E/VALUATING NEW MEDIA IN LANGUAGE DEVELOPMENT

Kenneth Reeder, Trude Heift, Jörg Roche, Shahbaz Tabyanian,
Stephan Schlickau, and Peter Gölz

1. Introduction: Two questions we are sometimes reluctant to ask

Language instructors are bombarded in professional conversations, conferences and publications with glowing reports and demonstrations of "leading edge", "new generation", "must-have" second or foreign language software packages. Indeed, with the advent of such interesting and attractive software as *A la rencontre de Philippe* or *Dans un quartier de Paris* for French, *Berliner Sehen* or *Pilot* for German, or *Ucuchi* for Quechua, and their ilk, it comes as little surprise that a great deal of discussion, often of a highly technical sort, surrounds these new tools for teaching and learning. A question that many of us in the profession are sometimes reluctant to ask about newer software packages is whether in fact the software has convincingly been shown to fulfill its educational purposes. What do we know about the educational effectiveness of the current generation of multimedia language learning software? And, underlying that question, how best do we go about finding out?

The present paper makes a modest claim. The authors - all working as language instructors/researchers in school and university settings and sometime designer-developers of language software - argue that a new approach is needed to the educational evaluation of language learning software that falls under the rubric "new media" or "multimedia" as distinct from previous generations of CALL software. The paper discusses the case for such a new approach by arguing that present approaches to the evaluation of CALL software, while reasonably adequate (although not wholly, we note) for earlier generations of CALL programs, are not appropriate for what we show to be a new genre of CALL software distinguished by its shared assumptions about language learning and teaching as well as by its technical design. We conclude by sketching a research-based program of what we term "E/Valuation" that aims to assist language educators to answer questions about the educational effectiveness of recent multimedia language learning software. We suggest that this needs to take into account not only the nature of the new media and its potential to promote language learning in novel ways, but also current professional knowledge about language learning and teaching.

2. New Media for Language Development

What do the "New Media" look like when applied to the design and development of language learning software? Our team has identified three different types of second or foreign language teaching and learning software as the basis for its investigation: microcosm simulations, microethnographies, and online programs.

The common pedagogical traits of these programs include their proximity to authentic or simulated linguistic and cultural settings, and as a consequence the programs often promote a high degree of interaction between program and user, promote learning of cultural content and intercultural communication skills using realistic experiences and artifacts, and often share features of immersion language education and content based instructional approaches. In addition, many of them are based upon constructivist assumptions about learning, and consequently the programs often promote autonomous learning. What they tend *not* to share are characteristics of grammar-translation, audiolingual or behaviouristic assumptions about learning and instructional design, such as programmed learning, drill and practice sequences or for that matter, much explicit scope and sequence to syllabus design.

Their common technical trait is that each of the three types of programs makes extensive use of the multimedia capacities of computers including complex graphic elements, such as streaming video or animation, and fairly sophisticated sound elements, delivered either from stand-alone media like CD-ROMs or from the World Wide Web, the multimedia manifestation of the Internet, Hypertext, or collaborative learning environments. Some make an interesting effort at incorporating artificial intelligence in the form of natural language processing in addition to earlier developments in help systems that are context-sensitive or that update themselves according to learners' progress. New media for language development, or what we will refer to as multimedia language learning (MMLL) software, represents the most recent generation of computer-assisted language learning (CALL) software.

Taken together, these shared characteristics suggest that MMLL software constitutes a recognizable genre of language learning software whose assumptions about language learning, and approaches to instructional design (a) are not adequately taken into account by present approaches to CALL software description and evaluation, and (b) should be taken into account in any comprehensive approach to the description and evaluation of CALL software. We review recent examples of these three categories in Section 5 below, when we turn to our modest proposal as to how to go about evaluating the new generation of language software in a more satisfactory fashion.

