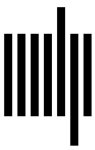
This is a section of doi:10.7551/mitpress/12200.001.0001

The Open Handbook of Linguistic Data Management

Edited by: Andrea L. Berez-Kroeker, Bradley McDonnell, Eve Koller, Lauren B. Collister

Citation:

The Open Handbook of Linguistic Data Management
Edited by: Andrea L. Berez-Kroeker, Bradley McDonnell, Eve Koller, Lauren B. Collister
DOI: 10.7551/mitpress/12200.001.0001
ISBN (electronic): 9780262366076
Publisher: The MIT Press
Published: 2022



The MIT Press

31 Managing Acquisition Data for Developing Large Sesotho, English, and French Corpora for CHILDES

Katherine Demuth

1 Introduction

Research into the acquisition of language began with diary data in the 1800s (Darwin 1877), constituting the early beginnings of corpus construction, which has become central to language acquisition research. Although experimental methods have also had a major impact on the field of developmental psycholinguistics, these often involve cross-sectional studies at different ages, where data from many participants are collated to provide a picture of what the "typical" child of a certain age might perceive or produce. Language acquisition corpora, which often involve the longitudinal study of only a few children, provide a complementary picture of development over time. As such they also provide a rich amount of detail regarding the different steps along the way to becoming a competent user of a language, how this is resolved simultaneously at multiple levels of linguistic structure, and how this developmental trajectory is similar to and/or varies from one individual to the next.

If the child corpora contain spoken input from parents, other caregivers, and/or siblings, these can also be extremely useful for addressing issues about the nature of child-directed speech, as well as the discourse use of the language. This is a critical aspect of the design of a corpus, essential for understanding what types of linguistic information the learner is exposed to, and how this may change as the child becomes a more competent user of the language.

Different types of language acquisition corpora are typically collected for different reasons, focusing on specific research questions. Brown's (1973) longitudinal corpus of Adam, Eve, and Sara's interactions with their parents was specifically designed to address the order of acquisition of grammatical morphemes, forming one of the first comprehensive studies of the acquisition of English grammar. The Manchester Corpus (Theakston et al. 2001), on the other hand, was designed to provide a "dense" corpus of frequent recordings of several children's speech to facilitate research questions of lexical use and how this develops over time. The Demuth Sesotho Corpus (Demuth 1992), described herein, was collected to address issues of children's acquisition of Sesotho morphosyntax, but has also yielded new insights about the structure of the language (e.g., Demuth, Machobane, & Moloi 2009; Demuth et al. 2005; Machobane, Moloi, & Demuth 2007). In contrast, the Providence and Lyon corpora were collected to examine the emergence of grammatical morphemes (Demuth, Culbertson, & Alter 2006; Demuth & Tremblay 2008).

The specific goal of each corpus thus determines the number and ages of the children in the corpus, the frequency and type of recording, and the nature of the interlocutors. Nonetheless, a corpus may be useful for addressing many research questions not originally envisioned during the initial corpus design. In many cases, corpus data can also provide pilot data for designing future experiments and/or future corpora designed to address issues not possible given current resources. This has been the case with all the corpora discussed in this chapter.

Given the different types of corpora collected for different purposes, different results may be found depending on the nature of the corpus and the research question. For example, spoken and written corpora typically yield very different distributions in terms of lexical frequencies. For example, the language used in newspaper text differs from that of poetry or novels in many obvious respects. The same is true of more formal versus less formal speech styles, including narratives, news broadcasts, and everyday speech interactions. None of these different types of language data is capable of capturing all of

what humans know about language. Thus, given a particular research question, new corpora may need to be collected. It is thus advisable to report the type of corpus used in any study, even if it was not specifically designed for the research question at hand. This can then address the important issue of replicability. But perhaps more important, before constructing a corpus, is to collect consent from those in the study to have their data made publicly available. This is especially critical for any data containing video.

The first corpus discussed here is the Demuth Sesotho Corpus (Demuth 1992). This was designed as an exploratory study, investigating the acquisition of morphosyntax in Sesotho, a southern Bantu language. It was collected as part of Demuth's dissertation research. Later funding and computational assistance from Mark Johnson enabled the entire corpus to be computerized, tagged, the audio files linked, and contributed to the Child Language Data Exchange System (CHILDES) database. It is still one of the only corpora on the acquisition of an African language.

