Corpus-based Investigations of Child Phonological Development:

Formal and Practical Considerations

Yvan Rose

Memorial University of Newfoundland

Acknowledgements:

I am grateful to Jacques Durand, Ulrike Gut and Gjert Kristoffersen for providing such a wonderful venue for the discipline of corpus-based research in phonology, as well as for inviting me to contribute to this important publication. I also owe special thanks to Elisabeth Delais-Roussarie for her useful feedback on an earlier draft of this chapter. I am also indebted to everyone behind the PhonBank initiative, including all the members of the Phon development team at Memorial University of Newfoundland and of the CHILDES project at Carnegie Mellon University, for their tremendous support over the last several years. Several aspects of the work discussed in this chapter have benefited from funding from the National Institute of Health, the National Science Foundation, the Canada Foundation for Innovation, the Social Sciences and Humanities Research Council of Canada, Memorial University of Newfoundland, and a Petro-Canada Young Innovator Award. Of course, all errors or omissions are my own.

Corpus-based Investigations of Child Phonological Development:

Formal and Practical Considerations

1. Introduction

Seminal works such as Jakobson (1941/1968) and the rise of generative and cognitive approaches to linguistics have provided researchers with a breath of fundamental questions about the human language faculty and the role of nature versus nurture in language and its development (e.g. Stampe 1969, Chomsky 1981, Pinker 1984, 1989, Elman, Bates, Johnson, Karmiloff-Smith, Parisi & Plunkett 1996, Tomasello 2003). In turn, many of these questions have fuelled the need to develop corpora fit for the study of language in general as well as more special-purpose corpora for the study of language acquisition or, more specific to this chapter, phonological development.

It is in this context that many of the first large-scale corpus studies took place, including the famous Providence study by Brown (1973). Concerning phonological development, pioneer work by Deville (1891), Smith (1973) and Compton & Streeter (1977) offered early grounds for broad-based investigations of child phonological data, albeit based on non-computerized means of data compilation. The unavailability of computerized methods at the time however imposed several limits on corpus building and data analysis. More generally, the discipline has been

hampered by technological issues, regarding for example the use of specialized character sets representing phonetic symbols such as those of the International Phonetic Alphabet (IPA). In spite of such limitations, and aside from works devoted specifically to learnability theory,[1] the vast majority of production-based studies of phonological development rely on some longitudinal or cross-sectional corpus data.

The need for corpus-based studies lies in the very nature of research in phonology, which requires the systematic study of segmental and prosodic patterns within and across various positions within the syllable, the word or larger prosodic domains such as the phrase, the intonational group or the utterance. The field of phonological development is concerned primarily with how children acquire these patterns. While certain studies, for example about the development of perceptual abilities, must follow tight experimental protocols, often independent of the learners' spoken abilities (if any, for example in the case of young infants), studies of the development of productive abilities generally rely on corpora of spoken forms, which range from experimentally elicited speech productions to recordings of undirected,

1 For further discussions of learnability in the context of phonology, see Dresher (1994), Dresher & van der Hulst (1995), Boersma (1998), Hale & Reiss (1998, 2008), Tesar & Smolensky (2000), Hayes & Wilson (2008), *inter alia*. While many of the discussions in learnability are theoretical in their nature, empirical tests for learnability models can be found in, e.g. Archibald (1993), Boersma & Hayes (2001) and Escudero & Boersma (2002).

spontaneous productions, many of which conducted in naturalistic settings (e.g. the learner's home).

Because the expression of phonological patterns is directly measurable in the physical domain, the substance of phonology can, and should, be addressed using the best observational methods available. Analyses of phonological data must also rely on strong theoretical underpinnings, in order to explain generalizations that transcend individual observations as well as to frame these observations within various grammatical domains such as those mentioned above. These claims are emphasized in a number of recent and inspiring papers highlighting the need, perhaps more important now than ever, to take advantage of the most advanced methods available, however without neglecting the more abstract, cognitive properties of these systems of knowledge representation and processing as well as the theoretical models required for their description (e.g. Goldsmith 2007, Durand 2009). The modern empiricist arguments formulated in these works convincingly demonstrate that it is in the contemporary development of our empirical knowledge that lies many of tomorrow's breakthroughs, both in theoretical phonology and in related experimental or more applied domains.

In the sections that follow, I wholeheartedly embrace this view. I argue, further, that one of the most efficient ways to progress as a field is through open data sharing coupled with unrestricted access to the tools required to explore phonological corpora. This argument breaks away from

much of the technologically challenged history of corpus-based research on phonology. I first discuss some of the issues that are central to research in phonological development. I then describe a few solutions, with an emphasis on the recently proposed PhonBank initiative (e.g. MacWhinney & Rose 2008; see also Rose & MacWhinney in Volume III of the Handbook) within the larger CHILDES project (e.g. MacWhinney 2000). Although aspects of the discussion will inevitably touch on some of the current theoretical debates in phonology, I endeavour to keep the ensuing sections as theory-neutral as possible. For further discussion of current theoretical debates in phonological development, see Rose & Inkelas (2011).

2. Issues in corpus-based research on phonological development

Research in phonological development is for the most part comparable to other kinds of phonological investigations. It focuses on the types of units that make up the learner's phonological system and how they manifest themselves in phonological productions. This research also requires a consideration of how phonological units interact with each other both in perception and production, as well as how they interface with other components of the learner's larger system (e.g. its morpho-syntax). The particular context of language acquisition, however, adds a time dimension to phonological descriptions. This further complicates research methodology, and can also impact the central issue of data interpretation. More specifically, developmental aspects of child phonological systems call for the consideration of issues such as the presence or absence of the units required to describe the target (adult) system and the

timing of their emergence or maturation during the course of development. Another difficulty inherent to phonological acquisition studies lies in the fact that children's productions often fluctuate, both within and across observational periods, be they calculated in days, weeks or longer time spans. I expand on these and related issues in the next two subsections.

2.1. Theoretical issues

From a theoretical perspective, the most controversial debate lies not in the existence of an inborn language faculty (no one can reasonably deny the fact that the human species is quite unique from the perspective of language and language acquisition), but on the understanding of this faculty. This chapter does not aim at addressing this controversial question. Regardless of the researcher's theoretical inclination in this pursuit, a number of challenges are likely to arise, each of which impacting on aspects of the research, and, potentially, its outcomes.

2.1.1. Phonological units

As in virtually all kinds of scientific endeavours, one of the factors that may have drastic implications for research outcomes pertains to the identification of the units which are relevant for research. The field of phonology enjoys a rich tradition of descriptive units, including the phonological segment (e.g. Pāṇini, circa 520–460 BC) and its internal components, the features, which have been firmly implemented into most phonological descriptions after Jakobson, Fant & Halle (1952). Interestingly, while the true nature of segments and features as well as their

psychological reality is still a matter of theoretical controversy (e.g. Bybee 2001, Vihman & Croft 2007), features are considered the most basic phonological units by a majority of scholars.

Since the advent of multilinear phonology (e.g. Kahn 1976, Goldsmith 1976) and, especially, the theory of Prosodic Phonology, a number of other, more phonetically elusive units have also been proposed, including the syllable and its constituents as well as higher-level categories such as the Foot, Prosodic Word, Phonological Phrase, and so on (e.g. Selkirk 1980a,b, 1982). In each case, arguments have been formulated, at times independent of phonetic content itself (e.g. onsets of empty-headed syllables; Kaye, Lowenstamm & Vergnaud 1990; cf. Piggott 1999). Proposals making reference to abstract phonological structure offer interesting avenues to analyze adult phonological systems as well as their development (e.g. Rose 2000, Goad & Brannen 2003 and Barlow 2003). As with any theoretical proposal, an appeal to abstract units however yields a number of fundamental questions, for example, about their psychological validity in phonological representation or processing, or their emergence in the learner's system. Questions such as these lie beyond the scope of this chapter. However, all studies of phonological development should minimally pay attention to the constructs offered by phonological theory, if only to test their relevance in the context of acquisition. This is particularly important in that a desirable theory of phonology (or grammar, more generally) must also be fully compatible with learnability issues, and testable on empirical grounds.

