The Oral Language Archive (OLA)
A Digital Audio Database for Foreign Language Study

Christopher M. Jones and Mathew M. McNally
Carnegie Mellon University

[Originally published in the *CALL Journal*, University of Exeter, U.K., Vol. 9, No. 2-3, 1996, 235-250]

The Oral Language Archive (OLA)
A Digital Audio Database for Foreign Language Study

ABSTRACT

The Oral Language Archive is a project underway at Carnegie Mellon University to create a centralized database of digitized sound accessible from distributed personal computers for the study of foreign languages. In addition to the sounds themselves, OLA is comprised of the tools to manage and distribute those sounds. This combination of information, and the tools to interactively access that information, constitute a unique attempt to provide a universal structure to describe, categorize and listen to all major languages via a single access system. The OLA user will be able to listen to hours of Japanese, French and Russian dialogues of varying complexity from their personal computer using a single application.

RATIONALE

The study of language and culture is a multi-faceted endeavor, ultimately as wide-ranging as the targeted culture itself. Educators have increasingly realized that textbooks, while a necessary guide for classroom learners, can mislead students into a narrow linguistic definition of their goal . That goal has in fact continually expanded over the last two decades, changing from a basic mastery of a linguistic system to a familiarity with the texts and artifacts, customs and history of the often diverse peoples sharing a common language. This change in the goal of language instruction has not always been accompanied by an increase in the cultural and linguistic resources available to students and instructors. In many areas of the United States consistent access to foreign or immigrant language communities is impractical, leaving the supply of appropriate materials to instructors and commercial publishing houses, who have become increasingly sensitive to the needs outlined above. Time and cost structures still conspire to limit the richness desirable in authentic materials available to learners.

The movement toward oral communication as a primary focus of instruction has accentuated this difficulty. The traditional emphasis on literature as the exclusive bearer of culture was much more convenient: literature was easily transportable, with established canons and arbiters of taste; libraries and language and literature programs were comfortable in the assumption that their literary collections represented foreign cultures in a fair and relatively complete manner. That assumption no longer holds. Foreign language classroom emphasis on speech and the societal movement away from literature as the exclusive bearer of culture have both contributed to a diversification of the resources required in contemporary study of language and culture.

Outside the classroom, students' experience of the spoken language is typically through the audio and video cassette, both of which have proven advantages in terms of cost and convenience. The principal disadvantage of these materials is their linear nature: they must be listened to from beginning to end; accessing portions in the middle of any recording is possible, but inconvenient. They also are often limited in their scope - constrained by media limitations such as length. For these reasons these media are not adaptable to the interactive dimension currently targeted by instructors and language materials developers. Interactivity is diversely defined, but essentially implies student control over at least the pace and sequence, and ideally the content of instruction. Supplying this dimension outside of the computer domain has proven problematic, since the essential component of interactivity is the necessity for instant and random access to components of the learning resource in use. The concept of random access is a given in the computer environment and, with its promise of constant and multiple restructurings of a learning experience, is ideologically attractive. It is only in the last two or three years that the computer has begun to help us imagine random access to multimedia events, however, rather than just text. Today computers are increasingly sound and video-capable. Computer authoring systems for languages which offer the interactive dimension also exist, but without readily available sources of quality authentic audio or video. The Oral Language Archive at Carnegie Mellon University is helping to address that gap by

offering a versatile and extensive source of digitized oral language to institutions of higher learning and the tools to interactively access these resources. The OLA will thus supply not only a large and growing source of culturally diverse materials, but also a distributive mechanism which makes their appropriate use possible with a facility heretofore unthinkable.

The Oral Language Archive as a concept distinguishes itself from the many hypermedia development efforts currently under way in that it is network-based from beginning to end. Unlike individual workstation-based courseware, which often requires a copy of the software and the media (often laser disc) to be able to work, the OLA requires only log-on privileges. The information stored in the OLA will be huge in comparison to that of a single multimedia project, and it will be simultaneously accessible by a growing number of concurrent users.

PROJECT CONTENT

1. Language input and organization

a. What language?