The second two corpora (the Providence Corpus [American English] and the Lyon Corpus [French]) were collected at the same time as part of the same project, and they used similar workflow and data management procedures (Demuth, Culbertson, & Alter 2006; Demuth & Tremblay 2008). These "parallel" corpora were collected for the purpose of exploring the emergence of grammatical morphemes in children's early speech in two prosodically different languages (stress-timed English, with trochaic feet vs. syllable-timed French with phrase-final prominence). Many of the data collection methods employed and transcription procedures used were thus designed to be as similar as possible. As it was always envisioned that these data would be contributed to the public domain, parental consent to do so was obtained during enrollment in the study. Further details of these three different corpora are outlined in sections 2–4.

2 The acquisition of morphosyntax: The Demuth Sesotho Corpus

The Demuth Sesotho Corpus (Demuth 1992) was collected in the southern African country of Lesotho from 1980 to 1982. Like other Bantu languages, Sesotho is highly agglutinative, with pervasive agreement on nominal modifiers and the verb: the verb can thus constitute a full sentence on its own, with resulting free word order (cf. Demuth 1992; Doke & Mofokeng 1985 for a description of Sesotho). This raises many interesting questions regarding how the language is acquired. The data for this corpus were collected in a small mountain village of 550 people in the district of Mokhotlong, where it was possible to establish close rapport with both the children and their families. The corpus contains a longitudinal study of four target children's language development as they interacted with members of their extended family including mothers and/or grandmothers, older siblings, cousins, and peers. Three of the target children were aged two to three years, with an older child aged three to four years. This older girl and one of the younger ones were cousins living in the same household and were, therefore, recorded together. Monthly, three- to four-hour recordings of spontaneous speech took place over one year per child, resulting in a corpus of ninety-eight hours of speech containing approximately 13,250 utterances with lexical verbs (approximately 500,000 morphemes).

The original goal of the study was to examine the acquisition of the tense/aspect system. However, it quickly became apparent that these systems are very complex, and that the semantic theories of the day were not up to addressing such issues. Nonetheless, it has been possible to address many issues of morphosyntactic acquisition, leading to early studies of passives (Demuth 1989, 1990) and relative clauses (Demuth 1995), as well as an overall assessment of morphosyntactic abilities in general (Demuth 1992). Because about 40% of the corpus contains utterances from the four target children, and about 40% are adult utterances, with the remaining 20% from peers or older siblings, it has also been possible to have an excellent understanding of the characteristics of the child-directed speech these children hear, providing much-needed new insight into the nature of the target grammar (Demuth et al. 2005; Demuth, Machobane, & Moloi 2009; Machobane, Moloi, & Demuth 2007). Thus, although part of the data management challenge for a lesser studied language is to know what learners may be hearing in the environment around them, investigating this issue is often only made possible by exploring discourse interactions with children.

This corpus was collected and transcribed in the early 1980s, before the advent of portable computers and/or solar batteries for powering electronic devices. Thus, orthographic and broad phonemic transcription was

carried out by hand (by Demuth, in consultation with the mothers/grandmothers of the children), and then verified by a Sesotho native speaker at the University of Lesotho who listened to the tape recordings.

Sesotho is also a tonal language, like all other Bantu languages (except Kiswahili), and like (most) other Niger-Kordofanian languages. A series of follow-up grants were therefore written to explore the acquisition of Sesotho tone, and other syntactic constructions, complementing the use of the Demuth Sesotho Corpus with new experiments. These grants then also provided funding to computerize and morphologically tag the Demuth Sesotho Corpus using the Codes for the Human Analysis of Transcripts and accompanying formatting tools, to prepare it for donation to the CHILDES database (see MacWhinney 2000; https://childes.talkbank.org). It was also then possible to digitize and link the audio files, providing the mechanism for listening to the audio and conducting analysis of the children's acquisition of tone (Demuth 1993). Mark Johnson was instrumental in constructing the Sesotho Morphological Parser to morphologically tag the corpus, 'learning' from an initial hand-tagged set of data to provide options for a human to select the appropriate parse from a list of alternatives. Morphemes that constitute 'words' are connected with a hyphen in the gloss: ke-a-e-rek-a sm1s-t^p-om9-v^buy-m^in 'I'm buying it.' Fusion of two morphemes was indicated with a slash / in the morphological tags: ke-u-rek-ets-e sm1s-om2s-v^buy-ap/t^pf-m^in 'I bought (it) for you.' A full list of tags was then provided in the manual and resulting publications.

