CHAPTER 19

THE PHONBANK PROJECT

Data and Software-Assisted Methods for the Study of Phonology and Phonological Development

YVAN ROSE AND BRIAN MACWHINNEY

19.1 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] 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 analysed 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 (see Rose, this volume). While phonological aspects of the data are best analysed within Phon, the researcher may also 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

[1] For more information about TalkBank, see: <http://talkbank.org>.

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. This is preceded by discussions of 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.

19.2 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 sixteen 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 the 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 contributions to this publication. 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, the majority of the PhonBank corpora have been collected from very young children, most of them documented between the ages of 10 and 36 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.

19.2.1 PhonBank: The State of the Art

Currently, the PhonBank database includes the corpora of data on child phonological development listed in Table 19.1.

Table 19.1 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 (6 children) and the Lyon French corpus (4 children).

The majority of these corpora also 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.[2]

19.2.2 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.

19.2.2.1 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 Table 19.1, 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.[3] Once the data are transferred, members of

[2] <http://talkbank.org/browser>.

[3] <http://talkbank.org/share/irb/>.

Table 19.1 PhonBank database: current state

Language	Corpus ID	Size (MB)	# Participants	Audio	Adult
Cree	CCLAS	17	1	✓	–
Dutch	Zink	19.4	4	✓	–
Dutch	CLPF	53.7	12	✓[i]	–
English	Chiat	2.9	3	✓	–
English	Compton-Pater	66	3	–	–
English	Davis	13.5	21	✓	–
English	Goad	27.9	2	✓	–
English	Inkelas	5	1	–	–
English	Smith	8.5	1	–	–
English	Stanford	3.6	5	–	–
French (Québec)	Goad-Rose	18.8	2	✓	–
French	Kern	17.6	4	✓	–
French	Stanford	3.7	6	–	–
French	Paris	12.4	3	✓	✓
German	Stuttgart	6.3	6	✓	–
German	TAKI	3.9	5	✓	–
Japanese	Ota	1.5	3	✓	–
Japanese	Stanford	1.3	5	–	–
Portuguese	CCF	73.6	5	✓	–
Portuguese	Freitas	33.5	7	–	–
Portuguese	MCF	0.9	2	✓	✓
Romanian	Kern	11.4	4	✓	–
Swedish	Lacerda	3.1	3	✓	–
Swedish	Stanford	1.5	4	–	–
Tunisian Arabic	Kern	24.6	4	✓	–

[i]Audio data are currently available for 5 of the 12 documented children.

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. In short, data sharing is one of the most compelling ways to move the entire field of research forward.

19.2.2.2 Transcription

As discussed in Rose, this volume, the representation of speech data through phonetic transcription is a double-edged sword. On the one hand, the use of phonetic transcription to represent speech may introduce certain biases in the data. On the other hand, phonetic 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 consensus-based transcript 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.

19.2.2.3 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 analyse 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.

19.2.2.4 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.

19.3 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 and O'Brien 2004). This work offered the original blueprint for Phon. Additional interdisciplinary research laid the ground work for later integration of data annotation algorithms in the program (e.g. Maddocks 2005; Gedge et al. 2007). The development of Phon has also been heavily influenced by the central needs of PhonBank, as per the foundational principles listed in Table 19.2.

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 Table 19.2, according to which Phon is fully compatible with Mac OS X, Windows, and Linux operating systems that support the Java Runtime Environment, 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.[4]

In 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

[4] The source code is available from <https://www.phon.ca/>.

Table 19.2 Phon: design and implementation criteria

a. Reliability
b. Theoretical flexibility/neutrality
c. Simplicity
d. Compatibility
e. Extensibility
f. Availability