The Oral Language Archive is most immediately concerned with the collection of authentic speech in the languages targeted for the purpose of supporting basic language instruction. This has been achieved initially through the hiring of native-speakers resident in the Pittsburgh area, normally in pairs, who are given functional or situational priming and recorded.

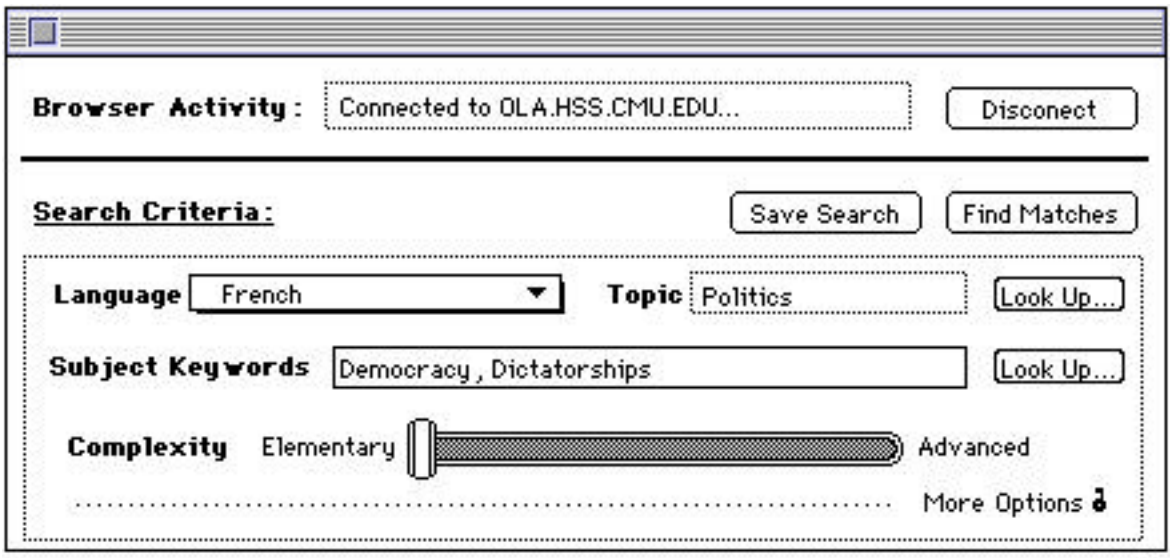
We are principally guided in our collection by a topical taxonomy originally developed by J.A. Van Ek under the aegis of the European Community which detailed the topics, functions and concepts essential to successful negotiation of the demands of a foreign culture.[1] The recording is accomplished by proposing role-playing situations to two or more native speakers with a common cultural background. Thus to begin collecting for the topical category of travel we might give two French speakers from Quebec a map of Montréal. One would ask the other for directions on the Métro to the Stade Olympique, on foot to Le Musée d'Art Moderne, by car to L'Insectarium. In

other situations, we ask our subjects to role-play varying degrees of formality and relative status. All speakers recorded have signed a release authorizing the Oral Language Archive unrestricted educational use of the recordings.

We also intend to travel to record target-language speakers in their countries of origin. The project team is cognizant of the fact that it will be impossible to satisfy the speech source requirements of our database design with on-campus or local recruits. If we intend to have significant variation of age, gender, class and social status, and national or regional origins in our speakers, we must supplement our on-campus collection with speech collected abroad. This will be critical to making the final Archive rich enough to justify the flexibility of the database and the multiple demands we anticipate from our users. This diversity will be of major importance to OLA users interested in the francophone world, for example, or for later users from Hispanic studies.

