The Multimedia Adult Learner Corpus

(published in TESOL Quarterly (2003), v. 37, # 3 pp. 546-557)
Posted to this website with the permission of TESOL Quarterly and the authors.

Stephen Reder
Kathryn Harris
Kristen Setzler
Portland State University
Portland, Oregon, United States

This report describes an innovative corpus project that will add several important dimensions to the emerging connections between corpus linguistics and TESOL. A multimedia learner corpus, the Multimedia Adult ESL Learner Corpus (MAELC), is being collected within an adult ESL instructional environment. This "Lab School" environment (see http://www.labschool.pdx.edu) is jointly operated by the Applied Linguistics Department at Portland State University and Portland Community College, an adult ESL provider. Low-level adult ESL classrooms within a regular program are continuously recorded with multiple video cameras and microphones. By the end of the 5-year project period (August 1, 2001- July 31, 2006), the resulting corpus will contain approximately 5000 hours of classroom language and instruction involving approximately 1000 adult learners. With software developed to attach transcriptions and classroom activity codes to the digital media corpus, users can readily used to search for and play back video-audio clips that illustrate particular points of second language acquisition (SLA) or L2 pedagogy. This multimedia corpus and associated software tools will be available online to scholars and practitioners for research and professional development activities.

INNOVATIVE CHARACTERISTICS OF THE CORPUS

MAELC will add considerably to existing L2 learner corpora. Although a range of L2 learner corpora are already available (e.g., Granger, 1998; Granger, Hung & Petch-Tyson, 2002), this corpus adds several aspects to the connections between corpus linguistics and TESOL: (a) It focuses on the early stages of adult SLA; (b) it is highly extensible and searchable in terms of both transcribed language and coded pedagogical activities**;** and (3) with associated software, it maintains persistent links between transcriptions and original audio-video recordings.

Focus on Early Stages of SLA

A multimedia corpus is a particularly appropriate way to capture language in the early stages of acquisition. Low-level learner language has traditionally been very difficult to research, in part because emergent L2 forms and non-verbally conducted communication are difficult to represent in transcripts. MAELC represents learner through both transcripts and the associated video and audio recordings. The corpus includes learners from the very beginning stages throughout their acquisition process, making longitudinal studies possible on large numbers of learners. The recording of individual learners on a regular basis over time will make possible, in-depth research on learner language development.

Corpus Extensibility and Searchability

The digitally recorded multimedia corpus is very large, with numerous cameras and microphones used to record each class (see Appendix A). Such a corpus would be extremely time-consuming for either practitioners or scholars to use unless its contents were indexed in ways that allow ready access for use in research and/or professional development. The project has developed specialized software, called ClassAction (see Appendix B), to attach activity and content codes and transcriptions of classroom language to the multimedia corpus. Our classroom activity coding framework, described below, indexes and helps locate clips from the recorded language classrooms reflecting particular participation patterns, pedagogical activities, and so forth. Searches based on these activity codes make it easy to examine various aspects of learner acquisition processes in the classroom context.

We have also developed a transcription framework appropriate for early stages of SLA. Transcripts include information on what students actually said and what their target utterance was, facilitating the ready identification of certain types of errors in learner language. Users of the corpus can search and analyze the transcribed language data using corpus linguistics software. ClassAction enables other researchers to add their own structured codes, open-ended annotations or transcription details (e.g., a layer of phonetic transcription or a layer of grammatical tags) to the corpus, which will be available to a community of users. The corpus can be searched in terms of combinations of classroom codes and student language, facilitating research directed at relationships between pedagogical activities and student language development.