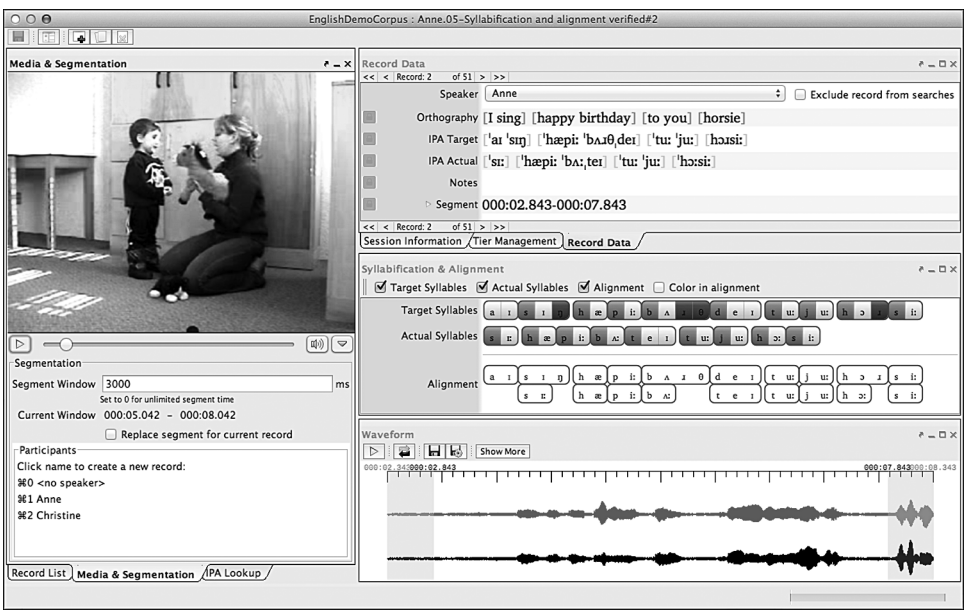

FIGURE 19.1 Phon's general graphical user interface.

described in previous publications (e.g. Rose et al. 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 Figure 19.1. The figure displays 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 Figure 19.1, the Record Data view is superposed over two additional panels, namely Session Information and Tier Management. Similarly, 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.

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.

19.3.1 Corpus Organization and Transcription

19.3.1.1 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 Figure 19.2, whereby a project contains one or many corpora, each of which contains a set of data transcripts.

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.

19.3.1.2 Media Linkage and Segmentation

After the creation of a Phon project, the next step involves associating a transcript to a media file. Once the file is located on the drive and loaded into Phon 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

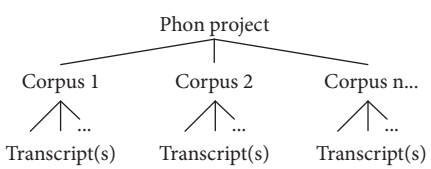

FIGURE 19.2 Phon data structure.

with data records in the transcript. In typical workflows, media segmentation precedes all transcription-related tasks. In contrast to this, experimental protocols, for example 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 Figure 19.1.

19.3.1.3 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).[5]

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 transcription 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 to maximize 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

[5] For more information about Praat, see <http://www.fon.hum.uva.nl/praat/>.

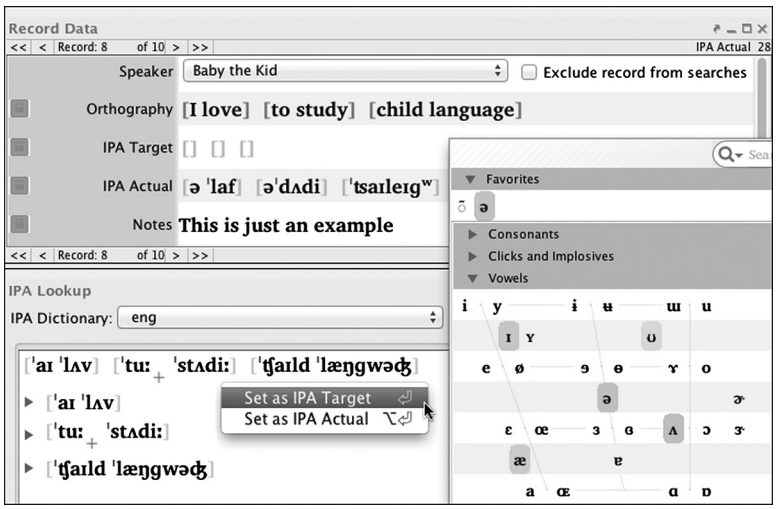

FIGURE 19.3 Transcription within Phon: a snapshot.

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.[6]

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 screenshot in Figure 19.3 offers a glimpse of 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.

While these facilities offer several advantages for phonetic transcription, the system does not solve all the 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. 'going to' versus 'gonna'). 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.

[6] 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.