b. Structure of the language archives

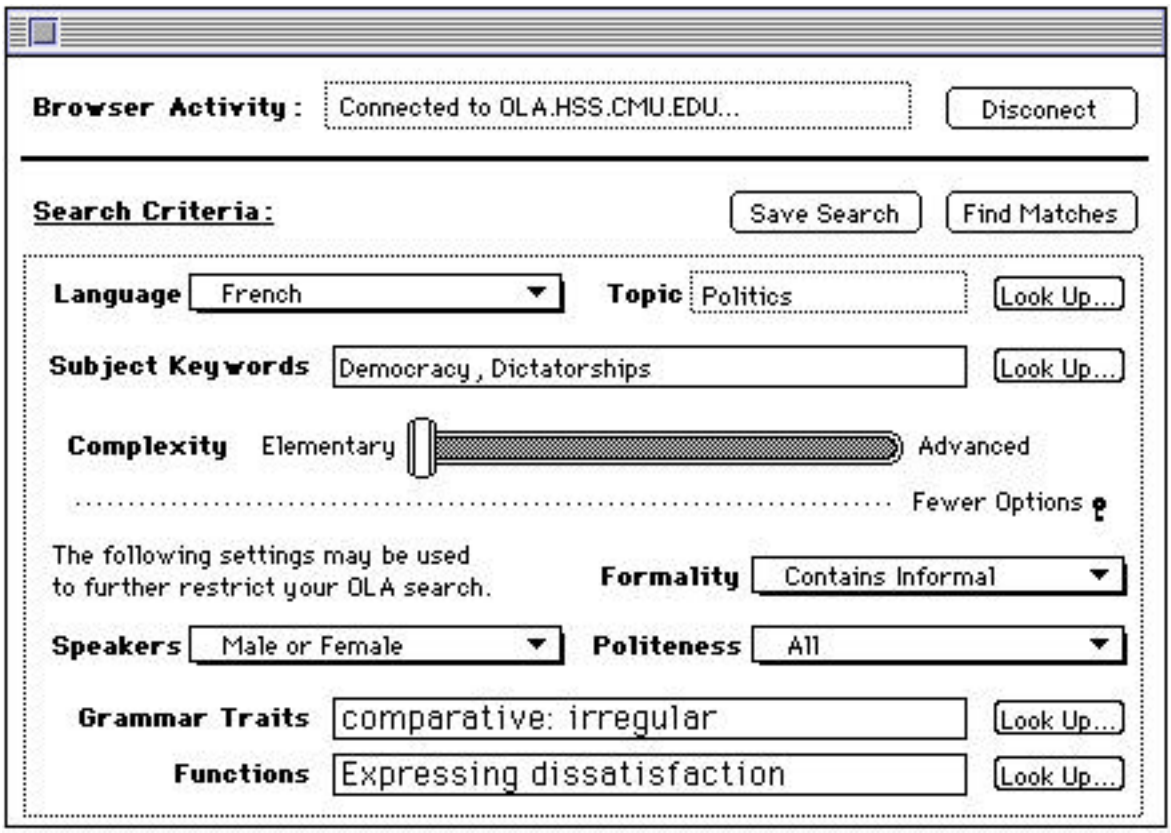
Our intention is to allow users of the OLA to search for language using a wide variety of criteria, some of which are slanted toward instructional needs, others toward potential research interests. Instructors and researchers will have the OLA Browser as their primary access to the Archive. This will be a software interface, requiring no technical competence to manipulate, allowing menu-driven Query By Example searches. The searches can be limited by a variety of factors, much like on-line catalogue searches are currently done in university libraries. The diagram below taken from the OLA Browser contains an example OLA search.



Primary Search Constuctor Window

In that the user will be searching a database of speech and not text, however, other variables are necessary. Among the selectable variables will be (a possible search delimiter is indicated for each variable) : Language (French), Gender of speaker /s (Female), Grammar trait (Indirect object pronouns), Functions (Asking for help), Topic (Travel), Formality (Formal), Subject keywords (Lodging services), Lexical difficulty (Intermediate). In a subsequent iteration other criteria will be added, including Regional Origin of speaker/s (Québec), Age of speaker/s (Adult), Date of

Recording (1994) and Setting (Street). Most current searches do not productively invoke such a large number of variables, but as the OLA grows in a given language, this type of search will be correspondingly encouraged and rewarded. A more complex variation of the search displayed in the diagram below follows:



"More Options" Search Constructor Window

The search results are a list of the dialogues corresponding to the above search criteria. The user can then listen to the dialogue(s) and have access to information such as length, format, collector and so forth of any given segment or dialogue. The instructor is also able to save the retrieved digital audio to disc or establish a link from OLA to another software environment (a Dasher exercise or a HyperCard stack, for example) by retaining the catalogue number of the dialogue. This allows the re-purposing of the audio dialogue or segment into a context imagined by the instructor/user.