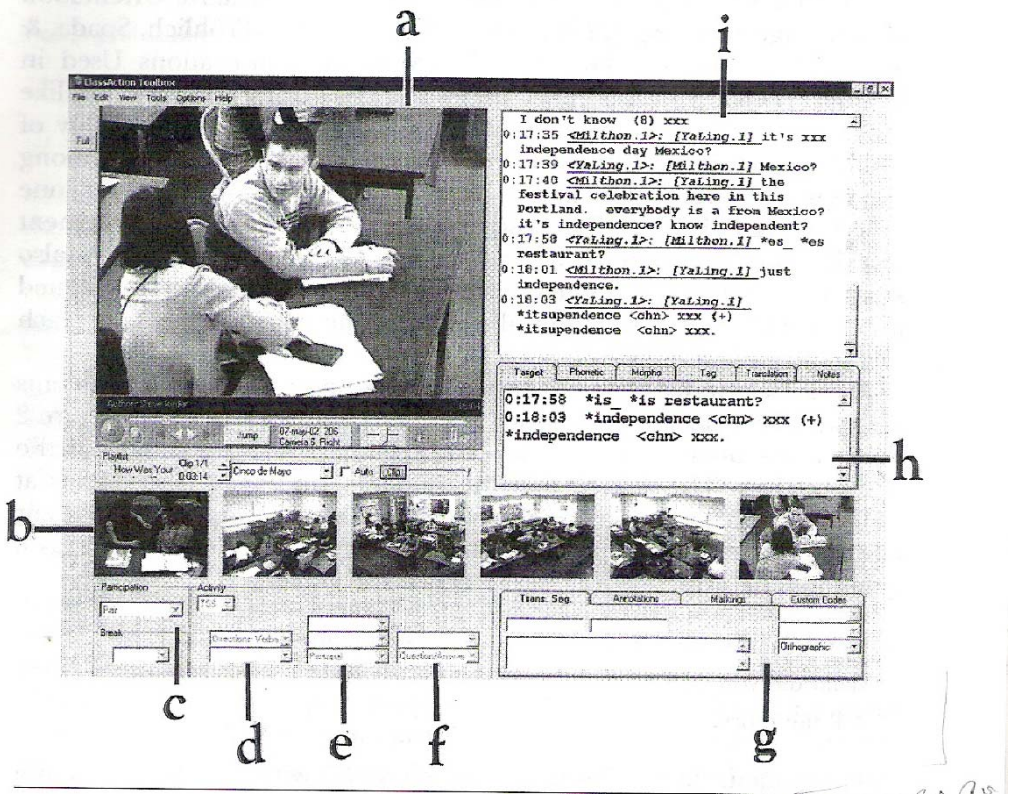
Persistent Links with Recorded Media

Persistent links between the corpus of transcribed language and the original media recordings add a number of important dimensions to corpus-based research and professional development in TESOL. First, when researchers query the corpus, they not only receive the selected linguistic data but can also view and listen to the associated clips from the classrooms. This retrieval will greatly extend researchers' ability to interpret transcriptions and codes in the corpus as well as to add additional transcriptions or activity codes to the corpus. Language teachers and teacher educators can search the corpus and display selected clips to use in preservice and in-service education and professional development activities. Modules that have been developed on working with low level learners and teachable moments (Kurzet, 2000) have included illustrative video clips, related readings and discussion questions.

Figure 1, a screen shot from the Toolbox module of ClassAction, illustrates how the transcription and coding of classroom language and activity are persistently are linked to the recorded media. (See Lab School, 2003a, for the same clip and associated corpus data referred to in this report). Onscreen is the "Cinco de Mayo" clip from a class on May 7, 2002 (0:18:04 into a 3-hour class). Six synchronized camera views are displayed (b), any one of which can be clicked upon to enlarge its size (a) and activate its audio. Classroom activity codes associated with this moment are also shown (c-f). The real-time scrolling transcript appears in the upper right window (i), with a variety of transcript layers displayed in the tabbed window below it (h); here the *target* layer is shown. Users can view, enter, or edit additional data – within the standard MAELC framework or within a custom framework—through the tabbed window (g).