A summer of research by Demuth, Johnson, and research assistants was needed to render the Sesotho Corpus computerized and morphologically tagged. This, plus the contribution to the CHILDES database, has put these data in the public domain, where others can now freely use them for exploring issues of broad theoretical interest, including research on child language acquisition, the typology of linguistic systems/learnability, Bantu linguistic structures, and computational issues involving morphological parsing and/or machine translation.

Initial analyses using the Demuth Sesotho Corpus, such as the early work on the acquisition of passives (Demuth 1989, 1990), was carried out by hand. Once the Demuth Sesotho Corpus was tagged and computerized, it was possible to return to these initial analyses and check them for reliability. The correlation was very strong (cf. Demuth & Kline 2006; Kline & Demuth 2010), providing a reliability check for these early analyses. These findings have since been confirmed with a series of perception, production, and generalization experiments using novel verbs (cf. Demuth, Moloi, & Machobane 2010). Thus, although paper transcripts have been extremely useful throughout the ages, the computerization (and morphological tagging) of such corpora not only preserve them for future generations, but also make it possible for both the original researchers, as well as those to come, to explore the data much more systematically and reliably than would otherwise be possible.

3 The emergence of grammatical morphemes in American English: The Providence Corpus

Since at least Brown's (1973) corpus of Adam, Eve, and Sarah, researchers had realized that children acquire grammatical morphemes gradually during the preschool years, and that there are periods in development when children may produce a given morpheme only variably—perhaps 50% of the time, or less. The mechanisms for this gradual acquisition process had been attributed to a lack of either semantic (Brown 1973) or syntactic (Wexler 1994) knowledge. However, the type of longitudinal corpus data needed to evaluate these claims, and explore the potential phonological contributions to this process, did not exist as these previously collected corpora began too late (after the age of two years) and/or had neither the phonetic transcriptions of children's imperfect attempts at grammatical morphemes nor the linked acoustic files needed to conduct a phonetic analysis of what was said. To be able to address the possibility that the emergence of grammatical morphemes might be influenced by phonological or prosodic factors (à la Gerken 1996), it was therefore necessary to collect a new corpus of English, starting earlier (from the age of one year), with audio and video files linked. But this would also be a large undertaking and not possible without several years of funding.

With the help of a large grant and the assistance of talented undergraduates, the Providence Corpus was compiled over a period of six years (Demuth, Culbertson, & Alter 2006). The corpus contains longitudinal audio/video recordings of six monolingual English-speaking children's language development from one to

three years during spontaneous interactions with their parents (usually the mother) at home. The aim of the study was to provide a corpus of phonetically transcribed data, with linked acoustic files, for the purpose of studying early phonological and morphological development. The participants included three girls and three boys, each recorded for approximately one hour a week between the ages of one and three years, beginning at the onset of first words. The full corpus consists of 364 hours of audio/video data, linked to PhonBank (Rose & MacWhinney 2014).

Both adult and child utterances were orthographically transcribed using CHILDES transcription conventions, with the audio/video files linked. Trained transcribers then carried out broad phonemic transcription using the Speech Assessment Methods Phonetic Alphabet, a computer-readable phonetic script transcription of the child utterances. These were then later transferred to Unicode/International Phonetic Alphabet. Stress was only transcribed if it occurred in an unpredictable location (as in the case of child Lily "stressing" articles at age 1;10 [cf. Demuth & McCullough 2009b]). A second trained coder then retranscribed 10% of each file. Reliability scores ranged from 80% to 98% (discounting voicing errors).