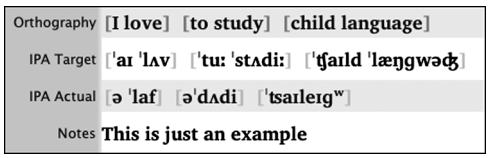

FIGURE 19.4 Example word groups.

19.3.1.4 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 Figure 19.4.[7]

The alignment of word groups across all relevant tiers also effectively provides a system for 'vertical' alignments of portions of the transcribed utterances, which facilitate data queries on specific subsets of the transcribed utterances. For example, queries can be restricted to word groups containing specific strings of symbols in a given tier, or groups which are located in a given position within the utterance (e.g. utterance-medial vs. final word groups).[8]

19.3.2 Automatized Data Annotation Systems

As discussed in Rose, this volume, 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.

19.3.2.1 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.

[7] Note also the absence of bracketing in the Notes tier, which is not group-aligned.

[8] 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.

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},[9] irrespective of other potential phone characteristics such as its voicing or exact labial articulation.

19.3.2.2 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.

19.3.2.3 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 Figure 19.5, relies on a 'best-guess' dynamic programming algorithm (Maddocks 2005).

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 Figure 19.5b, 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

[9] Feature sets in Phon are listed between braces; capitalization of the feature labels is optional.

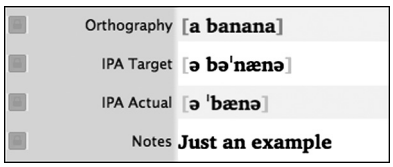

FIGURE 19.5a Phone alignment: Transcriptions.

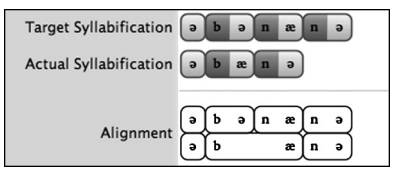

FIGURE 19.5b Phone alignment: Aligned phones.

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.

19.3.3 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 meta-data 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 meta-data 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 prosodic markers (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.

19.4 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.

19.4.1 Data Query

19.4.1.1 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.[10] 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 Figure 19.6, morpheme counts

[10] 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.

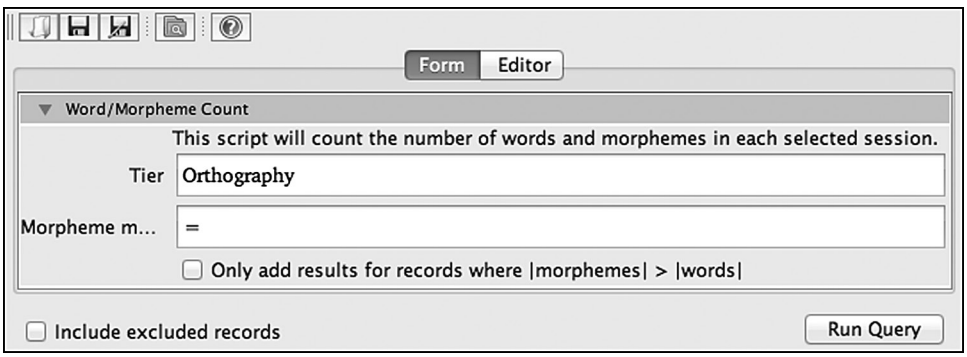

FIGURE 19.6a Example MLU search: Search parameters.

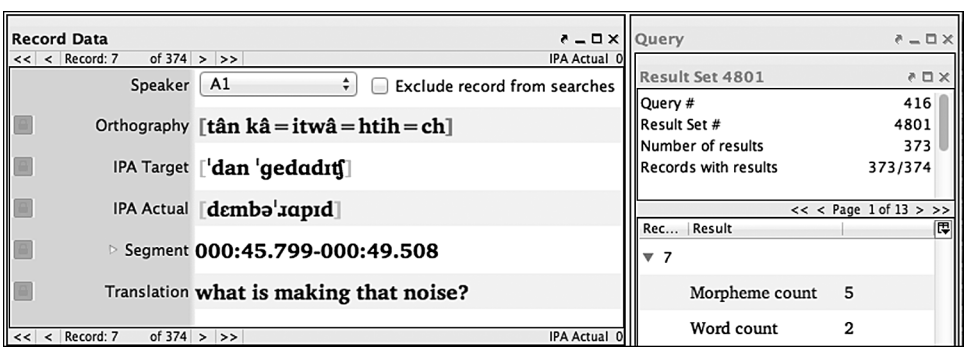

FIGURE 19.6b Example MLU search: Sample of per-record results visualization.