Even fairly concrete issues such as the use of phonetic transcriptions are subject to controversy. Above and beyond methodological issues (see below), central theoretical debates revolve around the status we should attribute to the units that phonetic symbols are designed to represent. Within the generative framework, researchers most often utilize phonetic symbols to refer to speech segments which are assumed to be psychologically valid abstractions for the set of units they represent in the speakers' phonological lexicon. Conversely, segments and/or features have virtually no place in exemplar-based approaches (e.g. Bybee 2001 and references therein). Yet, in a paradoxical way, even researchers who reject these constructs often make reference to them in their accounts of sound patterns in development and elsewhere (e.g. Vihman & Croft 2007). Minimally, this warrants the use of descriptive phonological units in ways that transcend debates about their theoretical interpretation.

Coming back to phonological features, it is common knowledge that natural groupings (classes) of speech sounds pattern together, in both adult languages and developing systems. Beyond articulatory descriptions, features are often used as labels to identify these natural classes. This can be illustrated in the context of developmental data through attested cases of long-distance consonant metathesis, whereby units such as consonants or consonantal features that are present in target (adult) forms are preserved but swapped across vowels in systematic ways. For example, as we can see in (1), child code-named W, a learner of English, produces every word-

initial target fricative in word-final position (original data by Leonard & McGregor 1991, as reported by Velleman 1996).

(1) Manner-conditioned metathesis patterns (Leonard & McGregor 1991)[2]

|zu|	[uz]	'zoo'
|faɪn|	[aɪnf]	'fine'
|sop|	[ops]	'soap'
|snupi|	[nupis]	'Snoopy'
|stɑp|	[taps]	'stop'

Independent from any analytical framework, the systematic study of phonological patterns becomes immensely easier through the use of labels enabling easy identification of descriptive units and positions (here, target word-initial fricatives produced in word-final position; see section 4 below for more discussion of positional effects).[3] From this perspective, phonetic transcriptions offer entry points into the phonological corpus, without a priori theoretical implications. (See below for a concrete application of this view within the context of PhonBank.)

2 Target forms are represented between vertical bars and actual forms between brackets.

3 As noted in various sections of this chapter, positional effects in phonological development, as in phonology more generally, can be observed within and across a number of grammatical domains within syllables, words or larger units, be they prosodically or syntactically defined.

2.1.2. Interaction between phonological units

Unit identification, while crucial, is only the first step toward the characterization of phonological systems and their development. Phonological accounts, regardless of their theoretical orientation, must also consider whether or how each unit identified interacts within the larger system. Phonology is indeed a complex system whose many components interact in often intricate ways (e.g. Chitoran, Pellegrino & Marsico 2009). We can indeed compare a phonological system to a refined diamond, every facet of which can be observed individually while also contributing to the larger whole. For example, contributions to Goad & Rose (2003) highlight how segmental behaviours observed in acquisition data are conditioned by prosodic aspects of developing phonological systems. Similarly, studies of prosody and prosodic development show clear conditioning at higher levels of grammatical organization (e.g. Delais-Roussarie 2005, Prieto 2011, Prieto, Estrella, Thorson & Vanrell 2012). In theory, such interactions may be explained in many different ways, for example through perceptual, articulatory or more representational factors. The corpus-based exploration of each of these avenues requires appropriate technological support, itself back by articulated theories enabling both the definition of the units involved as well as their integration within a coherent whole.

Corpus-based studies are also central to our understanding of phonological development in that they provide not only a basis for the qualitative observation of developmental patterns, but also

for the assessment of the relative importance of these patterns among and between individual learners, which also enable an assessment of the relative occurrence of these patterns both within particular languages and cross-linguistically. This is particularly relevant given the recent history of corpus-based investigations of phonological development. As discussed in section 3, it is indeed not until recently that scholars in the field have been in a position to make such assessments. In spite of recent progress in this area, typological descriptions of phonological development, especially from a longitudinal standpoint, remain generally limited to relatively few case studies, even for the best documented languages (e.g. Dutch, English, French or Portuguese). While these studies point to important observations about the course of development, much remains to be done in this area before we are in a position to grasp the full picture of how children acquire their phonological systems during their first years of life.

2.1.3. Development

In addition to the issues discussed above, which apply to virtually all types of phonological studies, acquisition studies come with their own sets of specific questions. Without going into the detail of data gathering protocols and issues that pertain to data sampling —all of which are inherent to every type of data-driven undertaking— the study of phonological development adds methodological components which are not typically needed in studies of steady-state phonological systems.

Perhaps the most important of these components is the time dimension. Aside from historical or socio-linguistic considerations, time is typically not an issue in phonological investigations.[4] The handling of this additional dimension depends on the type of study performed. For example, in longitudinal studies, time offers a crucial link between data observed across recording sessions that pertain to individual learners. In contrast to this, cross-sectional studies, which offer population-based assessments, generally preclude time-based tracking of grammatical states for any single participant. While general trends observed within and among data cross-sections do point to various aspects of grammatical development, most of the phonological patterns observed in these studies are difficult to interpret from a strictly grammatical perspective. The understanding of developmental patterns of phonological production indeed requires observations about their emergence (e.g. gradual versus categorical), details about their manifestation, including their stability (or potential variability) within and across word forms, the length of the period during which they are observed in the learner's productions, and the way in which they evolve, or disappear from the productions at the end of this period. Cross-sectional observations do not offer the type of real-time sampling required to address such questions. Outside of the establishment of developmental norms or similar population-based

4 Given well-known facts such as language change and, on what can be fairly small time scales, sociolinguistically driven patterns of variation, one cannot consider adult languages as completely steady-state. However, little is known about young children's sensitivity to micro parameters of variation, or how these parameters are acquired during the developmental period.

measures, it is thus generally ill-advised to assume a priori direct relationships between age and grammatical development. This is especially true given the important degree of variability observed between learners, throughout the developmental period, irrespective of the target language (e.g. Smit 1993, Jongstra 2003, Costa 2010).

2.2. Methodological challenges

Beyond theoretical issues, the study of phonological development also raises a number of methodological challenges. In this section, I discuss the issue of data transcription more in depth, in light of the related challenge of data assessment. Together, they provide strong motivation for open data sharing as well as for the development of scientific standards for data annotation, two topics which I address subsequently.

2.2.1. Data transcription

Corpus-based phonological studies typically rely on phonetically transcribed speech samples, where the level of phonetic transcriptions may vary from broad to relatively narrow. The interpretation of these data often relies, at least in part, on the assumption that the speech segments and related segmental or prosodic diacritics we represent symbolically in transcribed corpora are relevant to phonology and phonological processing. At a very basic level, the use of transcribed units such as speech segments and the various diacritics used to adorn them greatly facilitates the description of phonological patterns. However, as discussed in section 2.1.1, the

definition of these units can be subject to controversy. In addition, the use of a symbolic, thus inherently categorical, system may impose ad hoc splits within what would otherwise be data continua. This problem is further compounded by the fact that transcriptions are in the vast majority of cases based on impressionistic judgements performed by adult transcribers. As such, aspects of the speech signal may be filtered out by the transcriber's own perceptual system, potentially introducing unwanted biases to the transcriptions.

Such issues have been discussed in depth in the literature on covert contrasts (e.g. Scobbie, Gibbon, Hardcastle & Fletcher 1996) and related publications, most of them from proponents of laboratory approaches to phonology (see Munson, Edwards, Schellinger, Beckman & Meyer 2010 for a recent summary and discussion). The main criticisms against research relying on phonetic transcriptions point to the fact that human transcribers are subject to a number of perceptual biases that may affect speech perception and, consequently, its transcription. This is best demonstrated in studies which show that transcribers may perceive identical speech samples in different ways, depending on a series of factors such as their own linguistic background, their experience with phonetic transcription, visual cues, or even the transcribers' own expectations about the samples they are transcribing. The effects of such factors on transcription are also visible in studies of inter-transcriber accuracy, which show that even trained transcribers are not always able to accomplish their work in systematic ways (e.g. Edwards & Beckman 2008). Such issues are presumably even more problematic in the context

of phonological development, as language learners are often unable to render phonological contrasts or allophonic alternations in ways that mature/proficient listeners are likely to expect. Hence the possibility that phonological patterns produced by the learner may be partially or completely missed by the transcriber (e.g. Scobbie et al. 1996; see also the chapter by Post & Delais-Roussarie in the present volume).