These data were then contributed to the CHILDES database. These have been useful for addressing a wide range of issues from phonological/prosodic effects on the acquisition of both inflectional morphemes (e.g., third-person singular) (Song, Sundara, & Demuth 2009) and articles (Demuth & McCullough 2009b). Other studies have explored some of the acoustics of the children's and mothers' speech, with a special focus on the acquisition of coda consonants (Song et al. 2013, Song, Shattuck-Hufnagel, & Demuth 2015). The Providence Corpus has also been useful for teaching language acquisition classes, where students can log on to the CHILDES database and conduct either cross-sectional and/or longitudinal group and/or individual research projects. Some of these have resulted in honors theses/publications (e.g., Evans & Demuth 2012 on pronoun reversal). Many others have used the Providence Corpus for a wide range of studies of child speech, child-mother interactions, and computational modeling of word segmentation (Börschinger, Johnson, & Demuth 2013; Johnson et al. 2014). The Providence Corpus was also one of the first to be included in the new Databrary database (https://www.databrary.org), which is available to researchers of child development. Thus, contribution to the public domain has again made these data widely available to a broad range of researchers for the investigation of issues not originally envisioned.

4 The emergence of grammatical morphemes in French: The Lyon Corpus

Addressing theoretical issues regarding the nature of the acquisition process in one language raises many questions about the generalizability of these processes to other languages. Thus, as it became clear that phonological and prosodic factors might influence how and when grammatical morphemes were acquired, it became necessary to explore these issues crosslinguistically. Because English and French are prosodically very different (English has [trochaic] lexical stress, and French has phrase-final prominence), the collection of a comparable corpus of French acquisition became paramount.

The methods for collecting and annotating the Lyon Corpus were thus as similar as possible to those used for the Providence Corpus. It contains longitudinal audio/video recordings of five monolingual French-speaking children's language development from ages one to three years during spontaneous interactions with their mothers at home. (Two additional children were also recorded; transcription is still in progress). All files are audio/video linked and available in PhonBank (Rose & MacWhinney 2014).

Once again, the aim of the study was to provide a corpus of phonetically transcribed data, with linked acoustic files, for the purpose of studying early phonological and morphological development. The participants included two boys and three girls, each recorded for one hour every two weeks beginning at the onset of first words (around one year) and continuing until the age of three. The corpus currently consists of 185 hours of speech.

Both adult and child utterances were orthographically transcribed using CHILDES transcription conventions, with the audio/video files linked. Trained transcribers then carried out a broad phonemic (Speech Assessment Methods Phonetic Alphabet > Unicode) transcription of the child utterances. Ten percent of each recording was then retranscribed by a second trained coder, with segmental reliability scores ranging from 90% to 98%.

Issues of reliability can now be addressed across different French corpora as well, where article/determiner acquisition has been well documented by several different groups of researchers (e.g., Veneziano & Sinclair

2000; Bassano, Maillochon, & Mottet 2008; Demuth & Tremblay 2008). The corpus has also provided the means for exploring other issues of phonological interest, such as the acquisition of consonant clusters, both in French and crosslinguistically (Demuth & Kehoe 2006; Demuth & McCullough 2009a; Kehoe et al. 2008).

5 Conclusions

In sum, the collection, annotation, and preparation of language acquisition corpora, though extremely labor-intensive, have made, and continue to make, an enormous contribution to science. Access to underlying data sets allows not only for reliability and verification of original results, but also provides a wealth of additional information that can continue to be tapped in years to come to address a wide range of research questions. These data resources are enhanced by the inclusion now not only of the transcriptions, but also audio files so that researchers can hear what was actually said. This, plus the inclusion of video files, allows for a better understanding of the discourse context as well—that is, who was talking to whom, looking where, and so on. (See Holton, Leonard, & Pulsifer, chapter 4, this volume, for further discussion of ethical issues of video sharing.)

As an example, intonation (and information structure more generally) play a critical role in understanding the meaning of what was said, both in English, and in other languages. Without access to the acoustic signal, all this information is missing. As Ochs (1979) once noted, the very act of transcription involves a reduction of information. Thus, the inclusion of audio files, as well as video of those interacting, provides a much more complete picture of the discourse interactions, including potentially helpful or even essential information regarding the nature of language use, including its development. Technology is now available for ensuring that future corpus construction can include both audio and video information, preserving a more complete record of the discourse situation. Both types of information also facilitate accurate transcription, which is still one of the bottlenecks for corpus development.