Students will most often use OLA resources in the third-party environments mentioned above, for which the OLA project will provide seamless integration.

c. Linguistic difficulty

A critical factor in the selection of the language used in instruction is relative difficulty. Material that is topically appropriate for the goals of the instructional activity must also be appropriate to the learners' level of proficiency. Thus the instructor must have the means to search and retrieve dialogues based on relative difficulty of the language. In the OLA, difficulty is a function of lexical characteristics of the individual vocabulary items in the dialogue. Word frequency reflects how often a word is used in everyday communication and thus is an important predictor of difficulty: more frequently encountered items are learned sooner and used more often. The OLA project team has selected internationally recognized frequency lists--originally established as a guide to instructors-- as the basis of determining our lexical complexity index. These word lists will serve as the basis for a computer text analysis of the transcripts which will in turn produce a difficulty rating usable as a search delimiter. Three levels will be defined, Basic, Intermediate, and Advanced, with higher level lists subsuming the lists below. The contents of the lists vary across languages. For the initial development of Japanese archives, for example, the three levels will be defined by vocabulary lists published by the National Language Research Institute of Toyko in 1984.[2] The size of the lists for the Japanese archives are as follows: Basic Level, 2000 words; Intermediate, 4000 words. Hence each segment of the OLA will have two descriptors that define the percentage of words in the segment that fall into the Elementary and Intermediate, plus a third--Advanced--for words that appear in neither list . The lists in use for French are *Le français fondamental (*1st and 2nd degrees) developed at the *Institut National Pédagogique* in Paris.[3] For German we are using Alan J. Pfeffer's *Grundstufe* and *Mittelstufe*, developed at the University of Pittsburgh.[4]

d. Speech transcription and coding

OLA dialogues will always be transcribed in the recorded language, first, and contain narrative titles in that language for reference. In addition, both transcription and titles are translated into English. An OLA database file thus looks like this:

DialogueID: 33
DialogueContainer: FR003.DIR
DefaultTitle: "Le Musée Picasso"
EnglishTitle: "The Picasso Museum"

SegmentID: 381
SegmentFile:
"Christophe invite Laurence à aller au musée Picasso."
"Christophe invites Laurence to go to the Picasso Museum."
"Eh, qu'est-ce que tu dirais ce week-end d'aller au musée Picasso?"
"Hey, what would you say to going to the Picasso Museum this week-end?"

SegmentID: 382
SegmentFile: FR003002.AII
"Laurence accepte l'invitation et dit ne jamais y avoir été."
"Laurence accepts the invitation and says she's never been there."
"Ah, c'est une bonne idée! Je...n'y suis jamais allée."
"Ah, what a good idea! I've never been there."

etc

When the dialogue is presented by the Browser however, titles and transcription are presented separately, and the dialogue can be skimmed for content by looking at the segment titles, as well as played segment-by-segment.

In each case--transcription and titling--multiple containers are available, so that the potential exists for offering titles and translation in German, Japanese and Korean for the French dialogue above, making the potential internationalization of these archives a reality. These alternative titles will be selectable from pop-up menus in the Browser that allow the user to select the language of the title or transcription areas of the Browser. Likewise, more research-oriented transcriptions, like the International Phonetic Alphabet or CHILDES conventions,[5] can be entered. This flexibility is in line with the non-predictive orientation of the OLA. At some point a researcher may decide to do a phonetic transcription and analysis of modern spoken French in Quebec, based on OLA

recordings. These transcriptions could then be entered into the database and be offered for the use of other language researchers.