Unfortunately, no solution currently exists for any of these problems. While laboratory phonologists often suggest that transcriptions should be backed by acoustic measurements, few go as far as rejecting transcriptions altogether. It is also important to note in this context that while acoustic measurements can indeed provide objective verifications of acoustic dimensions of speech, they simply cannot be considered a panacea. Foremost in this argument is the fact that reliance on acoustic measurements, despite their a priori objectivity, does involve some degree of impressionistic judgement as well, if only in the determination of which acoustic properties are relevant to the problem. For example, while first and second formant values are generally considered an absolute must in the assessment of vowel place of articulation, no such consensus seems to exist about reference to third (or fourth) formant values, the interpretation of which often depends on the types of vocalic contrasts that exist in the languages under investigation. In addition, the interpretation of acoustic measurements is often based on theoretical models yielding expectations about the types of relationships that must exist between speech acoustics and the articulations that underlie them. Such expectations are

however not always warranted in the context of child phonological development, given that established models are generally based on adult vocal tract sizes and configuration, and that young children's vocal tracts differ from these in significant ways (e.g. Fletcher 1973, Kent 1976, 1992, Kent & Murray 1982, Crelin 1987, Kent & Miolo 1995). Any study of child speech acoustics thus requires an adaptation of these models to the still-maturing properties of children's vocal tracts (e.g. Ménard 2002, Stelt, Zajdó & Wempe 2005; see also Lintfert 2009). Finally, even when all precautions are taken with regard to the relationships between acoustic correlates and related speech articulations, we must keep in mind that the datapoints obtained through instrumental analysis must often receive a relative interpretation. For example, fairly easy measures such as Voice Onset Time (VOT) must be assessed in relation to variables such as speech rate, prosodic position within the word and phrase, sentential focus, and so on. Such complex information can make the interpretation of acoustic data extremely challenging at times. Coming full circle, it is interesting to note as well that these complex computations are exactly what the human perceptual system is particularly good at, especially for cue combinations that are relevant to the transcriber's mother tongue. In this respect, human transcriber impressions, despite their inherent limitations, do provide a breath of information, some of which central to the study of phonological systems.

In light of these considerations, no one can claim to interpret speech data, whichever their nature, in a strictly objective way. It follows from this that researchers should select the method,

or combination of methods, that best fits their needs, and restrict data interpretation according to the limitations inherent to their methods, be they based on impressionistic transcriptions, instrumental measurements, or combinations of these or other means. Until computational systems become reliable enough to generate phonetic transcriptions or other types of data representations directly from audio-recorded data, we will not be in a position to do away with human transcriptions. In the meantime, we must embrace the problem of data representation in a pragmatic way, and strive to reduce its potential impacts on our interpretation of computer (and human) readable renditions of speech sounds.

2.2.2. Data assessment

Once the data are represented in some form (via phonetic transcriptions and/or other means), the next step in the research typically consists of scrutinizing them in light of the research questions at stake. From a strictly developmental perspective, studies of child phonology typically combine two types of measures: productivity and accuracy. While these two concepts are not inherently complex, their application in the context of child language raises a number of issues. For example, how can one describe phonological productivity? In terms of sheer number of attempts? In the variety of attempted, or successfully produced, units? Similarly, accuracy measures pose a number of challenges, as they generally require some evaluation of the phonological patterns observed in the learner's productions against a given norm.

In most corpus-based studies, transcriptions of the learner's attempted forms are compared against transcribed forms representing target (typically adult) renditions of the same forms. As discussed above, studies may also incorporate acoustic or articulatory descriptions of the target speech. While such comparisons are increasingly easy to attain through software-assisted methods (some of which are discussed below; see also the chapter by Rose & MacWhinney in Volume III of the Handbook), many questions remain on how to exactly assess patterns of phonological development (e.g. Ingram 1989; see also Preston, Ramsdell, Oller, Edwards & Tobin 2011 as well as references therein).

Irrespective of the methods and criteria favoured by the researcher, however, even accessing the vast amounts of data needed for systematic research often proves a daunting challenge. The obvious solution to this lies in data sharing, which I address in the next subsection.

2.2.3. Data sharing

Until recently, open access to corpora documenting phonological development was practically nonexistent. Except from a handful of corpora such as the CLPF corpus of phonological development in Dutch (Fikkert 1994, Levelt 1994), most corpora available offered very little support for large-scale investigations. This problem, however, was not related to the absence of such corpora as much as to various roadblocks preventing their access. In the paragraphs that follow, I formulate a brief argument in favour of data sharing, with the hope of encouraging the

research community to engage in further efforts towards broader and easier access to research data.

The foremost consideration in this argument consists of the scientific validity of corpus-based phonological investigations. While it is true, at least to some extent, that academic reputation during medieval times was assessed by how much knowledge someone 'possessed', a mentality which gave rise to the elaboration of (often obscure) data mines jealously protected from outsider's scrutiny, modern views of science generally reject such an archaic mentality. In virtually all fields of empirically driven science today, the disclosure of raw data is minimally considered good practice, while it is a strict requirement is most areas of 'hard' science. A comparison between astronomy and phonology is appropriate in this context, as both disciplines often discuss units (e.g. an extrasolar planet; an underlying phoneme) on the basis of indirect evidence (here, cosmic light effects; phonological patterning). In astronomy, confirmation of research results —and their publication— requires the disclosure of all of the raw evidence for independent assessment. The proof is in the data. Unfortunately, and for no other reason than established research traditions in phonology (and many areas of linguistics), peer-reviewed appraisals of phonological investigations seldom require disclosure of the raw data. As a result, it is rare that one can evaluate an analysis beyond the evidence provided by its proponent(s). This obviously limits the scope of the review. While this practice facilitates the publication process, it certainly does not help the reinforcement of our scientific standards.

Our research practices should, in an ideal world, incorporate full data disclosure, despite two basic objections often heard in this context. The first relates to legal or ethical considerations, typically about the absence of the informed consent required for the publication of the corpus. While presumptions about such legalities often impose a priori constraints on data release, it is important to note that such matters can be successfully addressed in most academic jurisdictions. This view is in fact supported by many research institutions and granting agencies, which are now requiring the release of non-sensitive data by the end of the funded project. This practice can, and should, be extended to all newly funded initiatives. The second objection relates to the costs involved in the preparation of corpora for public release, too often considered a valid reason for 'private' ownership. I contend that this matter must be considered in light of the long-term financial advantages of data sharing. A quick consideration of a medieval-like situation suffices to illustrate this view. Imagine a group of 20 independent, medieval researchers, each one occupying a secluded data stronghold containing five gold coins worth of data. Research in this imaginary (though not so remote) world can only progress at a very slow pace, if at all, because the work produced by these researchers can be based on no more than five-coin datasets. Time-porting our medieval researchers into a modern-day world blessed with open data sharing would immediately give each of them access to 20 times more data, and thus able to draw research results from a significantly more extensive empirical foundation. The quality of the research generated would benefit from stronger empirical

grounding as well as from more open competition between the researchers, for example about the peer assessment of their proposals.