Acknowledgments

Collection of the Demuth Sesotho Corpus was supported in part by Fulbright-Hays Doctoral Dissertation Grant and Social Science Research Council International Doctoral Grant for Research in Africa dissertation funding. Computerization and tagging of the corpus was supported in part by National Science Foundation grants BNS-08709938 and SBR-9727897. Collection and annotation of the Providence Corpus and the Lyon Corpus were funded by National Institutes of Health grant R01MH60922 (Demuth & Johnson, Jisa). Additional funding for the Lyon Corpus was provided by two grants from Action Concertée Incitative (Terrains, techniques et théories et Internationalisation des sciences humaines et sociales), as well as support from the Délégation générale à la langue française and the Ministère de l'Enseignement supérieur et de la Recherche. None of these corpora would exist without the ongoing participation of the children and their families who have allowed us into their homes for several years: they have provided abundant insight into their language, their culture, and their children's language development. We thank them, as well as the many students and research assistants in Lesotho, the United States, and France, who contributed hours to the transcription and preparation of these corpora to ready them for donation to the CHILDES database.

References

Bassano, Dominique, Isabelle Maillochon, and Sylvain Mottet. 2008. Noun grammaticalization and determiner use in French children's speech: A gradual development with prosodic and lexical influences. *Journal of Child Language* 35 (2): 403–438.

Börschinger, Benjamin, Mark Johnson, and Katherine Demuth. 2013. A joint model of word segmentation and phonological variation for English word-final/t/-deletion. *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics* 1:1508–1516.

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

Darwin, Charles. 1877. A biographical sketch of an infant. *Mind* 2 (7): 285–294.

Demuth, Katherine. 1989. Maturation and the acquisition of the Sesotho passive. *Language* 65 (1): 56–80.

Demuth, Katherine. 1990. Subject, topic and Sesotho passive. *Journal of Child Language* 17 (1): 67–84.

Demuth, Katherine. 1992. Acquisition of Sesotho. In *The Cross-Linguistic Study of Language Acquisition*, ed. Dan Slobin, 557–638. Hillsdale, NJ: Lawrence Erlbaum Associates.

Demuth, Katherine. 1993. Issues in the acquisition of the Sesotho tonal system. *Journal of Child Language* 20 (2): 275–301.

Demuth, Katherine. 1995. Questions, relatives, and minimal projection. *Language Acquisition* 4 (1–2): 49–71.

Demuth, Katherine, Jennifer Culbertson, and Jennifer Alter. 2006. Word-minimality, epenthesis and coda licensing in the early acquisition of English. *Language* 49 (2): 137–173.

Demuth, Katherine, and Margaret Kehoe. 2006. The acquisition of word-final clusters in French. *Catalan Journal of Linguistics* 5 (1): 59–81.

Demuth, Katherine, and Melissa Kline. 2006. The distribution of passives in spoken Sesotho. *Southern African Linguistics Applied Language Studies* 24 (3): 377–388.

Demuth, Katherine, Malillo Machobane, and Francina Moloi. 2009. Learning how to license null noun-class prefixes in Sesotho. *Language* 85 (4): 864–883.

Demuth, Katherine, Malillo Machobane, Francina Moloi, and Christopher Odato. 2005. Learning animacy hierarchy effects in Sesotho double object applicatives. *Language* 81 (2): 421–447.

Demuth, Katherine, and Elizabeth McCullough. 2009a. The acquisition of clusters in French. *Journal of Child Language* 36 (2): 425–448.

Demuth, Katherine, and Elizabeth McCullough. 2009b. The prosodic (re)organization of children's early English articles. *Journal of Child Language* 36 (1): 173–200.

Demuth, Katherine, Francina Moloi, and Malillo Machobane. 2010. 3-Year-olds' comprehension, production, and generalization of Sesotho passives. *Cognition* 115 (2): 238–251.

Demuth, Katherine, and Annie Tremblay. 2008. Prosodically-conditioned variability in children's production of French determiners. *Journal of Child Language* 35 (1): 99–127.

Doke, Clement Martyn, and S. Machabe Mofokeng. 1985. *Textbook of Southern Sotho Grammar*. 2nd ed. Cape Town: Longman.