FIGURE 1

Screen shot of a MAELC Clip in ClassAction Toolbox



Note. "cinco de Mayo" clip from a class on May 7, 2002 (0:18:04 into 3-hour class). (a) current camera view; (b) six synchronized camera views (two pair-focused views on the ends separated by four fixed, whole-class views); (c) coded *participation pattern* ("Pair"); (d) coded *prompt* ("Directions: Verbal"); (e) coded *information* ("Personal"); (f) coded *language* ("Question/Answer"); (g) tabbed panel for viewing/editing transcriptions, annotations, markings, and custom codes linked to media; (h) tabbed panel for viewing synchronized nonorthographic layers of transcription (*target, phonetic, morphophonemic, grammatical tag, translation,* and *notes* layers; target layer shown); (i) synchronized orthographic transcription window.
Key to transcript notation: <Name.#>= speaker ID; <chn>=Chinese code switch [Name.#>=addressee ID;_= false start; xxx = unrecoverable speech; (+) = pause of 0.5-1.0 second; ? = rising intonation; (#) = pause of 1 second or more; . = falling intonation; * = emergent lexical item.

CODING SYSTEM

The design of the MAELC coding system reflects the Lab School project's research focus on SLA as seen through student interaction in classrooms. In developing the coding system, we were guided by the need to index as many hours of classroom

recordings as possible while maintaining consistency and reliability among numerous coders. We chose categories and category labels to maximize usability by researchers in SLA and L2 pedagogy as well as by L2 educational practitioners.

The Lab School coding system draws on the Communicative Orientation of Language Teaching (COLT) observation scheme (Fröhlich, Spada & Allen, 1985) and the Foci for Observing Communications Used in Settings (FOCUS) (Fanselow, 1977) for categories. However, unlike those systems, our coding system take advantages of the flexibility of coding recorded as opposed to real-time classes. We divide time along overlapping, parallel dimensions (termed *segment types*). Within any one timeline, time is divided into segments, with the end of one segment marking the beginning of the next. (The COLT scheme, Part A, also segments the time line but allows multiple category labels to be assigned to each segment. ClassAction allows only one category label for each segment within a segment type.)

At the beginning of the clip shown in Figure 1 (18:04), two students are engaged in pair work that starts at 15:01 and ends at 22:28. Figure 2 shows a schematic of the overlapping segments that are coded for the entire activity. The activity that the students are engaged in begins at 11:04 with the teacher giving directions for the students to discuss their weekend. The teacher provides language for the students to use as a guide for their discussion. On the board she writes:

"How was your weekend?"
"What did you do?"
"Tell me more."

The students have a few minutes each to talk with their partner about their weekends. Beyond the initial question, the students use their own language, not needing the support

of the language provided by the teacher. At 22:28, the teacher brings the class back together to wrap up the activity by eliciting details from the student pairs. The activity ends at 32:30. A new activity begins at 32:30 with the teacher asking the students questions on a different topic.

FIGURE 2
Overlapping Segments in the MAELC Coding System

	Time						
Segment Type	11:04	15:01		22:28			32:30
Participation Pattern	Teacher fronted	Pair		Teacher fronted			
Pedagogical Activity	Prompt: Teacher provided directions Information: Student provided personal information Language: Student provided questions and answers						New activity

The segment types or categories chosen for the MAELC coding system describe the organization of the classroom and the instructional activities in which the teachers and students engage. The codes describe what is observable from video recordings rather than inferences about the intentions of the teachers within the instructional process. The goal of the coding system is not to compare the teacher to an established set of expectations but to index a large corpus of recorded classes so that corpus users can readily locate and observe periods during which particular student or teacher behaviors of interest are likely to occur. Twenty-four hours per week of classes have been recorded continuously since September, 2001. At this writing, half of these classes have been indexed using the MAELC coding system. A portion of each coded class has been transcribed.

Participation Pattern

Participation pattern codes, shown in Part (c) of Figure 1, reflect the grouping the grouping of the class. Examples include *teacher fronted, student fronted, individual private* (students are working alone at their desks), *individual public* (students are working alone but in the public space; e.g., they are writing on the board), *pair, group* and *free movement*. Indexing language classrooms in this way allows project researchers to locate and further analyze periods during which, for example, students are working and talking as pairs in the context of varied pedagogical activities.