Scientific competition should be about ideas, not data. An empirically level field of scientific competition can only foster innovations which, if combined with stringent peer reviewing, offers all the seeds to propel the field towards new horizons. While breakthroughs were possible in the medieval ages, it is mostly because most areas of research were still in their infancy, let alone the development of empirical and rational methods of investigation, or even the promotion of the idea of science. In contrast to this, the investigation of modern-day, typically subtler research questions often requires the consideration of large and diverse sources of data. Hence the absolute need to openly share our empirical knowledge.[5]

Generalized data sharing is also warranted on more practical grounds. It is a euphemism to say that we live in a world with limited funding available for linguistic research. In light of this, it is easy to balk at the time and resources involved in the audio/video recording, transcription and coding of multimedia corpora. However, the most appropriate reaction to this challenge should be towards an optimization of the returns on the costs. As data sharing offers free access

5 Of course, there are corpora that are sensitive in nature (e.g. clinical) and whose anonymity cannot be sufficiently protected within a public database. It certainly makes sense that such corpora be kept from free public access. Crucially, however, the peer review process should never be compromised.

to empirical evidence against which we can evaluate research hypotheses, any shared corpus generally reduces the costs of this research to data assessment itself, much of which can be alleviated through computational means of data processing, unless additional coding is required for hypothesis testing. In this latter case, the investment into the 'new' coding should be contributed as part of an augmented version of the original corpus, such that it can in turn translate into an investment for future research.[6] Societal arguments can also be formulated in this context, given that most research funding comes from public sources.

Concrete examples of the virtues of data sharing are numerous. In the context of acquisition studies, the CHILDES project has supported, through data sharing, thousands of publications in various fields of theoretical and applied linguistics, psychology, computational modelling and a number of related disciplines. Concerning phonological development, corpora such as that of Smith (1973), originally published as an appendix, and the CLPF corpus of Dutch development mentioned above, are referred to in some form or other in hundreds of publications. These corpora, which served as precursor models demonstrating the virtues of data sharing, have recently been incorporated into the PhonBank database (http://childes.psy.cmu.edu/phon/), which now offers new ways to explore them.

---

6 This possibility also opens the issue of data 'authorship'. However, as in all cases of potentially sensitive issues discussed in this chapter, a healthy combination of goodwill and pragmatism (e.g. shared authorship, corpus versioning) can bring solutions to virtually all problems.

Despite the optimistic vision of data sharing expressed above, some challenges remain. Indeed, the open sourcing of data can also yield counter-productive issues, a few of which are discussed in the next subsection, alongside working solutions.

2.2.4. Data standardization

Throughout the steps involved in corpus building, the researcher often makes decisions which, while they are perfectly sensible on an individual basis, can create incompatibilities between data sets. For example, even the use of a standardized transcription system (e.g. the IPA) does not guarantee compatibility at all levels of coding in the corpora, primarily because of the conventional nature of symbolic representations. While this and other technical issues can be addressed on a case-by-case basis, any combination of them poses as many complex problems. This often prevents ready compatibility across datasets, which undermines the incorporation of individual corpora into large data pools for broad-based investigations. Quite fortunately, many solutions to this challenge have emerged in recent years, most of them connected to the democratization of personal computing, the development of open formats, and the advent of the Internet.

The best solution to such problems is in fact similar to the solution to data lockup itself. It requires the establishments of open standards, themselves developed by the scientists involved

in data sharing. Such standards have begun to emerge in recent years. For example, the CHILDES database is now fully compatible with the TalkBank XML format. While this format is administered from within the TalkBank project (http://talkbank.org/), it is developed in close collaboration with researchers in the field. Similar efforts have been made through research consortiums such as the Text Encoding Initiative (http://www.tei-c.org/).[7] Beyond these success stories, the establishment of standards for corpus building and sharing remains a must. With powerful technology now readily accessible, we are set to see rapid changes in methods of phonological investigation. I discuss this potential in the next section, where I focus more narrowly on the corpus-based study of child phonology.

3. Corpus-based research in phonological development

As mentioned in section 2, corpus-based research in phonology has traditionally been poorly supported. This holds especially true of child language data. While the advent of computerized systems in the 1980s set the stage for more powerful means of corpus-wide data coding and querying, early advances in these areas were more beneficial to morphological and syntactic investigations than to phonological research. Words and morphemes are, from a technical perspective, easier to represent than speech sounds as the former can often be achieved through

7 Concerning phonology in particular, discussions of an open format for phonological data encoding have taken place during the *Workshop on Developing Standards for Phonological Corpora*, which was held in summer 2009 at the University of Augsburg.

established orthographic systems. Because of this, even sophisticated database systems such as CHILDES traditionally lacked much of the support required for phonology. Three recent developments from computer science have offered keys to solve these issues: an industry-wide adoption of standards such as Unicode character encoding, XML data specification formats and, more generally, the rise of the open-source movement. Together, and in combination with advances in phonological theory, which provide a rich array of labels enabling articulated descriptions of sound patterns, these technologies enable powerful applications for the corpus-based study of phonology, many of which are now implemented within the PhonBank initiative. I provide an outlook of this project below. In order to provide the relevant background, I first revisit, if only briefly, some concrete issues affecting the systematic investigation of phonological development.

3.1. Challenges in the characterization of developmental patterns

Many studies of adult phonological systems rely on paradigms of transcribed data describing (morpho-) phonological alternations. Such paradigms often come from structuralist descriptions, many of which are documented through systematic methods of data elicitation. In contrast to this, studies of phonological development require the compilation of data that can be relatively difficult to organize at times. For example, data elicitation techniques are generally limited to picture naming or word repetition. Most recordings of child language are in fact entirely or

partially undirected, following naturalistic methods of data gathering. This yields inherent complications, for example concerning the basic structuring of data corpora.

As discussed above, a significant challenge posed by phonological development consists of identifying the relevant units and analyzing both their emergence and the ways in which they interact with one another within recording sessions and, in the case of longitudinal studies, across multiple sessions. As discussed in section 2.2.1 in the context of transcription, some phenomena that occur in child phonological data can be difficult to represent, let alone to even detect. Because of this, no a priori data classification can be envisioned. Another major difference that exists between child and adult phonology is that the former is inherently more transitional than the latter. As noted above, while the phonology of adult languages does evolve over their history, the nature and rate of this evolution can generally be measured over relatively large time spans (e.g. generations of speakers). In contrast to this, child phonological systems can display systematic differences within weeks or even days (e.g. Smith 1973, Levelt 1994, Fikkert 1994, Freitas 1997, Pater 1997, Rose 2000, Inkelas 2003, Inkelas & Rose 2008, McAllister 2009). However, stage-like effects are seldom clean-cut; conservative and innovative patterns of speech production often overlap within individual recording sessions, which makes child phonological systems often seem like complicated sets of moving, fuzzy targets to describe.

3.2. Software-assisted methods for data annotation

As already discussed, we cannot eliminate the human factor in tasks such as phonetic transcription. However, for all data that can be derived automatically, the favoured approach should be to use computer-assisted methods, and restrict human intervention to annotation verification, whenever necessary. Indeed, if we were to pit the human coder against any computer to perform tasks that can be automatized through algorithms, for example in the assignment of syllable constituent labels to consonants and vowels present in the transcript, the computer would easily win both in speed and consistency. Even in cases where the computer generates erroneous labelling, such errors can typically be associated to a particular algorithmic behaviour in response to a given type of data input; appropriate fixes to the algorithm eliminate all relevant errors once the data are scanned anew. In many cases, the algorithms can even be designed to explicitly flag potentially problematic cases.

Each time a computer-assisted method of data annotation can be envisioned, this solution should thus have prevalence. On this view, phonetically transcribed corpora of child speech should minimally consist of a set of tiers containing the target words in orthographic and in IPA transcriptions, the latter paired with the IPA transcriptions of the child's actual renditions of these target forms. Optionally, tiers containing other types of annotations that cannot be derived from orthographic or phonetic transcriptions should be added as needed (e.g. to describe utterances produced spontaneously or through imitation). Conversely, data annotations

that can be derived by means of algorithms should not require manual entry. For example, in all research making reference to descriptive phonological features, the features can be derived from articulatory descriptions of IPA symbols and diacritics. In a similar way, descriptions pertaining to syllable structure or stress should be obtained directly from the transcriptions, given the generally predictable nature of these aspects of phonological systems. This approach to corpus annotation is the one embraced in the design of the Phon software application, which provides most of the technological support behind the PhonBank public database, to which I turn next.