3. Current Practices in Language Software Evaluation

3.1 General issues for software evaluation research and practice

A critique of current approaches to software evaluation requires some description of current practices in the field. Generally, there are two main approaches to software assessment: introspective (checklists, reviews) and empirical evaluations. While checklists constitute a more-or-less systematic and structured evaluation involving the use of a printed form, the typical review includes basic information about the program and the reviewer's subjective description. However, the types of criteria found in a checklist and a review largely overlap (see Knowles, 1992; Schmueckler & Shuell, 1989; Hubbard, 1992). Often the criteria employed in a review will vary according to the reviewer rather than derive from a theory of evaluation.

In contrast, empirical evaluations require that the materials have to be used for some time by

actual learners in a learning situation and thus the approach moves away from the introspective approach (see Scholfield, 2000 for a detailed discussion). Our survey of recent evaluation projects of an empirical type, together with some reflection upon our own recent practices and experiences with the assessment of our own projects, suggests a lack of methodological rigor or at very least a lack of agreed-upon methodological protocols that can create what we term "idiosyncratic assessment."

An introspective assessment by learners, for example, can prove to be unreliable, producing contradictory results at times. Child's (1997, 1998) studies, for example, indicate that even apparently highly interdependent factors vary to a considerable degree. In a multimedia-based language course taught by the same instructor during 1990 and 1995, students were asked to evaluate factors like "% 'strongly agreeing' course materials are useful" (1991: 71%; 1995: 91%) and "% rating course 'superior'" (1991: 74%; 1995: 65%). While both factors would seem on the face of it to be highly interdependent the numeric results show little correlation, raising serious technical questions about both reliability of measurement as well as the correlational validity of the measures themselves. We return to further underlying weaknesses - of a less technical nature but concerning gaps in the *content* or *construct validity* of a great deal of software evaluation - in Section 4 below.

In the analysis of learning potentials by experts, the material itself is in the focus of attention. An expert carefully analyses all potential effects of the materials. Although this assessment method is fairly accurate in terms of learning potentials, it still does not allow any clear statement on actual achievements on the side of the learners. The difference between potential and actual outcomes, however, is of utmost importance since it is widely known that experts and novices employ different strategies when dealing with texts. As a consequence, the competence of the instructors to deal with software is likely to have a significant effect on the achievement levels reached by the learners. However, the competence level of instructors is difficult to measure in itself.

In investigating the main elements found in introspective and empirical evaluations, our research group's survey (Tabyanian and Reeder, in preparation) revealed two main components: product-related components and instructional design and learning approaches. We found that assessments of both sorts of components appeared in some cases to lack coherent connections to best practices in language teaching or current understandings of language learning.

3.2 Product-related components in software evaluation

Most of the evaluation approaches we surveyed include an evaluation of the technical or 'usability' features of software. At this level, evaluation is concerned with the general characteristics of the software itself and the ease with which it can be used. This part of the evaluation, which can be done by educational software experts, determines the presence (or absence) of technical features and the content of the software. Such product-related components can include:

- *Technical aspects*: Technical aspects include implementation considerations, and documentation and packaging. Hardware specifications, cost effectiveness, instructional

and operational manual, suggested classroom activities, and the description of the links and branching techniques between data are among the criteria at this level.

- *Content considerations*: Content considerations include: The accuracy and presentation of the material, consistency regarding the level and nature of the content presented, and the general appropriateness of the material for the typical users (Gros & Spector, 1994).

- *General use considerations*: These often include the quality of the user interface, including menu types, items covered and terminology used in the interface, and support material availability, including adaptability to the Internet. Interaction with users is also covered as part of the usability considerations because of the dominance of communicative language teaching (CLT) approaches.

3.3 Instructional design and learning considerations in software evaluation

- *Instructional design*: While the majority of software developers as well as the evaluation systems agree on the significance of instructional considerations in the development and evaluation of educational courseware, there is no agreement among researchers and evaluators as to what *criteria* to use to assess this aspect of language software. Part of this difficulty could be caused by the absence of instructors on software design teams, creating a gap between design, development, and classroom implementation as noted by Hubbard (1992). Moreover, since the instructional needs of any given classroom are context-dependent, building accurate evaluative criteria into software evaluation systems presents a serious technical challenge to most software evaluation systems. As Leu, Hillinger, Loseby, Balcom & Dinkin (1998, p. 204) put it, "Although new technologies are becoming more widely available they are not always appropriated by teachers and systematically integrated into the curriculum". For the same reasons, software designers can remain unaware of instructional concerns of language educators. A notable exception to this rule is reported in Leu et al. (1998), in which six elementary school teachers involved in designing software for sixth-grade students proposed features that could accommodate their instructional needs. The researchers and teachers involved in this project identified software design themes that guided their decisions at each stage of the design.