Activity

The second way in which the time line is divided reflects the pedagogical activity. Activities are the components of daily instruction that are organized and planned by the teacher. In the MAELC coding system, each activity is described in three dimensions: the prompt that starts the activity (as indicated in Figure1, Part [d]), the information used in the activity (as indicated in Figure 1, Part [e]) and the language students use to participate in the activity (as indicated in Figure 1, Part [f]). Each dimension includes information about who (teacher or student) provides it and about what is provided. In Figure 1 these dimensions are coded respectively as a *teacher-provided prompt* that is a *set of directions; student-provided information* that is *personal* and *student-provided language* that is *question/answer*.

Using the MAELC coding system, users of the corpus can locate segments of time during which patterns of language or pedagogical behavior of interest may occur. An example is the location of pair participation pattern segments occurring during

pedagogical activities that utilize students' personal information. Lab School researchers are comparing the language produced in these activities with the language produced in activities using information from other sources, such as textbooks. Current research focuses on analyzing the degree of information transfer that occurs, the negotiation of meaning that happens, and the ways in which the students work to co-construct meaning in these and other activity types (e.g., Garland, 2002).

TRANSCRIPTION SYSTEM

Portions of classroom students' language are also being transcribed. The nature of the data in the corpus requires a broader notion of *transcript* than the field has previously associated with that term. A transcript in our project is not an isolated representation of language; it is a linked, multimedia combination of audio, video, and written representations. This type of corpus creates a powerful, context for discussing our own representations and analyses of students' emerging language as well as how such representations limit and facilitate the various understandings of student language within the TESOL field.

The transcription includes an identification of the speaker and addressee and the modality (written or oral) as well as representations of the language produced. Because the transcription can be searched, corpus users can access high-quality language samples for use in longitudinal language acquisition research. Making this kind of research possible required careful consideration of the transcription system used in the corpus. The development of the MAELC transcription system was guided by the difficulties of representing both low-level student language and the dynamic classroom environment.

Furthermore, the system had to strike a balance between the level of transcription detail and limited project resources. The Lab School project's research focus on student language as seen in student-student interaction guides the types of information included in the transcription system as well as the principled selection of time periods for transcription.

Because the corpus is Web-based, other researchers will be able to use ClassAction to transcribe additional material and add layers of transcription detail to transcribed portions of the corpus (e.g., phonetic or morphophonemic layers; see Figure 1, Part [g]). (For information about obtaining access to the corpus and ClassAction software, see Lab School, 2003b).

Transcription Segments and Bubbles

In order to anchor the transcription notation to its audio-video context, a transcription *segment* specifies a given camera, microphone, starting point and ending point within a particular media file. The base transcription layer is a general, orthographic representation of language, including more detailed features than non-technical transcripts usually include, and fewer features than most conversation analysis approaches include. To avoid oversimplifying the complex nature of language produced by low-level learners in the classroom, we developed notation features specific to low-level discourse and a protocols for transcribing classroom interaction that maximize the utility and extensibility of the initial transcription.

Rather than attempt to represent all language that might be heard in a particular camera view, we used the notion of a language *bubble* to focus our transcription efforts.

A bubble contains language that is produced by a student wearing a wireless microphone, the students' interlocutors, and anyone else audible through that microphone who provides a discourse context for that student. The bubble concept sets up priorities within the multitude of simultaneous conversations that occur in the ESOL classroom, resulting in a usable transcript for the researcher/end user of the corpus. The system also has the benefit of targeting language recorded with high audio quality, that is, language that takes place around the wireless microphones.

Discourse that is transcribed in the student-student language bubbles includes information about the speaker and addressee, code switching behaviors, oral or written modality of the language produced, information on major phrase-level intonation information, intraturn pauses, miscues and repairs, paralinguistic vocalizations (e.g., laughter), and nonlexical information used in lieu of lexical items. Although a transcription of protocol for this type of language could include additional aspects of low-level learner language, we have reasoned that, as a first pass, this level of detail will provide a platform for our current analysis and a basis for future research.