4. The PhonBank initiative

PhonBank seeks to broaden the scope of the current CHILDES database system to include the analysis of phonological development in first and second languages for language learners with and without language disorders. This initiative, which is backed by an international consortium of researchers, encompasses the visions expressed above for computer-assisted investigation of phonological development as well as the open sourcing of corpus data.[8]

8 While most software support required for PhonBank is included in Phon, additional issues related to data sharing and format standardization are handled through the CHILDES database, via the TalkBank XML data format, handled through utilities such as CLAN2XML and XML2Phon.

PhonBank offers an open platform for phonological data exchange, in accordance with CHILDES principles and rules of conduct. Such principles include, from a technological standpoint, reliance on open standards and maximum availability of both the data and the tools required for their processing. For example, collegial decisions involving members of the scientific consortium provide the basis for the development of practical solutions for the databasing and querying of phonological annotations. Also, in the context that relatively few child language researchers have formal training in computer science, one important component of the open access philosophy consists of facilitating access to even the most advanced technology. This is achieved through Phon's availability as free, open source software as well as its convivial graphical user interface. In order to best serve the needs of PhonBank, Phon was also designed to facilitate software-assisted approaches to data annotation. I illustrate concrete implementations of this vision below, after a brief overview of some precursor applications. The ensuing illustrations focus on the design of Phon, rather than its use. For more information about PhonBank and the full workflow supported in Phon, the interested reader is invited to consult the chapter by Rose & MacWhinney in Volume III of the Handbook.

4.1. Precursor solutions

A fair number of attempts have been made to address some of the methodological needs described in preceding sections. These include spreadsheet layouts as well as several different database systems. I briefly overview some of these systems in the next paragraphs.

Two independent in-house projects, both of them coincidentally named ChildPhon, were designed during the early 1990s. These systems, one of them developed in the Netherlands, the other in Canada, were both programmed within proprietary database layout systems (4th Dimension and FileMaker Pro, respectively) and used in research on the acquisition of Dutch, Portuguese and Québec French (e.g. Fikkert 1994, Levelt 1994, Freitas 1997, Rose 2000). These databases offered basic templates for textual (orthographic and IPA) annotations. They were however plagued by technical limitations. For example, none of these applications offered time-aligned information between the data transcript and the original recording. Also, because they were based on ASCII fonts, none of the IPA data supported by these systems were compatible across computer platforms. The development of these systems has since been abandoned (see Rose 2003a for further discussion).

The LIPP system (Logical International Phonetics Program) is commercial software that provides a wide array of functions for phonological analysis. These include IPA support as well as relatively powerful means to compile datasets based on these transcriptions. However, LIPP was built nearly 15 years ago and has not significantly changed in the ensuing years. For example, LIPP offers no multimedia capability; it is also limited to Windows operating systems. Finally, because data transcribed within LIPP are stored in a complex and proprietary format, it is difficult to envision the use of this software for data sharing initiatives.

A more recent commercial application for phonological analysis is CAPES (Computerized Articulation and Phonology Evaluation System), an application developed for clinical purposes and related research. CAPES enables the phonetic transcription of speech productions as well as some comparisons between these and corresponding target forms. While it does offer multimedia support, CAPES is, similar to LIPP, a Windows-only system. It also has a closed and limited data structure, which hampers its usefulness for database building.

Finally, the EMU system, described in Volume III of the Handbook, offers an appealing combination of facilities for phonological as well as for acoustic investigations of speech corpora. Among other possibilities, EMU supports the compilation of acoustic measurements based on positional criteria (e.g. positions within the syllable, word, phrase or utterance). Freely available as open source software, EMU offers cross-platform data compatibility, however through the pseudo-IPA SAMPA transcription system.

Despite their functionalities, some of which are rather powerful, especially in the case of LIPP and EMU, the systems described above also present a number of shortcomings. First, these applications require hefty amounts of manual data entry that could be avoided altogether, as they lack algorithms to derive predictable phonological annotations. Second, to the exception of EMU, these systems contradict the open source philosophy in that they are in part or fully based

on commercial, proprietary technology. My aim is however not to emphasize a critical view of these systems; they all have intrinsic value in that they all offer innovative approaches to corpus-based investigations of child language. In the next section, I turn to a recent addition to this family of applications, the Phon software program (Hedlund & O'Brien 2004),[9] whose development has benefited from lessons learned from all of its precursors. It is important to note in this context that Phon does not aim at replacing any of these systems; rather, it supplements these systems through a new combination of methods and related facilities.

4.2. Some illustrations

In this section, I discuss how some of the concrete problems discussed above are addressed within Phon. While the emphasis is on this specific software program, the discussion can be extended to virtually all computerized solutions to corpus-based approaches to the study of phonological development that are currently conceivable (e.g. Kunath & Weinberger 2009 for an alternative software design targeting similar issues). More detailed descriptions of the functionality supported by Phon are provided in Volume III of the Handbook.

9 While the original design of Phon is documented in Hedlund & O'Brien (2004), additional descriptions of the application's functionality can be found in Rose, MacWhinney, Byrne, Hedlund, Maddocks, O'Brien & Wareham (2006) and Rose, Hedlund, Byrne, Wareham & MacWhinney (2007) as well as in Rose (2008).

4.2.1. Computer-assisted data annotation

As suggested in section 3, an efficient database is one that does not require manual coding for each and every aspect of the data that may be relevant for research. Given the generally compositional nature of phonological systems, much of the phonological information useful for data compilations can be derived directly from the phonetic transcriptions. In the next subsections, I discuss how some methods of data annotation in Phon capitalize on this fact.

The design of Phon builds upon the key observation that in virtually all representational theories of phonology, phonetic symbols are associated with a series of descriptive features. Phonological theory thus provides the grounds for automatic systems of data labelling on a sizeable number of phonological dimensions. Such algorithms are available in Phon, which can derive rich information about segmental and prosodic patterning without the need for specific annotation beyond phonetic transcription (e.g. Rose et al. 2006). Within Phon, each phonetic symbol and diacritic is associated with a set of descriptive features and segmental strings are labelled for position within the syllable (e.g. onset versus coda).[10] Strings of symbols can also be labelled for syllable information based on their relative position within the string, with groupings of stressed and unstressed syllables forming metrical constituents describing stress patterns and other prosodic properties of word forms. For example, using this system, the

10 Other syllable labels and positions within the string can also be obtained.

researcher can specify a query making reference to obstruent consonants located in syllable onsets. Similarly, the status of the syllables with regard to stress or position within the word can easily be specified as well.

As discussed in section 2.2.2, the particular, developmental nature of child language phonology requires an assessment of development over time. For longitudinal studies, this implies the tracking of phonological performance across recording sessions; in the case of cross-sectional studies, group performance can be assessed within or across populations, age groups, and so on. Whichever the approach selected by the researcher, performance evaluation can be addressed through comparisons between target (adult/model) and actual (child/produced) form. Phon supports this approach through an algorithm that performs best-guess segmental alignments between corresponding target and actual IPA forms (Maddocks 2005). This alignment system is illustrated in (2), where we can see that the phones transcribed in the target form *Gaspard* [gasˈpɑʁ] automatically align with the most plausibly corresponding sets of actual phones [baˈpæː] despite the distortions observed between target and actual forms.[11]

11 See Kunath & Weinberger (2009) for a description of an alternative system of pair-wise phonological alignment.

(2) Automatic target-actual phone alignment system

Orthography:	*Gaspard*					
IPA Target:	ɡ	a	s	ˈp	ɑ	ʁ
IPA Actual:	b	a		ˈp	æː	

Of course, the user can ultimately modify all automated data annotations provided by Phon (e.g. features, syllabification and target-actual IPA alignments). User input is thus restricted to a bare minimum, in spite of the unavoidable resort to user-performed transcriptions and other annotations which cannot be derived through algorithmic means.

4.2.2. Complex phonological queries

Beyond basic documentation, the goal of virtually all corpus-building initiatives typically consists of the systematic extraction of results offering a empirical basis for scientific observation or hypothesis testing. As discussed in section 2, phonological systems can often be decomposed into subsystems which interact with one another in systematic ways. This implies that we cannot limit ourselves to assessing performance on individual units. Queries combining multiple criteria are paramount, in order to best characterize the systematic relationships that exist between the various facets of the phonological systems under investigation.