After transcription, the dialogue is analyzed for lexical complexity (see above) and coded for topical and syntactic content. For purposes of subject keyword searches, we have supplemented the Van Ek taxonomy with an adaptation of a more exhaustive listing prepared from transcripts of television programming by Sharon Black at the Annenberg School of the University of Pennsylvania.[6] We have developed our own lists of syntactic keywords in common instructional use in each language, though, as with transcriptions, the possibility of later entry of linguistics research-oriented syntactic coding exists.

e. Recording and sound

Though recording direct-to-disk is currently possible, it is not convenient to rely on it in diverse recording environments. Thus most recording is done with high-quality portable tape decks and then digitized at a 16 bit, 44K sampling rate onto Macintoshes and edited using the sound editing software SoundEdit 16 from MacroMedia. The segmented working sound format will make use of Apple's QuickTime and the IMA 4:1 audio compression algorithm that is cross-platform compatible (PC and Macintosh). In that QuickTime supports both audio and video, this also insures that the extension of the OLA functionality to video recordings will be relatively seamless.

2. Instructional implications

a. Existing and potential applications

Recorded speech is currently in use in a multitude of computer-based contexts, including for record/playback/compare pronunciation checking, for dictation, for translation and interpretation practice, for aural comprehension checks, and for cloze and fill-in-the-blank exercises. That this wealth of vehicles exists is indeed one of the primary justifications for the Oral Language Archive.

We have agreements in principle with the directors of the four major authoring systems currently in use in higher education in the United States to insure their integration with the OLA project. Contacts are also being sought with international authoring system developers. This integration will essentially be accomplished by making the OLA dialogue a new media type for authors using these systems. Author/instructors using Dasher, Libra, SuperMacLang or WinCalis will enter links to dialogues from the OLA archive as they create their exercises. Student users will then invoke these links (i.e., listen to OLA dialogues) in these contexts without knowing that the sound files are coming from a distant network location. This will greatly enhance the quantity and variety of recorded language available for use by instructors and students in computer-based contexts.

Additionally, Carnegie Mellon project participants with experience in computer-assisted language learning will further increase the stock of available software templates in such commonly used authoring environments as HyperCard, MacroMedia Director, Toolbook and Visual Basic. These will be distributed with OLA subscription agreements to aid in the successful exploitation of OLA resources.

The browser, a primary interface to the OLA, allows the downloading of audio samples into any of the existing vehicles named above (or vehicles to be created). This will be often a less desirable option, however, given the size of the sound samples concerned. Our primary model will be to export only the “pathway” to the desired sound, without copying it from its OLA server location. Thus the OLA will handle recording, organization and storage problems associated with authentic digitized speech, while leaving individual instructors to create learning contexts unfettered by these concerns. These saved pathways will automatically find the OLA resources required for an exercise when opened by a student user.

The latter capacity will open up the use of the OLA in the classroom. An instructor will be able search the Archive, saving an unlimited number of dialogues as titles with “pathway” descriptions

supplied by the browser (the pathway descriptions will be invisible to both instructor and student; they will be a highly complex, but background function). By inserting the diskette with these pathway descriptions into any classroom computer with an Internet connection, the instructor will be able to directly access pre-selected OLA sound resources for in-class modeling, comprehension and other activities. That is, full access to the entirety of the OLA will be available from any networked location.

b. Non-predictive nature of the OLA

As is perhaps evident from the descriptions above, the uses of the Oral Language Archive will by no means be exhausted by the suggestions of the project team. Users of the Archive at other institutions will quickly and easily create additional uses and exercise templates employing the Archive resources. It is indeed one of the primary goals of the project team to furnish a virtually open development environment, with eventual applications which we can only begin to imagine.

3. Computational aspects

From the user's perspective, the OLA is composed of three logical components:

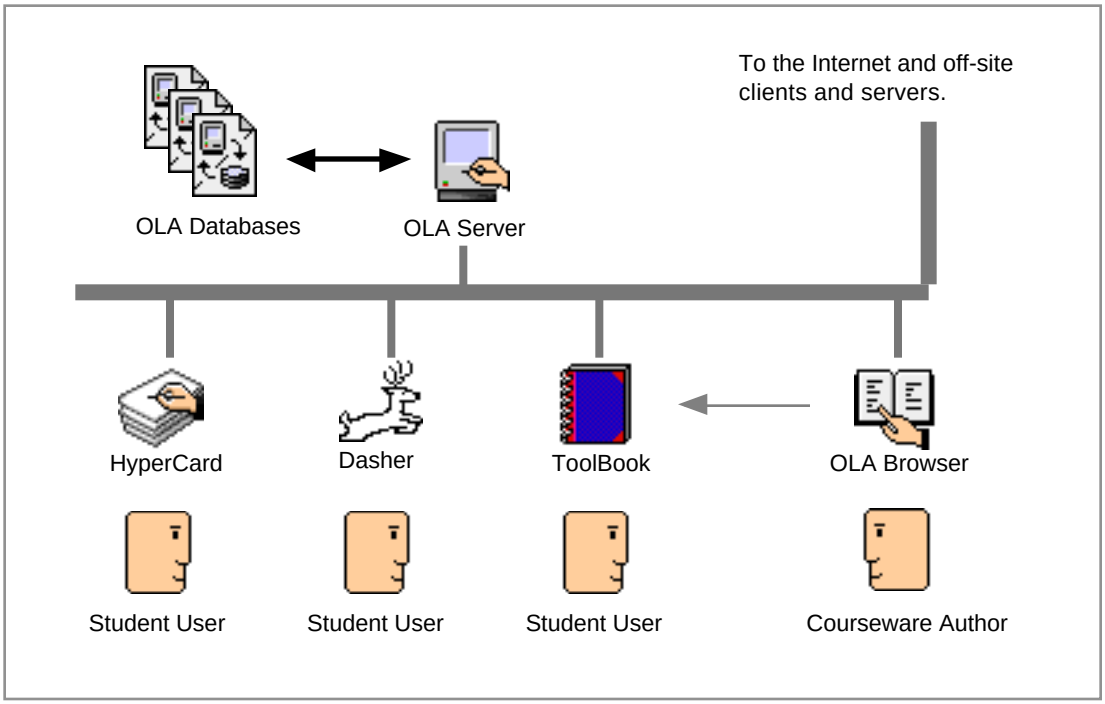
- OLA Servers which maintain the centralized storage of all sound and video segments, provide the computing power for searches through the archive and are capable of working with multiple clients (users who are requesting information) at the same time.
- OLA Client Modules that are added to existing authoring environments on both PC compatibles and Macintoshes to extend the functionality of those environments by enabling them to communicate with OLA Servers and providing retrieval and playback of OLA dialogues.
- An OLA Browser that is used by faculty or researchers on either personal computers to quickly browse through the OLA Servers for appropriate sound or video segments. The browsers will allow a user to select a segment from a server and place either the actual segment or "pathway" to that segment in the authoring tool of their choice.

These three high-level components work together to provide the user with interactive access to the OLA. OLA Servers run on centralized machines that are maintained by the administrators who create, edit and transcribe the audio and video segments. The OLA Clients which are found on the faculty or researchers workstation will transparently communicate with the servers to search for sounds and subsequently play them on demand.

This logical model allows faculty to develop exercises in tools such as HyperCard or Toolbook. After using an OLA Browser to navigate through the OLA to identify segments of interest, faculty build requests to access those segments from within the authoring tool of their choice. Faculty can then distribute these exercises to their students as courseware that provides their students with access to the same OLA segments from any machine networked to an OLA Server just as if each student had a copy of all the appropriate dialogue segments on their own computer.

It is important to note that the underlying technology that presents the logical user model described above is invisible to the end user. From the user's perspective, they simply sit down in front of a workstation and seem to have the entire OLA on their personal machine. Only the OLA client environment’s application code will know that the sound and video segments actually reside somewhere across the network and not on the local machine - even the authoring tools used to access the OLA will be unaware that the segments are not stored locally. Hence, as the OLA Client Modules are simply extensions to the authoring tool of the users’ choice, the users’ conceptual model will be that the tool they already know (such as HyperCard) has simply been extended to include the OLA as part of its underlying functionality. An example for HyperCard scriptors would be the addition of a new command primitive called "OLASoundPlay dialogueID" that would play the appropriate dialogue segment after being attached to a high-level button.

OLA High Level Architecture



Student Users - Student Users running courseware developed in standard authoring environments have transparent access to dialogues from the OLA.

Courseware Author - Using an OLA Browser, courseware authors search the OLA for dialogues to access from their materials via OLA aware authoring tools.