Evans, Karen E., and Katherine Demuth. 2012. Individual differences in pronoun reversal: Evidence from two longitudinal case studies. *Journal of Child Language* 39 (1): 162–191.

Gerken, LouAnn. 1996. Prosodic structure in young children's language production. *Language* 72 (4): 683–712.

Johnson, Mark, Anne Christophe, Emmanuel Dupoux, and Katherine Demuth. 2014. Modelling function words improves unsupervised word segmentation. *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics*. Vol. 1, *Long Papers*, 282–292.

Kehoe, Margaret, Geraldine Hilaire-Debove, Katherine Demuth, and Conxita Lleó. 2008. The structure of branching onsets and rising diphthongs: Evidence from the acquisition of French and Spanish. *Language Acquisition* 15 (1): 5–57.

Kline, Melissa, and Katherine Demuth. 2010. Factors facilitating implicit learning: The case of the Sesotho passive. *Language Acquisition* 17 (4): 220–234.

Machobane, Malillo, Francina Moloi, and Katherine Demuth. 2007. Some restrictions on Sesotho null noun class prefixes. *South African Journal of African Languages* 27 (4): 166–180.

MacWhinney, Brian. 2000. *The CHILDES Project: Tools for Analyzing Talk*. 3rd ed. Mahwah, NJ: Lawrence Erlbaum Associates.

Ochs, Elinor. 1979. Transcription as theory. *Developmental Pragmatics* 10 (1): 43–72.

Rose, Yvan, and Brian MacWhinney. 2014. The PhonBank Project: Data and software-assisted methods for the study of phonology and phonological development. In *The Oxford Handbook of Corpus Phonology*, ed. Jacques Durand, Ulrike Gut, and Gjert Kristoffersen, 380–401. Oxford: Oxford University Press.

Song, Jae Yung, Katherine Demuth, Karen Evans, and Stefanie Shattuck-Hufnagel. 2013. Durational cues to fricative codas in 2-year-olds' American English: Voicing and morphemic factors. *Journal of the Acoustical Society of America* 133 (5): 2931–2946.

Song, Jae Yung, Stefanie Shattuck-Hufnagel, and Katherine Demuth. 2015. Development of phonetic variants (allophones) in 2-year-olds learning American English: A study of alveolar stop/t, d/codas. *Journal of Phonetics* 52:152–169.

Song, Jae Yung, Megha Sundara, and Katherine Demuth. 2009. Phonological constraints on children's production of English third person singular *-s*. *Journal of Speech, Language, Hearing Research* 52 (3): 623–642.

Theakston, Anna L., Elena V. M. Lieven, Julian M. Pine, and Caroline F. Rowland. 2001. The role of performance limitations in the acquisition of verb-argument structure: An alternative account. *Journal of Child Language* 28 (1): 127–152.

Veneziano, Edy, and Hermine Sinclair. 2000. The changing status of "filler syllables" on the way to grammatical morphemes. *Journal of Child Language* 27 (3): 461–500.

Wexler, Ken. 1994. Optional infinitives, head movement and the economy of derivations in child grammar. In *Verb Movement*, ed. David Lightfoot and Norbert Hornstein, 305–350. Cambridge: Cambridge University Press.

This book was set in Stone Serif and Stone Sans by Westchester Publishing Services.

Library of Congress Cataloging-in-Publication Data

Names: Berez-Kroeker, Andrea L., editor. | McDonnell, Bradley James, editor. | Koller, Eve, editor. | Collister, Lauren B., editor.
Title: The open handbook of linguistic data management / edited by Andrea L. Berez-Kroeker, Bradley McDonnell, Eve Koller and Lauren B. Collister.
Description: Cambridge, Massachusetts : The MIT Press, [2021] | Series: Open handbooks in linguistics series | Includes bibliographical references and index.
Identifiers: LCCN 2020044363 | ISBN 9780262045261 (hardcover)
Subjects: LCSH: Computational linguistics. | Natural language processing (Computer science) | Data mining.
Classification: LCC P98 .O64 2021 | DDC 410.285—dc23
LC record available at https://lccn.loc.gov/2020044363