An example of phonological patterning whose study requires the simultaneous consideration of multiple conditioning factors is positional velar fronting. This pattern consists of the realization of target velar consonants as coronals in prosodically strong positions (e.g. in word-initial or otherwise stressed onsets) but not in weak positions (e.g. non-initial onsets of unstressed syllables; codas), as exemplified in (3).

(3) Positional velar fronting (data from Inkelas & Rose 2008)

a. Prosodically strong onsets

[ˈtʰʌp]	‘cup’	1;09.23
[ˈdoː]	‘go’	1;10.01
[ˈhɛksəˌdɔn]	‘hexagon’	2;02.22

b. Prosodically weak onsets; codas

[ˈmɑŋki]	‘monkey’	1;08.10
[ˈbejgu]	‘bagel’	1;09.23
[ˈpædjɔk]	‘padlock’	2;04.09

Data compilations of such patterns are made tremendously easier in Phon, which supports a wide variety of search criteria as well as the possibility to combine these criteria within individual queries.

4.2.3. Pattern detection

Patterns such as positional velar fronting can be described and, thus, queried, in very specific ways. However, data mining operations at times require more encompassing searches, for example to detect the presence of phenomena in a corpus, in a preliminary step leading to more specific data compilations. The study of consonant harmony (e.g. Smith 1973, Spencer 1986) provides a good illustration of this. Consonant harmony consists of the sharing of features between consonants across intervening vowels. For example, in a word like 'duck' produced as [gʌk], we observe a neutralization of the place of articulation of the first consonant ([d]), which is substituted by the velar articulation that independently appears on the second consonant in the target form ([k]). While consonant harmony is mostly discussed in light of place of articulation, it can manifest itself in a number of different ways across children. It can for example affect manner instead of place features (e.g. dos Santos 2007, Altvater, dos Santos & Fikkert 2010). Further, the manifestation of consonant harmony can be subject to a certain amount of variability, especially during transitions between developmental stages (e.g. Rose 2000, Pater & Werle 2003, dos Santos 2007). Given all of these observations, the study of this and other similar patterns of phonological productions can be particularly challenging. However, consonant harmony is both robustly attested in the literature and theoretically intriguing enough to be worthy of investigation, a fact evidenced by the rich and controversial body of literature devoted to it. For example, the question as to whether vowels intervening

between 'harmonizing' consonants are relevant to the pattern is still subject to controversy (e.g. Fikkert & Levelt 2008 for a recent discussion) and, thus, warrants further research.[12]

Given the non-local nature of consonant harmony, its detection requires simultaneous comparisons between aligned target and actual consonants across two or more positions. Despite its complexity, this and similar patterns involving relations between non-adjacent positions (e.g. vowel harmony, consonant metathesis) lend themselves particularly well for computer-assisted algorithmic detection. Such systems have already been implemented within the Phon query system (e.g. Gedge, Hedlund, Rose & Wareham 2007). Using these algorithms, the researcher can have a first, relatively blind, pass at the data, in order to detect the presence of consonant harmony. If patterns are detected, further searches can be performed to narrow down the results until proper data description is obtained. Algorithmic detection thus enables quick and reliable observations which, were they performed manually, would require significantly more time and efforts. Again here, the appeal to computer-assisted methods alleviates the need for any substantial user input beyond phonetic transcription and verification of computer-generated alignments between IPA target and actual forms.

---

12 See Smith (1973), Stemberger & Stoel-Gammon (1991), Macken (1992, 1995), Goad (1997, 2001), Dinnsen, Barlow & Morrissette (1997), Pater (1997), Bernhardt & Stemberger (1998), Rose (2000, 2003b), and Dunphy (2006) for additional characterizations of consonant harmony and related discussions. See also Hansson (2001) for an excellent discussion in the context of adult languages.

4.2.4. Sharing all the way

The skeptical reader may infer from the above that the system of data compilation in Phon is more reflective of current knowledge than offering functionality for future research, since it has been inspired by our current knowledge of child phonological patterning. This, however, could not be farther from the truth. This system is indeed flexible enough to enable virtually all types of queries that one can logically imagine from transcribed corpora of child language. It is also noteworthy that Phon offers tremendous potential for observational breakthroughs. This is particularly true given that most of our current empirical knowledge comes from a relatively limited number of studies, the scope of which has traditionally been limited by technological obstacles. On an optimistic note, now that many of these obstacles have been overcome, and given the unprecedented availability of data through PhonBank, the potential for advances in the field is perhaps better now than it has ever been.

In line with the vision expressed throughout this chapter in favour of open data sharing, Phon has been designed with concrete needs related to the PhonBank data sharing initiative in mind. First, all data transcribed and annotated within Phon are transportable across contemporary Mac OS X, Windows and Linux platforms. Second, in order to further facilitate collaborative work and contribute to the open sourcing effort, the query scripts can be shared between researchers, who can build, modify and share them as needed. As a result, data compilations

can now be published in combination with the methods used to obtain them, which tremendously facilitates post hoc verifications of the evidence. Given the data facilities built into Phon, and the infrastructure for data sharing supported by PhonBank, compelling solutions are now readily available for generalized data sourcing.

5. Future outlook

In this chapter, I addressed a series of current theoretical and related methodological challenges affecting research on child language acquisition. I then discussed some concrete problems in the context of computer-assisted, corpus-based investigations of phonological development. Throughout the chapter, I highlighted the fact that open sharing of both the data and the technological means to explore them offers some of the best potential to propel the field of research on phonological development towards new horizons.

Similar to the current trend in research on adult phonological systems, we observe in acquisition studies a convergence between the traditionally more distant fields of phonetics and phonology. This is also an important direction in which Phon and PhonBank are set to evolve. The vision entertained at this stage is to expand the Phon data structure to incorporate acoustic measurement data obtained from open-source software for speech analysis such as Praat and

Speech Filing System.[13] Among other benefits, this evolution will facilitate the investigation of the important relationships that exist between speech acoustics and phonological conditioning. More generally, it will open additional avenues to improve methodological aspects of corpus-based phonology, for example in enabling easier verifications of transcription accuracy through instrumental speech analysis. Assuming that models of speech acoustics can be reliable enough for the study of child language, this outlook incorporates much of the vision that phonetic symbols be used essentially as gateways into more subtle aspects of the phonological corpus.

Finally, in this and other similar projects, reliance on open scientific standards is an absolute must, hence the need to develop our research methodologies and the tools supporting them in a collegial way. It is indeed my hope that public forums for methodological advances as well as open access to the empirical base become the norm in the field, in order to maximize the outcomes of our scientific explorations of phonological development.

---

13 More information about these two projects can be found at the following web addresses:

— Praat: http://www.fon.hum.uva.nl/praat/

— Speech Filing System: http://www.phon.ucl.ac.uk/resource/sfs/

References

Altvater, N., C. dos Santos & P. Fikkert (2010), A Cross-linguistic Perspective on the Role of Prosodic Structure in the Acquisition of Manner of Articulation Features. Unpublished Manuscript.

Archibald, J. (1993), *Language Learnability and L2 Phonology: the Acquisition of Metrical Parameters*. (Dordrecht: Kluwer Academic Publishers)

Barlow, J. (2003), Asymmetries in the Acquisition of Consonant Clusters in Spanish. *Canadian Journal of Linguistics* 48, 179-210.

Bernhardt, B. & J. Stemberger (1998), *Handbook of Phonological Development from the Perspective of Constraint-Based Nonlinear Phonology*. (San Diego: Academic Press)

Boersma, P. (1998), *Functional Phonology. Formalizing the Interactions between Articulatory and Perceptual Drives*. (The Hague: Holland Academic Graphics)

Boersma, P. & B. Hayes (2001), Empirical tests of the Gradual Learning Algorithm. *Linguistic Inquiry* 32, 45-86.