OLA Server - Maintains the OLA databases, manages the requests from student users to play dialogues, and responds to courseware authors or researchers who are using an OLA Browser to navigate through the OLA.

OLA Browser - Used by faculty developers and researchers alike, the Browser provides search capabilities that allow the user to locate apposite dialogues within the OLA and to "export" links to those dialogues to other authoring environments.

OLA Databases - Relational databases provide universal access to the entire OLA - regardless of language type or format. Both the dialogues themselves and descriptive, searchable attributes for those dialogues are maintained by the server.

During the design and implementation of OLA, several underlying goals focused the project's implementation and final deliverables. These aspects represent the present-day technological

stumbling blocks that have been addressed to ensure the creation of both a user friendly and, equally as important in the long term, user efficient language teaching resource.

The system we have designed is targeted to meet the needs of users whose first priority is language, not computers. Great care has been taken to ensure that all of the technical implications that grow out of these goals are hidden from the final OLA user, who will only perceive that they have access to an incredibly vast resource.

a. Domain-specific design

In developing any system, specific domain oriented issues must be addressed. With OLA, the design has focused on the fast and reliable delivery of time-based media. That is, special consideration has been taken to ensure that both video and sound can be delivered (served to the user) in a timely fashion to maintain a high level of interactivity. The current design goal is to ensure that for any request the time-based media starts "playing" well within 1 second with a worst case delay of 2 seconds, and can continue unbroken until completion. Clearly video will require extensive pre-caching on existing network technology, but we are confident that such goals are attainable for audio-only segments.

b. Universality of structure

The system we are currently implementing is general enough to describe and define dialog interactions from all major language types (e.g.. Chinese and French), yet detailed enough to ensure language specific nuances are not lost in an oversimplified abstraction. Each element, whether an audio segment or video clip, of the OLA environment is described in terms of attributes (e.g.. language, gender of speaker, etc. etc.). Work has been completed to define the set of attributes that fully describe each element and its subsequent relationships to every other element. This ‘universal’ set of attributes has been defined across the various language types and facilitates a single interface to browse across languages in the OLA.

c. Global access

Another key to the system is its underlying design to provide global access to the centralized datastore (see below as separate point) from any TCP/IP network client (e.g.. Macintosh, PC compatibles, UNIX Workstations, and so on). This requirement not only dictates that the design must be based on an open technology, but that extreme care must be taken to ensure that the network is used to its maximum potential. Creating the right Client/Server balance via the extensive use of compression and intelligent caching of sounds are being investigated to facilitate Global Access.

d. Ubiquitous clients

In order to facilitate the largest number of possible OLA users, regardless of target platform, the system has been designed with the smallest client system interface possible. The initial target is to create independent code resources (i.e.. XCMD'esque code resources for the Macintosh and DLLs for PC's running windows) that will provide complete access to the centralized server from as many applications and authoring environments on the client machine as possible (e.g. HyperCard, MacroMedia Director, 4th Dimension on the Macintosh and Toolbook, Visual Basic on the PC). These code resources, called OLA Client Modules, provide a simple means of access to the OLA for authoring environments, relieve them of the responsibility of managing the underlying files that make up the dialogues, and insulate those environments from any future changes to the OLA architecture.

e. Centralized logical datastore

Even with the approaching realm of ubiquitous CD-ROM drives, conservative estimates quickly demonstrate that any useful collection of sound and video archives for use in language teaching and research will quickly outstrip the 600 megabyte limitations of the CD media. The question then becomes, how does one provide access to an archive that will be orders of magnitude greater than

the capacity of a CD-ROM disc? Attempting to distribute the gigabytes of space, not to mention the updating of that information, to each individual user would be next to impossible. Only by centralizing the storage of information, can the overall system grow beyond current end-user workstation constraints and remain maintainable in the process. This “Centralize the Data, but Distribute the Work” model is the basis for distributed computing and the only logical direction to pursue given the current computing usage trends and constraints.

By taking advantage of the fact that the expected usage of the OLA at any given time will be based on relatively small (compared to the size of the entire archive) pieces of information - such as an individual dialogue - a networked computer (OLA Client) attached to an OLA Server is able to dynamically retrieve sound on demand. This allows any OLA Client to dynamically access the entire OLA with minimal personal resources - especially local storage resources.