require explicit data coding, to identify morpheme boundaries. Morphemes are identified with the symbol '=' in the Orthography tier, as in Figure 19.6a; word boundaries are calculated based on the spaces that occur between the orthographically transcribed words. The screen shot in Figure 19.6b illustrates the visualization of one record alongside related search results, which contains two words and a total of five morphemes. Finally, the Word/Morpheme count function also generates mean values for each session selected for query, in Figure 19.6c. These latter results can easily be formatted in spreadsheets or other software for statistical analysis.[11]

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 et al. 2005 for further discussion of this measure).[12]

[11] This custom script can be downloaded from our public repository of scripts, available at: <https://www.phon.ca/phontrac/wiki/search/scriptlibrary>.

[12] We are currently investigating ways to best implement this algorithm into Phon.

```
Session: A1.A1-2006-10-18-v27
------------------------------------------------
      Word total: 735
       Word avg: 1.96524064171123
Morpheme total: 1032
 Morpheme avg: 2.7593582887700534
```

FIGURE 19.6c Example MLU search: Session summary results.

19.4.1.2 *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 and Davis 2009 on babbling development; Levelt et al. 1999/2000 on the acquisition of syllable types). The interface for such searches is illustrated in Figure 19.7.

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).

19.4.1.3 *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 and Davis 2009) to more abstract accounts of these abilities in terms of feature co-occurrence constraints (e.g. Levelt and 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

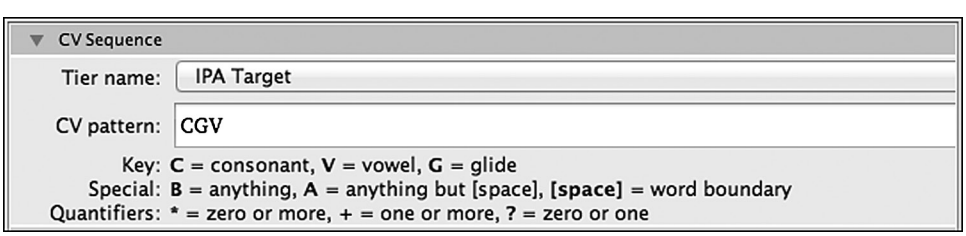

FIGURE 19.7 General syllable shapes.

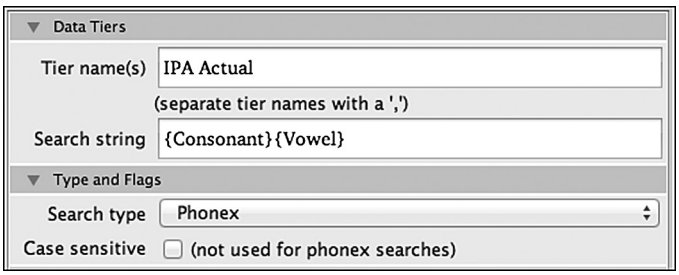

FIGURE 19.8 Query returning CV sequences found in IPA Actual forms.

(e.g. Fikkert and Levelt 2008; Kern and Davis 2009). As the screenshot in Figure 19.8 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.

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.

19.4.1.4 Feature-Based Queries

Like 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,[13] 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.

[13] See Rose and Inkelas (2011) for recent discussion.

19.4.1.5 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: [gʌk] 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 and Golston 2002, Inkelas and Rose 2007, McAllister 2009 and references therein) syllable-based labelling can be used to observe renditions of velar consonants across various positions. As Inkelas and Rose (2007: 707) observe, child E, a learner of English, produces coronals for target velars located in word-initial and stressed onsets (e.g. cake [ˈk̲eɪk] → [ˈt̲eɪk]; again [əˈg̲ɛn] → [əˈd̲ɛn]). However, velars in word-medial onsets of unstressed syllables and in codas are realized in target-like fashions (e.g. bagel [ˈbejg̲u]; back [ˈbæk̲] → [ˈbæk̲]). 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 Figure 19.9 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 searches can thus be achieved easily, through the setting of criteria within the search interface.