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

Bybee, J. (2001), *Phonology and Language Use*. (Cambridge: Cambridge University Press)

Chitoran, I., F. Pellegrino & E. Marsico (eds.) (2009). *Approaches to Phonological Complexity*. (Berlin: Mouton de Gruyter)

Chomsky, N. (1981), *Lectures on Government and Binding*. (Dordrecht: Foris)

Compton, A. & M. Streeter (1977). Child Phonology: Data Collection and Preliminary Analyses. *Papers and Reports on Child Language Development* 7, 99-109.

Costa, T. da (2010), *The Acquisition of the Consonantal System in European Portuguese: Focus on Place and Manner Features*. Ph.D. Dissertation, Universidade de Lisboa.

Crelin, E. (1987), *The Human Vocal Tract: Anatomy, Function, Development, and Evolution*. (New York: Vantage Press)

Deville, G. (1891), Notes sur le développement du langage II. *Revue de linguistique et de philologie comparée* 24, 10-42; 128-143; 242-257; 300-320.

Delais-Roussarie, E. (2005), Interface phonologie/syntaxe: des domaines phonologiques à l'organisation de la grammaire. In Durand, J., N. Nguyen, V. Rey & S. Wauquier-Gravelines (eds.), *Phonologie et phonétique: approches actuelles*, 159-183. (Paris: Hermès)

Dinnsen, D., J. Barlow & M. Morrisette (1997), Long-distance Place Assimilation with an Interacting Error Pattern in Phonological Acquisition. *Clinical Linguistics & Phonetics* 11, 319-338.

Dresher, E. (1994), Child Phonology, Learnability, and Phonological Theory. Ms.

Dresher, E. & H. van der Hulst (1995), Global Determinacy and Learnability in Phonology. In Archibald, J. (ed.), *Phonological Acquisition and Phonological Theory*, 1-21. (Hillsdale, NJ: Lawrence Erlbaum)

Dunphy, C. (2006), *Another Perspective on Consonant Harmony in Dutch*. M.A. Thesis, Memorial University of Newfoundland.

Durand, J. (2009), On the scope of linguistics: Data, intuitions, corpora. In Kawaguchi, Y., M. Minegishi & J. Durand (eds.), *Corpus Analysis and Variation in Linguistics*, 25-52. (Amsterdam: John Benjamins)

Edwards, J. & M. Beckman (2008), Methodological Questions in Studying Consonant Acquisition. Ms.

Elman, J., E. Bates, M. Johnson, A. Karmiloff-Smith, D. Parisi & K. Plunkett (1996), *Rethinking Innateness : A Connectionist Perspective on Development*. (Cambridge, MA: MIT Press)

Escudero, P. & P. Boersma (2002), The Subset Problem in L2 Perceptual Development: Multiple-Category Assimilation by Dutch Learners of Spanish. In Skarabela, B., S. Fish & A. Do (eds.), *Proceedings of the 26th Annual Boston University Conference on Language Development*, 208-219. (Somerville, MA: Cascadilla Press)

Fikkert, P. (1994), *On the Acquisition of Prosodic Structure* (HIL Dissertations in Linguistics 6). (The Hague: Holland Academic Graphics)

Fikkert, P. & C. Levelt (2008), How does Place Fall into Place? The Lexicon and Emergent Constraints in Children's Developing Grammars. In Avery, P., E. Dresher & K. Rice (eds.), *Contrast in Phonology: Theory, Perception, Acquisition*, 231-268. (Berlin: Mouton de Gruyter)

Fletcher, S. (1973), Maturation of the Speech Mechanism. *Folia Phoniatrica* 25, 161-172.

Freitas, M.J. (1997), *Aquisição da Estrutura Silábica do Português Europeu*. Ph.D. Dissertation, University of Lisbon.

Gedge, J., G. Hedlund, Y. Rose & T. Wareham (2007), Natural Language Process Detection: From Conception to Implementation. Paper presented at the *17th Annual Newfoundland Electrical and Computer Engineering Conference (NECEC)*, St. John's NL.

Goad, H. (1997), Consonant Harmony in Child Language: An Optimality-theoretic Account. In Hannahs, S.J. & M. Young-Sholten (eds.), *Focus on Phonological Acquisition*, 113-142. (Amsterdam: John Benjamins)

Goad, H. (2001), Assimilation Phenomena and Initial Constraint Ranking in Early Grammars. In Do, A., L. Domínguez & A. Johansen (eds.), *Proceedings of the 25th Annual Boston University Conference on Language Development*, 307-318. (Somerville, MA: Cascadilla Press)

Goad, H. & K. Brannen (2003), Phonetic Evidence for Phonological Structure in Syllabification. In Weijer, J. van de, V. van Heuven & H. van der Hulst (eds.), *The Phonological Spectrum, Vol. 2*, 3-30. (Amsterdam: John Benjamins)

Goad, H. & Y. Rose (eds.) (2003), Segmental-prosodic Interaction in Phonological Development: A Comparative Investigation. Special Issue, *Canadian Journal of Linguistics* 48(3/4): 139-452.

Goldsmith, J. (1976), An Overview of Autosegmental Phonology. *Linguistic Analysis* 2, 23-68.

Goldsmith, J. (2007), Towards a New Empiricism. Ms.

Hale, M. & C. Reiss (1998), Formal and Empirical Arguments Concerning Phonological Acquisition. *Linguistic Inquiry* 29, 656-683.

Hale, M. & C. Reiss (2008), *The Phonological Enterprise*. (Oxford: Oxford University Press)

Hansson, G. (2001), *Theoretical and Typological Issues in Consonant Harmony*. Ph.D. Dissertation, (University of California, Berkeley)

Hayes, B. & C. Wilson (2008), A Maximum Entropy Model of Phonotactics and Phonotactic learning. *Linguistic Inquiry* 39, 379-440.

Hedlund, G. & P. O'Brien (2004), *A Software System for Linguistic Data Capture and Analysis*. B.Sc. Honour's Thesis, Memorial University of Newfoundland.

Ingram, D. (1989), *First Language Acquisition: Method, Description, and Explanation*. (Cambridge, MA: Cambridge University Press)

Inkelas, S. (2003), J's Rhymes: A longitudinal Case Study of Language Play. *Journal of Child Language* 30, 557-581.

Inkelas, S. & Y. Rose (2008), Positional Neutralization: A Case Study from Child Language. *Language* 83, 707-736.

Jakobson, R. (1941/1968), *Kindersprache, Aphasie, und allgemeine Lautgesetze*. (Uppsala: Almqvist & Wiksell) [Translated (1968) by R. Keiler. *Child Language, Aphasia, and Phonological Universals*. (The Hague: Mouton)]

Jakobson, R., G. Fant & M. Halle (1952), *Preliminaries to Speech Analysis*. (Cambridge, MA: MIT Press)

Jongstra, W. (2003), *Variation in Reduction Strategies of Dutch Word-initial Consonant Clusters*. Ph.D. Dissertation, University of Toronto.

Kahn, D. (1976), *Syllable-based Generalizations in English Phonology*. Ph.D. Dissertation, MIT.

Kaye, J., J. Lowenstamm & J.-R. Vergnaud (1990), Constituent Structure and Government Phonology. *Phonology* 7, 193-231.

Kent, R. (1976), Anatomical and Neuromuscular Maturation of the Speech Mechanisms: Evidence from Acoustic Studies. *Journal of Speech and Hearing Research* 19, 421-447.

Kent, R. (1992), The Biology of Phonological Development. In Ferguson, C., L. Menn & C. Stoel-Gammon (eds.), *Phonological Development: Models, Research, Implications*, 65-90. (Timonium, Maryland: York Press)

Kent, R. & G. Miolo (1995), Phonetic Abilities in the First Year of Life. In Fletcher, P. & B. MacWhinney (eds.), *The Handbook of Child Language*, 303-334. (Cambridge, MA: Blackwell)

Kent, R. & A. Murray (1982), Acoustic Features of Infant Vocal Utterances at 3, 6 and 9 Months. *Journal of the Acoustical Society of America* 72, 353-365.

