usefulness in phonological data descriptions is undeniable.

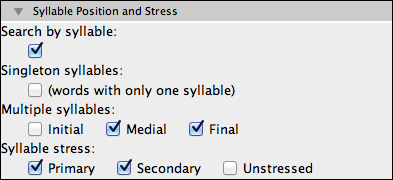
As mentioned in section 3.2.1, each IPA symbol supported in Phon is associated with a set of descriptive features. Diacritics added to IPA consonants or vowels also supplement their featural descriptions. For example, adding a tilde ( ̃ ) to the vowel [o] adds {Nasal} to this vowel’s feature set: [õ] = {Vowel, Mid, Back, …, Nasal}. As alluded to in the previous section, feature sets can be used to compile information about categories of phones and phone sequences. For example, a Phonex matcher such as “{Consonant, Labial}{Vowel, Round}” will return all sequences of labial consonants followed by round vowels, irrespective of any other factor such as the consonant manner of articulation or the vowel lingual place of articulation (e.g. [y] versus [u]).

Finally, similar to the other examples discussed above, feature-based searches can be specified for positions within the utterance, word group, word, and syllable, including syllable stress and position within the syllable. This system can also be combined with information about target-actual phone alignment, as we discuss next.

### Comparisons between IPA target and actual forms

Except in the case of babbling, for which no target forms can be identified by the researcher, phonological patterns discussed in the acquisition literature (e.g. rhotic gliding: [wɛd] for ‘red’; consonant harmony: [ɡʌk] for ‘duck’) cannot be studied without explicit reference to target forms. Also, since phonological patterns are often positionally determined, the incorporation of criteria defining particular positions within the word or syllable is required.

For example, in studies of positional velar fronting (e.g. Chiat 1983, Stoel-Gammon 1996, Bills & Golston 2002, Inkelas & Rose 2008, MacAllister 2009 and references therein) syllable-based labelling can be used to observe renditions of velar consonants across various positions. As Inkelas & Rose (2008: 707) observe, child E, a learner of English, produces coronals for target velars located in word-initial and stressed onsets (e.g. cake [ˈkeɪk] → [ˈteɪk]; again [əˈɡɛn] → [əˈdɛn]). However, velars in word-medial onsets of unstressed syllables and in codas are realized in target-like fashions (e.g. bagel [ˈbejɡu]; back [ˈbæk] → [ˈbæk]). The investigation of such positional conditioning requires comparisons of the behaviour of velars across syllable and word positions. Support for such queries is illustrated in (11) where we can see the criteria set for non-initial onsets of (primary or secondary) stressed syllables, the context that would return a value such as [ˈhɛksəˌdɔn] for ‘hexagon’ and [əˈdɛn] for ‘again’.

Context-sensitive detection of phonological patterns  


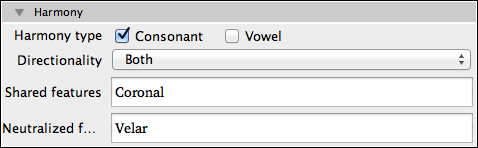
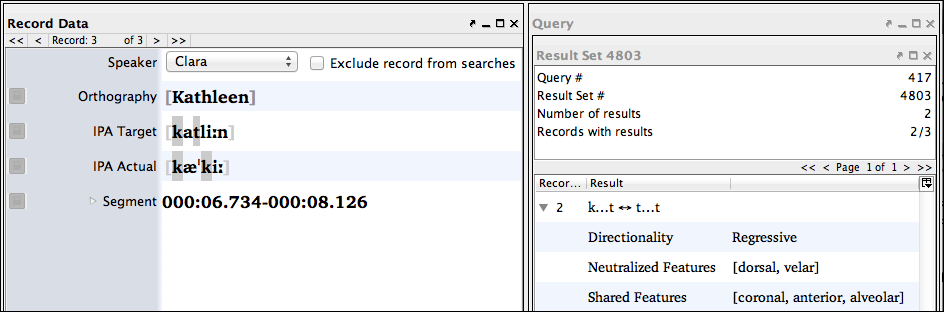
Context-sensitive searches can thus be achieved easily, through the setting of criteria within the search interface.

### Harmony and metathesis pattern detection

The use of specific search criteria such as those illustrated above is, however, not always useful, especially when we look for patterns whose definition may vary based on too many parameters. This is the case of both consonant harmony and consonant metathesis. Consonant harmony consists of the sharing of features between consonants across intervening vowels. For example, in the word ‘Kathleen’ produced as [tæti], in which we observe a neutralization of the first consonant’s place of articulation, which is substituted by the coronal articulation that independently appears on the second consonant in the target form. Consonant metathesis consists, instead, of the swapping of consonants (or consonantal features) across intervening vowels (e.g. ‘cat’ produced as [tæk]).