Flagging of Emergent Language

Many of the issues we encounted around how to of represent of emergent language were also confronted by researchers in L1 acquisition. How could we represent lexical items that were imperfectly acquired? How could we represent what the learner intended to express so as to facilitate research in interlanguage development? We often found that attempts to represent emergent forms either were in conflict with the need to economize effort within our large-scale transcription enterprise or failed to meet

reasonable standards of accuracy and reliability. Ultimately, to promote the broad and flexible use of the corpus by scholars, we decided to indicate systematically the occurrence of emergent forms in a way that would facilitate future research in this area even if we do not fully characterized such forms in the base transcription layer; these emergent items are flagged with asterisks.

The asterisks serve as an easy way to locate nontarget items with meanings that may be reasonably inferred. In the interaction shown in Figure 1, a student attempts to understand the explanation of a Cinco de Mayo festival:

Actual transcript: 18:03 Chinh.1 *itsupendence <chn> xxx

Target: *independence <Chinese code switch> xxx

This item is flagged based on the sound substitution of [nd] for [ts] in the lexical item. Extra or missing consonants also warrant a flag, although partial lexical items do not unless a lexical substitution is also occurring. (The example above also shows the marking of code switching, a feature common in the low-level learner language in the corpus.) Another example, also from clip shown in Figure 1, shows the flagging of the substitution of [s] for [ks] and the missing final consonant of the target adjective *(Mexican)*:

Actual transcript: 16:07 Chinh.1 *Mes_ *Mesico holiday?
Target: *Mex_ *Mexican holiday?

This example illustrates that, even though our protocol specifies that a sound substitution has occurred, it does so broadly. A researcher who wanted to add a layer of coding with a more principled phonetic representation could easily locate the relevant items.

We include target transcriptions only in cases where the transcriber is reasonably sure of the representation; in cases of reasonable doubt, no target representation is attempted. In the example above, it is reasonable to assume that the target form in this case is the adjective form *Mexican*. The preceding and following utterances show the negotiation of the adjective form. However, it is not always a clear-cut. Transcribers make no attempt to determine what a speaker intended to say (e.g., whether a student intended to say *Mexican* or *Mexico*), and the decision to flag an item is based on consonant substitution, addition or omission.

We do not represent phonetically students' emergent language that contains sound substitutions (because doing so consistently throughout the corpus was not feasible), nor do we assume that such attempts are target forms, creating a misrepresentation of the student's attempt toward accuracy. We hope that systematically flagging these items for further principled inquiry will help the field to understand better the role played by these forms in SLA.

Vowels, which are somewhat more like points on a continuum than many of the consonant sounds are, entail more difficult transcription decisions. We are still in the process of developing a good way to identify vowels that need to be flagged. The need for a flag is apparent in cases where vowel omissions or substitutions result in a change in lexical meaning. Less salient vowel substitutions have been difficult to pin down. When do vowel approximations that are evident in many accents qualify as interfering

with comprehension, and when are they minor? We have not succeeded at defining, a priori, a consistent description of vowel errors that seem egregious within broader orthographic (not phonetic) representation used in the base layer of the transcription. It is hoped that items already flagged in the corpus can serve as the basis for a bottom-up analysis of native speaker judgments in flagging vowel errors.

THE FUTURE OF MAELC

MAELC provides a new and perhaps unique view of low-level L2 development and the pedagogical context within which it occurs. The recording environment makes it possible to focus on emerging language in student-student interaction within classroom (as opposed to experimental) settings over time. A set of coding and transcription tools along with specialized software permits the project team to analyze the corpus in conducting research in SLA and pedagogy. Because the corpus and software will be accessible to researchers via the World Wide Web, we hope that corpus will grow to include additional codes, transcripts, and other types of annotation provided by a community of researchers.

Several obvious directions for studies based on this corpus are morphology acquisition students that address how L1 backgrounds influence early language acquisition and studies of the development of form/function relationships. Because the corpus is searchable, researchers are able to look at how lexical choice varies according to activity type. Perhaps more than anything, this corpus will provide the data necessary for researchers to explore how emergent learner language differs from higher level learner language and native language acquisition.