By centralizing the datastore on a set of servers and distributing audio or video segments to client machines only as needed, the design meets the needs of both personalized access of information and the OLA goal of an extensive repository of information from which to work.

f. Scaleability

Yet another key to the system is its ability to handle thousands, if not hundreds of thousands, of entries (both text, sound and video). The system has been designed from the ground up to "scale" to such vast amounts of information. Use of relational database techniques, and design with an eye towards distributed datastores has been the norm (i.e.. all the information may not always be in the exact same place, although it would be in the same logical “datastore”). This ensures that the OLA can grow rapidly without requiring any changes to the client structure or underlying access paradigms.

g. Specialized browsers

In addition to creating a scaleable language archive, and the clients to allow access to the system via the most popular authoring tools, the OLA browser will help users wade through the vast amounts of data as they build their teaching materials. This 'browser' supports searching by any field in the database, the creation of filters and automatic extraction of selected information for insertion into the materials being created. The current database engine even provides an industry standard SQL interface as a method of access of the OLA datastore.

Future developments

1. Enrichment of the current archives. We will augment the current archives of French, German and Japanese to a threshold of usability. This will include additional recording sessions both in the U.S. and abroad as well as the ongoing transcribing, digitizing, coding and database entry of the recorded speech. Our quality and content criteria are expected to evolve with input from test sites and colleagues outside the project team.

2. Dissemination of information about OLA resources and access through multiple channels, including: 1) the establishment of a World-Wide Web site with project descriptions, updates, subscription information, access protocols and demonstration audio and software; 2) international conference presentation; 3) regional workshops for language instructors introducing the OLA as a tool for foreign language instruction and research.

3. Establishment of structures for continuing development and support of the OLA project. This will include an inexpensive system of site-license subscription to OLA with multiple levels of potential use of the OLA resources including: a) standard remote use of the archives, including their integration with other software environments; b) download and use of OLA resources locally, either through local area networks or on individual machines; c) integration of OLA software elements and/or audio resources into software developed at other institutions for academic or commercial

distribution; d) licensing of the OLA server and database-management software for the establishment of OLA database sites at other universities.

Though there will be continuing centralized development of the OLA project, we anticipate that academic users will ultimately base research projects and the development of pedagogical materials on OLA resources, and externally-funded events will thus indirectly influence the evolution of the archive. We intend to facilitate this process by making public as much of the process and documentation as is feasible in the hope of fomenting participation by a wide variety of language professionals. To guarantee the continuing validity of the OLA archive contents and database structure, international Steering Committees will be established for each language. We anticipate the lifespan of the project to be several decades, with the client/server software base evolving to accompany bandwidth and processor evolution, thereby supplying increasingly efficient access to the sound/video archive and its associated data.

Notes:

[1]Van Ek, J.A. The Threshold level for Modern Language Learning in Schools. Groningen:wolters-Noordoff-Longman, 1976.

[2]A Study of the Fundamental Vocabulary for Japanese Language Teaching. Tokyo: The National Language Research Institute, 1984.

[3]Le français fondamental. Ministère de l'Education Nationale; Direction de la Coopération avec la Communauté et l'Etranger. Paris: Institut Pédagogique National, 1959.

[4]Pfeffer, J. Alan. Basic (spoken) German word list: Grundstufe. Englewood Cliffs, N.J: Prentice-Hall, 1964. Basic (spoken) German word list: Mittelstufe. Pittsburgh: University of Pittsburgh Press, 1970.

[5]The Child Language Data Exchange System (CHILDES), is an international database of child language, language disorders, and second language acquisition data based at Carnegie Mellon. For transcription conventions, see MacWhinney, B. The CHILDES Project: Tools for Analyzing Talk. Hillsdale, NJ: Erlbaum, 1991.

[6]Black, Sharon. Thesaurus of Subject Headings for Television: A Vocabulary for Indexing Script Collections. Phoenix: Oryx Press, 1990.

Special thanks to Michael Harrington, G.Richard Tucker, Etsuko Takahashi, and Anne Green for their input during the OLA design process.