The detection of these long-distance patterns requires simultaneous comparisons between aligned target and actual consonants across two or more positions, a challenging task, even in the eyes of a trained phonologist. As illustrated in (12), a specialized algorithm designed by Gedge, Hedlund, Rose & Wareham (2007) is implemented in Phon to detect such patterns. This example of consonant harmony detection comes from the Goad-Rose corpus of Québec French acquisition available through PhonBank (see van ’t Veer 2010 for compilations of this corpus using Phon). As we can see in (12a), the algorithm’s interface offers a series of query options, all of which correspond to descriptive aspects of harmony patterns. In this illustration, the system is configured to seek coronal harmony patterns affecting velars in progressive and regressive directions.

Consonant harmony

* 1. Search interface  
     
  2. Sample result  
     

Records returned by these search parameters can be visualized alongside metadata describing both the features involved in the pattern and its directionality, as in (12b)[[15]](#footnote-15) Using this interface, the researcher can quickly browse through the results, visualize the relevant records, playback the speech segments they are associated with and, whenever needed, further annotate these records. Search results can also be exported through the Phon reporting system, to which we turn next.

## Data reporting

As we saw in the last example, the results returned by Phon queries can be visualized directly within the application. Another property of these results is that they are persistent; they are saved in a separate database, from which they can be accessed for further processing. In order for data compilations to remain fully interpretable, the scripts, Phonex matchers and search parameters used to specify them are saved alongside the results themselves. These results can also be exported for further analysis or presentation purposes. The reports are also fully configurable. The user can add/remove various section listings, for example, the search parameters, the number of results returned for each session included in the query as well as inventories of patterns and related record data. Examples of the latter two are provided in (13).

Report sample

* 1. Inventory of patterns b. Record data

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| ʁ ↔ j | 8 |  | Result | ʁ ↔ ∅ |
| ʁ ↔ w | 1 |  | Orthography | Gaspard |
| ʁ ↔ ʁ | 8 |  | IPA Target | [ɡaspɑʁ] |
| ʁ ↔ ∅ | 7 |  | IPA Actual | [ɡapaː] |

As we can see in (13a), the inventory section of the report provides a count of each of the patterns compiled by the script. Such data can be used to quickly assess the relative prominence of each of the patterns. (13b) exemplifies related corpus data. As the astute reader might see, this example shows a long vowel concomitant with [ʁ] deletion. In line with the maximally neutral approach we strive to maintain at all levels of data processing in Phon, the interpretation of this potential case of vowel lengthening is entirely left to the user. In case this observation could have an impact on the analysis proposed, the user could capitalize on the flexible query system built into Phon to design additional data queries addressing the systematicity of this observation across the dataset or, else, identify the relevant set of records for further processing, whose corresponding media clips could also be exported for acoustic analysis.

In order to address questions that pertain to development, for example through the tracking of patterns across a number of recording sessions, Phon supports the generation of aggregated reports. These reports organize results per patterns on a horizontal axis, where each column corresponds to a given recording session, as illustrated in (14) with a representative example from child Catootje documented in the CLPF corpus.

Aggregated inventories (behaviour of target [bʀ]; representative example from CLPF corpus)

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | 1990-11-28 | 1991-01-09 | 1991-01-23 | 1991-02-20 |
| bʀ ↔ b | 2 | 1 | 4 | 0 |
| bʀ ↔ bj | 2 | 0 | 0 | 0 |
| bʀ ↔ bḻ | 0 | 1 | 0 | 0 |
| bʀ ↔ bʀ | 0 | 0 | 1 | 4 |

As we can see in this example, the behaviour of Catootje’s target [bʀ] clusters evolves across recording sessions. Building on this report format, the user can track this type of developmental pattern. The user can also use these aggregated reports as a basis for other data processing needs, for example to perform statistics or generate graphs for data presentation. The data reports generated by Phon are formatted as CSV files, a format which easily lends itself to these tasks, as it can be imported by all major third-party software applications such as spreadsheets or statistical packages.

Finally, as we can see in this and other examples above, the Phon data compilation system supports query-generated metadata annotations. This meta-annotation capability generates useful labels for systematic data mining, independent of any purpose-specific annotation. Of course, as research develops on the corpus, needs for new user-entered annotations in data records may emerge. This is to be expected in all pragmatic approaches to corpus-based research (e.g. Voorman & Gut 2008 for further discussion).