ACKNOWLEDGMENTS

This project is supported in part by grant R309B6002 from, the U.S. Department of Education, Educational Research and Development Centers Program, to the National Center for the Study of Adult Learning and Literacy, in which Portland State University is a partner.

THE AUTHORS

Stephen Reder is professor and chair of the Department of Applied Linguistics at Portland State University. His interests include adult literacy and language development and the relationships among literacy, technological change, and language. He is currently doing research in the ESOL Lab School and the Longitudinal Study of Adult Learning projects.

Kathryn Harris is a research associate at the Adult ESOL Lab School and an assistant professor in the Department of Applied Linguistics at Portland State University. She is interested in process of adult beginning second language acquisition, in-and out-of-class influences on that acquisition and how the acquisition process can be seen in classroom student-student interaction.

Kristen Setzler is a research associate and MA TESOL student at Portland State University. Her interests include second language acquisition, adult learning and literacy development, and the issues involved in the transcription of lower-level learner language.

REFERENCE:

Fanselow, J.F. (1977). Beyond RASHOMON – Conceptualizing and describing the teaching act. *TESOL Quarterly, 11*, 17-39.

Fröhlich, M., Spada, N., & Allen, P. (1985). Differences in the communicative orientation of L2 classrooms. *TESOL Quarterly, 19*, 27-57.

Garland, J. (2002) *Co-construction of language and activity in low-level ESL pair activity*. Unpublished master's thesis, Portland State University, Portland, Oregon, USA.

Granger, S. (1998). The computerized learner corpus: a versatile new source of data for SLA research. In S. Granger (Ed.), *Learner English on computer.*(pp.3-18). New York: Addison Wesley Longman.

Granger, S., Hung, J., & Petch-Tyson, S. (Eds) (2002). *Computer learner corpora, second language acquisition and foreign language teaching*. Amsterdam: John Benjamins.

Kurzet, R. (2002). Teachable moments: Videos of adult ESOL classrooms. *Focus on Basics*, *5*(D), 8-11.

Lab School. (2003a). *Cinco de mayo* [Data file]. Portland, OR: Portland State University. Retrieved July 16, 2003, from http://www.labschool.pdx.edu/Viewer/viewer.php?TQ_XML3

Lab School. (2003b). Lab School software --- *ClassAction*. Portland, OR: Portland State University. Retrieved July 16, 2003, from http://www.labschool.pdx.edu/ClassAction/

APPENDIX A

Data Collection for MAELC

In the classroom, four fixed ceiling-mounted cameras focus on the whole class, and two remotely controlled, ceiling-mounted cameras focus on individual students wearing wireless microphones (rotated among the students in a class from day to day). The result is a view of the instruction, a close-up view of the students participating in the instructional activities, and high-quality audio recordings of their language.

Classes are selected for coding and transcription in a stratified random design to provide approximately equal amounts of data for classes taught by each teacher and at each instructional level.

APPENDIX B

ClassAction Software

ClassAction is a set of interrelated software programs:

- The Coder-Transcriber program is used by project staff to code and transcribe the data.
- The Toolbox program is used to view synchronized camera views (with audio) and to view, enter, and edit associated codes, transcripts, and annotations.
- The Query program is used to search the large ClassAction corpus of language and codes to identity clips illustrating selected features of language use and pedagogy. These selections (e.g., all the utterances of a particular speaker over time or Level A conversations between speakers with different L1s) are assembled into play lists (comprising one or more clips) that can be viewed using Toolbox.
- The Viewer program is a freely downloadable browser plug-in that has some but not all of the functionality of Toolbox. Play lists can be edited and published on the ClassAction server so that research articles and textbooks may incorporate multimedia examples of particular points of SLA, use, or pedagogy and make them publicly viewable with the Viewer program. Users can view all recorded classes, including those not coded by Lab School staff.