Kunath, S. & S. Weinberger (2009), STAT: Speech Transcription Analysis Tool. *Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics, Companion Volume: Demonstration Session*, 9-12. (Boulder, CO: Association for Computational Linguistics)

Leonard, L. & K. McGregor (1991), Unusual Phonological Patterns and their Underlying Representations: A Case Study. *Journal of Child Language* 18, 261-271.

Levelt, C. (1994), *On the Acquisition of Place* (HIL Dissertations in Linguistics 8). (The Hague: Holland Academic Graphics)

Lintfert, B. (2009), *Phonetic and Phonological Development of Stress in German*. Ph.D. Dissertation, University of Stuttgart.

Macken, M. (1992), Where's Phonology? In Ferguson, C., L. Menn & C. Stoel-Gammon (eds.), *Phonological Development: Models, Research, Implications*, 249-269. Timonium, MD: York Press.

Macken, M. (1995), Phonological Acquisition. In Goldsmith, J. (ed.), *The Handbook of Phonological Theory*, 671-696. (Cambridge, MA: Blackwell)

MacWhinney, B. (2000), *The CHILDES Project: Tools for Analyzing Talk. 3rd Edition.* (Mahwah, NJ: Lawrence Erlbaum Associates)

MacWhinney, B. & Y. Rose (2008), The Phon & PhonBank Initiative within CHILDES. Paper presented at the *International Association for the Study of Child Language*, Edinburgh.

Maddocks, K. (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, T. (2009), *The Articulatory Basis of Positional Asymmetries in Phonological Acquisition*. Ph.D. Dissertation, Massachusetts Institute of Technology.

Ménard, L. (2002), *Production et perception des voyelles au cours de la croissance du conduit vocal: variabilité, invariance et normalisation*. Ph.D. Dissertation, Université Stendhal Grenoble 3.

Munson, B., J. Edwards, S. Schellinger, M. Beckman & M. Meyer (2010), Deconstructing Phonetic Transcription: Language-Specificity, Covert Contrast, Perceptual Bias, and an Extraterrestrial View of Vox Humana. *Clinical Linguistics and Phonetics* 24, 245-260.

Pater, J. (1997), Minimal Violation and Phonological Development. *Language Acquisition* 6, 201-253.

Pater, J. & A. Werle (2003), Direction of Assimilation in Child Consonant Harmony. *Canadian Journal of Linguistics* 48, 385-408.

Piggott, G. (1999), At the Right Edge of Words. *The Linguistic Review* 16, 143-185.

Pinker, S. (1984), *Language Learnability and Language Development*. (Cambridge, MA: Harvard University Press)

Pinker, S. (1989), *Learnability and Cognition*. (Cambridge, MA: MIT Press)

Preston, J., H. Ramsdell, K. Oller, M.L. Edwards & S. Tobin (2011), Developing a Weighted Measure of Speech Sound Accuracy. *Journal of Speech, Language, and Hearing Research* 54, 1-18.

Prieto, P. (2011), Tonal alignment. In Oostendorp, M. van, C. Ewen, E. Hume & K. Rice (eds.), *Companion to Phonology*, 1185-1203. (Malden, MA: Wiley-Blackwell)

Prieto, P., A. Estrella, J. Thorson & M. Vanrell (2012). Is Prosodic Development Correlated with Grammatical and Lexical Development? Evidence from Emerging Intonation in Catalan and Spanish. *Journal of Child Language* 39, 221-257.

Rose, Y. (2000), *Headedness and Prosodic Licensing in the L1 Acquisition of Phonology*. Ph.D. Dissertation, McGill University.

Rose, Y. (2003a), ChildPhon: A Database Solution for the Study of Child Phonology. In Beachley, B., A. Brown & F. Conlin (eds.), *Proceedings of the 27th Annual Boston University Conference on Language Development*, 674-685. (Somerville, MA: Cascadilla Press)

Rose, Y. (2003b), Place Specification and Segmental Distribution in the Acquisition of Word-final Consonant Syllabification. *Canadian Journal of Linguistics* 48, 409-435.

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

Rose, Y., G. Hedlund, R. Byrne, T. Wareham & B. MacWhinney (2007), Phon 1.2: A Computational Basis for Phonological Database Elaboration and Model Testing. In Buttery, P., A. Villavicencio & A. 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: Association for Computational Linguistics)

Rose, Y. & S. Inkelas (2011), The Interpretation of Phonological Patterns in First Language Acquisition. In Ewen, C., E. Hume, M. van Oostendorp & K. Rice (eds.), *The Blackwell Companion to Phonology*, 2414-2438. (Malden, MA: Wiley-Blackwell)

Rose, Y., B. MacWhinney, R. Byrne, G. Hedlund, K. Maddocks, P. O'Brien & T. Wareham (2006), Introducing Phon: A Software Solution for the Study of Phonological Acquisition. In Bamman, D., T. Magnitskaia & C. Zaller (eds.), *Proceedings of the 30th Annual Boston University Conference on Language Development*, 489-500. (Somerville, MA: Cascadilla Press)

Santos, C. dos (2007), *Développement phonologique en français langue maternelle: une étude de cas.* Ph.D. Dissertation, University Lumière Lyon 2.

Scobbie, J., F. Gibbon, W. Hardcastle & P. Fletcher (1996), Covert Contrast as a Stage in the Acquisition of Phonetics and Phonology. In Broe, M. & J. Pierrehumbert (eds.), *Papers in Laboratory Phonology V: Acquisition and the Lexicon*, 43-62. (Cambridge: Cambridge University Press)

Selkirk, E. (1980a), The Role of Prosodic Categories in English Word Stress. *Linguistic Inquiry* 11, 563-605.

Selkirk, E. (1980b), Prosodic Domains in Phonology: Sanskrit Revisited. In M. Aronoff & M.-L. Kean (eds.), *Juncture: A Collection of Original Papers*, 107-129. Saratoga, CA: Anma Libri.

Selkirk, E. (1982), The Syllable. In Hulst, H. van der & N. Smith (eds.), *The Structure of Phonological Representation, Vol. 2,* 337-385. (Dordrecht: Foris)

Smit, A. (1993), Phonologic Error Distribution in the Iowa-Nebraska Articulation Norms Project: Consonant Singletons. *Journal of Speech and Hearing Research* 36, 533-547.

Smith, N. (1973), *The Acquisition of Phonology: A Case Study*. (Cambridge: Cambridge University Press)

Spencer, A. (1986), Towards a Theory of Phonological Development. *Lingua* 68, 3-38.

Stampe, D. (1969), The Acquisition of Phonetic Representation. In Binnick, R. (ed.), *Papers from the 5th Regional Meeting of the Chicago Linguistic Society*, 433-444. (Chicago: Chicago Linguistic Society)

Stelt, J. van der, K. Zajdó & T. Wempe (2005), Exploring the Acoustic Vowel Space in Two-Year-Old Children: Results for Dutch and Hungarian. *Speech Communication* 47, 143-159.

Stemberger, J. & C. Stoel-Gammon (1991), The Underspecification of Coronals: Evidence from Language Acquisition and Performance Errors. In Paradis, C. & J.-F. Prunet (eds.), *The Special Status of Coronals: Internal and External Evidence* (Phonetics and Phonology Series), 181-199. (San Diego: Academic Press)

Tesar, B. & P. Smolensky (2000), *Learnability in Optimality Theory*. (Cambridge, MA: MIT Press)

Tomasello, M. (2003), *Constructing a Language: A Usage-based Theory of Language Acquisition*. (Cambridge: Harvard University Press)

Velleman, S. (1996), Metathesis Highlights Feature-by-position Constraints. In Bernhardt, B., J. Gilbert & D. Ingram (eds.), *Proceedings of the UBC International Conference on Phonological Acquisition*, 173-186. (Somerville: Cascadilla Press)

Vihman, M. & W. Croft (2007), Phonological Development: Toward a “Radical” Templatic Phonology. *Linguistics* 45, 683-725.

Voormann, H. & U. Gut (2008), Agile Corpus Creation. *Corpus Linguistics and Linguistic Theory* 4, 235-251.