# Phon-CLAN compatibility

Going back to the larger context of PhonBank, as we mentioned in the introduction, one of the goals of this project consists of supporting integrated research on various components of the grammar, for example by making simultaneous observations at the phonological, morphological and syntactic levels. While Phon and CLAN utilize different file formats, the Phon and CHAT formats, respectively, we are supporting a high degree of connectivity between these two formats, via the more general TalkBank XML format. A corpus created within CLAN must first be converted into TalkBank XML and from there into the Phon format. Following the same logic, a Phon corpus will first be converted into TalkBank XML before it is converted into CLAN. Conversions between CHAT and TalkBank XML are performed by the Chatter utility; conversions between Phon and TalkBank XML are performed with the PhonTalk utility.[[16]](#footnote-16)

Conversions from the CLAN to the Phon format allow the fast incorporation of already-existing CHILDES corpora into PhonBank: old data can now be used in new ways, often with only minimal coding efforts needed. More generally, conversions between these two applications also optimize the synergy of analytic functions supported within CHILDES/PhonBank: the overlap in functionality between CLAN and Phon is minimal, especially outside of media linking and transcription functions.

# Conclusion

Although still in its infancy, the PhonBank initiative already provides, through Phon and related software packages such as CLAN, a sizeable number of tools, the combination of which offers systematic support for corpus-based research in phonology, phonological development and related disciplines. Corpora of phonological data can now be transcribed, annotated and compiled in ways that are both efficient and systematic. Often implicit to the descriptions of the various functions described above is the fact that several aspects of phonological theory, especially in the area of phonological representation, have provided a sound foundation for the development of these tools. The labels used for data transcription and other annotations contribute to systematic data representations required for fine-grained characterizations of the phonological patterns attested in the corpus. The interpretation of these labels, however, is left entirely to the user.

Phonology, as a research field, has been evolving at a rapid pace over the past few decades, especially with the still recent yet conspicuous trend towards the integration of phonetic evidence into phonological analyses. It is our intention to keep pace with this evolution in future revisions of the Phon functionality, for example through support for acoustic data compilations. More generally, Phon as well as all other tools designed within the larger CHILDES project, which are freely available as open-source software, contribute to significant savings in research costs. In order to further optimize the resources available and, at the same time, facilitate large-scale scientific investigations of child language data, the researchers who benefit from these tools should also actively engage in data sharing, for which PhonBank and its siblings within the larger TalkBank project provide a ready and compelling infrastructure.

**References**

Bills, Shannon & Chris Golston. 2002. Prosodic and Linear Licensing in English Acquisition. In L. Carmichael, C.-H. Huang & V. Samiian (eds.), *Proceedings of the Western Conference on Linguistics (2001)*, 13-26. Fresno: California State University.

Brown, Roger. 1973. *A First Language: The Early Stages*. Cambridge, MA: Harvard University Press.

Chiat, Shulamuth. 1983. Why Mikey’s Right and My Key’s Wrong: The Significance of Stress and Word Boundaries in a Child’s Output System. *Cognition* 14, 275-300.

Fikkert, Paula & Clara Levelt. 2008. How does Place Fall into Place? The Lexicon and Emergent Constraints in Children’s Developing Grammars. In Peter Avery, B. Elan Dresher & Keren Rice (eds.), C*ontrast in Phonology: Theory, Perception, Acquisition*, 231-268. Berlin: Mouton de Gruyter.

Gedge, Jason, Gregory J. Hedlund, Yvan Rose & Todd Wareham. 2007. Natural Language Process Detection: From Conception to Implementation. *17th Annual Newfoundland Electrical and Computer Engineering Conference* (NECEC).  
<http://necec.engr.mun.ca/ocs2007/papers.php?first_letter=G>

Hedlund, Gregory & Philip O’Brien. 2004. A Software System for Linguistic Data Capture and Analysis. B.Sc. Honour's Thesis, Memorial University of Newfoundland.

Ingram, David. 2002. The Measurement of Whole-word Productions. *Journal of Child Language* 29, 713-733.

Inkelas, Sharon & Yvan Rose. 2008. Positional Neutralization: A Case Study from Child Language. *Language* 83, 707-736.

Kern, Sophie & Barbara Davis. 2009. Emergent Complexity in Early Vocal Acquisition: Cross-linguistic Comparisons of Canonical Babbling. In Ioana Chitoran, François Pellegrino & Egidio Marsico (eds.), *Approaches to Phonological Complexity*, 353-375. Berlin: Mouton de Gruyter.

Levelt, Clara & Marc van Oostendorp. 2007. Feature Co-occurrence Constraints in L1 Acquisition. In Bettelou Los & Marjo van Koppen (eds.), *Linguistics in the Netherlands 2007*, 162-172. Amsterdam: John Benjamins.

Levelt, Clara C., Niels O. Schiller & Willem J. Levelt. 1999/2000. The Acquisition of Syllable Types. *Language Acquisition* 8, 237-264.

Maddocks, Keith. 2005. An Effective Algorithm for the Alignment of Target and Actual Syllables for the Study of Language Acquisition. B.Sc. Honour's Thesis, Memorial University of Newfoundland.