19.4.1.6 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

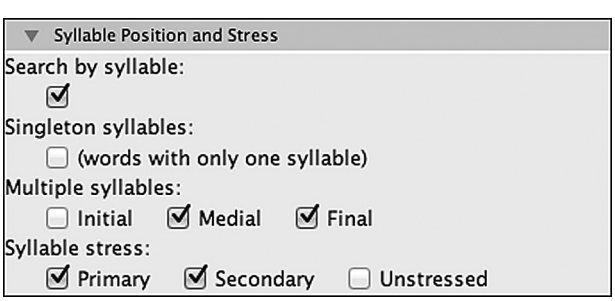

FIGURE 19.9 Context-sensitive detection of phonological patterns.

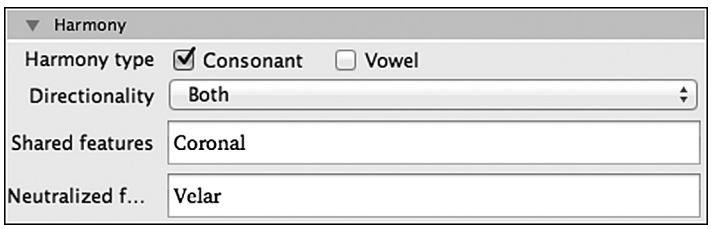

FIGURE 19.10a Consonant harmony: Search interface.

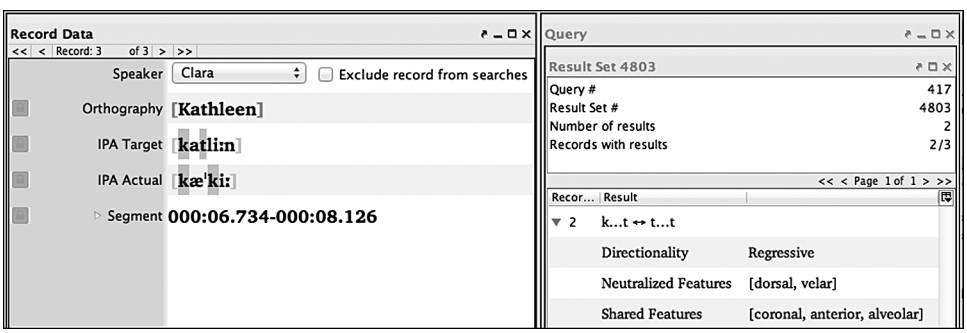

FIGURE 19.10b Consonant harmony: Sample result.

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 Figure 19.10, a specialized algorithm designed by Gedge et al. (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 Figure 19.10a, 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.

Records returned by these search parameters can be visualized alongside meta-data describing both the features involved in the pattern and its directionality, as in Figure 19.10b.[14] 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.

[14] As we can see in Figure 10b 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.

19.4.2 Data Reporting

As we saw in the last example, the results returned by Phon queries can be visualized directly within the application. These results can also be 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 Table 19.3.

As we can see in Table 19.3a, 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. Table 19.3b 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. The user could also identify the relevant set of records for further processing for example by exporting media clips 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 Table 19.4 with a representative example from child Catootje documented in the CLPF corpus.

As we can see in this example, the behaviour of Catootje's target [bʁ] 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

Table 19.3 Report sample

a. Inventory of patterns		b. Record data	
ʁ ↔ j	8	Result	ʁ ↔ ∅
ʁ ↔ w	1	Orthography	Gaspard
ʁ ↔ ʁ	8	IPA Target	[gaspɑʁ]
ʁ ↔ ∅	7	IPA Actual	[gapa:]

Table 19.4 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 ↔ bl^{-}	0	1	0	0
bʀ ↔ bʀ	0	0	1	4

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 meta-data 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. Voormann and Gut 2008 for further discussion).

19.5 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, itself integrated into Phon as a plugin.[15]

Conversions from the CLAN to the Phon for mat 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

[15] These utilities can be downloaded from <http://www.talkbank.org/software/>.

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.

19.6 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.[16]

[16] 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 Institutes 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 would not have been as fruitful without the tremendous feedback and support we received from the research community, for which we are grateful.