McAllister, Tara. 2009. The Articulatory Basis of Positional Asymmetries in Phonological Acquisition. Ph.D. Dissertation, Massachusetts Institute of Technology.

Prieto, Pilar. 2006. The Relevance of Metrical Information in Early Prosodic Word Acquisition: A Comparison of Catalan and Spanish. *Language and Speech* 49, 231-259.

Rose, Yvan. 2008. Phon 1.3: Current Features and Short Term Outlook. *Child Language Bulletin* 28.  
<http://iascl.talkbank.org/bulletins/bulletinV28N1.html>

Rose, Yvan. 2010. The PhonBank Initiative and Second Language Phonological Development: Innovative Tools for Research and Data Sharing. In Alice Henderson (ed.), *Proceedings of the International Conference on English Pronunciation: Issues & Practices*, 223-241. Chambéry: Université de Savoie.

Rose, Yvan, Gregory J. Hedlund, Rod Byrne, Todd Wareham & Brian MacWhinney. 2007. Phon 1.2: A Computational Basis for Phonological Database Elaboration and Model Testing. In Paula Buttery, Aline Villavicencio & Anna Korhonen (eds.), *Proceedings of the Workshop on Cognitive Aspects of Computational Language Acquisition, 45th Annual Meeting of the Association for Computational Linguistics*, 17-24. Stroudsburg, PA: ACL.

Rose, Yvan & Sharon Inkelas. 2011. The Interpretation of Phonological Patterns in First Language Acquisition. In Colin J. Ewen, Elizabeth Hume, Marc van Oostendorp & Keren Rice (eds.), *The Blackwell Companion to Phonology*, 2414-2438. Malden, MA: Wiley-Blackwell.

Selkirk, Elizabeth O. 1982. The Syllable. In Harry van der Hulst & Norval Smith (eds.), *The Structure of Phonological Representation*, vol. 2, 337-385. Dordrecht: Foris.

Stoel-Gammon, Carol. 1996. On the Acquisition of Velars in English. In Barbara H. Bernhardt, John Gilbert & David Ingram (eds.), *Proceedings of the UBC International Conference on Phonological Acquisition*, 201-214. Somerville: Cascadilla Press.

Taelman, Helena, Gert Durieux & Steven Gillis. 2005. Notes on Ingram's Whole-word Measures for Phonological Development. *Journal of Child Language* 32, 391-405.

Veer, Marijn van 't. 2010. Offset syllabification in Québec French and Dutch: Evidence from Acquisition. Paper presented at *7th Old World Conference in Phonology*, Nice, France.

Voormann, Holger & Ulrike Gut. 2008. Agile Corpus Creation. *Corpus Linguistics and Linguistic Theory* 4, 235-251.

1. For more information about TalkBank, see: <http://talkbank.org>. [↑](#footnote-ref-1)
2. Audio data are currently available for 5 of the 12 documented children. [↑](#footnote-ref-2)
3. The 1.3 MB figure is for the entire Stanford corpus. [↑](#footnote-ref-3)
4. <http://childes.psy.cmu.edu/browser>. [↑](#footnote-ref-4)
5. <http://talkbank.org/share/irb/>. [↑](#footnote-ref-5)
6. For more information about Praat, see <http://www.fon.hum.uva.nl/praat/>. [↑](#footnote-ref-6)
7. We are currently investigating the possibility of adding support for an inter-transcriber reliability assessment system. Through this system, we would be in a position to provide quantitative estimates of inter-transcriber reliability independent of consensus-based verifications. [↑](#footnote-ref-7)
8. Note also the absence of bracketing in the Notes tier, which is not group-aligned. [↑](#footnote-ref-8)
9. Because of non-trivial constraints pertaining to our data format, nested or overlapping word groups are currently not supported. We currently have no plans to tackle this issue. [↑](#footnote-ref-9)
10. Feature sets in Phon are listed between braces; capitalization of the feature labels is optional. [↑](#footnote-ref-10)
11. This system is fairly basic in that it does not incorporate all of the intricacies related to MLU calculation. CLAN offers a much more refined system. [↑](#footnote-ref-11)
12. This custom script can be downloaded from our public repository of scripts, available at: <http://phon.ling.mun.ca/phontrac/wiki/search/scriptlibrary>. [↑](#footnote-ref-12)
13. We are currently investigating ways to best implement this algorithm into Phon. [↑](#footnote-ref-13)
14. See Rose & Inkelas (2011) for recent discussion. [↑](#footnote-ref-14)
15. As we can see in (12b) as well, the list of descriptive features provided by Phon is redundant, with values such as {Coronal} and {Alveolar} both being returned. This redundancy, which may be perceived as a hindrance at first sight, is required to maximize the theoretical neutrality of the data compilations. [↑](#footnote-ref-15)
16. These utilities can be downloaded from <http://www.talkbank.org/software/>. [↑](#footnote-ref-16)