- *Learning process considerations*: A look at most software evaluation systems reveals the experimental nature of the evaluation approaches. "The prevailing methodology in the evaluation of software in the classroom is based on an experimental paradigm (control group, test, post-test control, etc.)" (Gros & Spector, 1994, p.38). Lack of a match between course objectives and instructional features included in the design of software seems to be the main reason for the unreliability of most student-gain based evaluations. Our survey of current evaluative approaches found no examples of formative evaluation of software in which not only the learning outcomes but also the learning processes leading to those outcomes were systematically examined, with the notable exception of Murray's (1998, 1999 a, b) innovative observational study of dyads who were using the French language multimedia software *A la rencontre de Philippe*.

4. Shortcomings of Current Approaches to Software Evaluation

In addition to noting the lack of evaluative criteria which measure not only learning outcomes but also learning processes, we identify a number of shortcomings in current evaluative practices in the sections that follow. The question remains as to whether current software evaluation guides could be adapted to address some of these concerns, or whether a qualitatively different approach would be needed to address these issues, to the extent that there is agreement that they are real issues for the future of software assessment.

4.1. Problems of validity and generalizability in experimental evaluation designs

As most empirical methods tend to aim at some degree of generalizability of their results to a population, they try to ascertain and control the effects of intervening factors. This, in turn, often leads to some experimental designs which consist of a test group, a control group and some standardized, highly pre-structured tools of investigation (e.g., a classical questionnaire). This type of experimental design, which has been adapted from the sciences, can prove problematic as it can also have a number of weaknesses:

- *Difficulty in attributing outcomes validly to the treatment.* Instruction processes are of a highly complex nature. It is therefore difficult, if possible at all, to account for all intervening variables. This difficulty normally leads to a design which does not reflect all variables or a lab design which selects a few but, as a consequence has little if anything in common with a 'normal' classroom setting.

- *Invalid reduction of complex learning processes.* A standardized questionnaire, particularly if poorly designed, will often reflect a highly attenuated or even behaviorist concept of learning if it presupposes an overly narrowly-defined set of language learning issues, and then translates these constructs into questions with foreseeable 'correct' answers. Thus, any learning result that has not been predicted will rarely become obvious in such an investigation. In addition, overly narrowly phrased questions may be so close to the topics discussed or tasks mastered in a classroom that little if any generalizability beyond either the test group or the control group is possible. This in turn argues for the usefulness of complementing such narrowly-defined measurement approaches with some more open-ended measures, such as open-ended questionnaire prompts or semi-structured interviews that can be analyzed in more qualitative methods, such as content analysis or discourse analysis. We return to idiosyncratic assessment in Section 5 below.

4.2. Current approaches to software evaluation fail to take educational goals into account

In numerous cases evaluative criteria fail to link the design of software to the instructional methodology of the program. Very few of the software evaluation examples we reviewed provide a methodological framework for CALL courseware development and evaluation. A proponent of such a systematic integration of development, evaluation and implementation, and an exception to our general findings, Hubbard (1992), defines a methodological framework as:

A framework for the description and analysis of methods, which are ultimately nothing more than a set of procedures applied in a consistent and reasoned fashion in the pursuit of a given goal, such as learning to speak and understand a foreign language (p. 41).

Accordingly, Hubbard (1987, 1992) argues for the application of a curriculum development approach to software evaluation as part of a triangular model for CALL courseware, the other two components being development and implementation. In his model, Hubbard (1992) emphasizes the interrelated nature of these components or "modules" as follows:
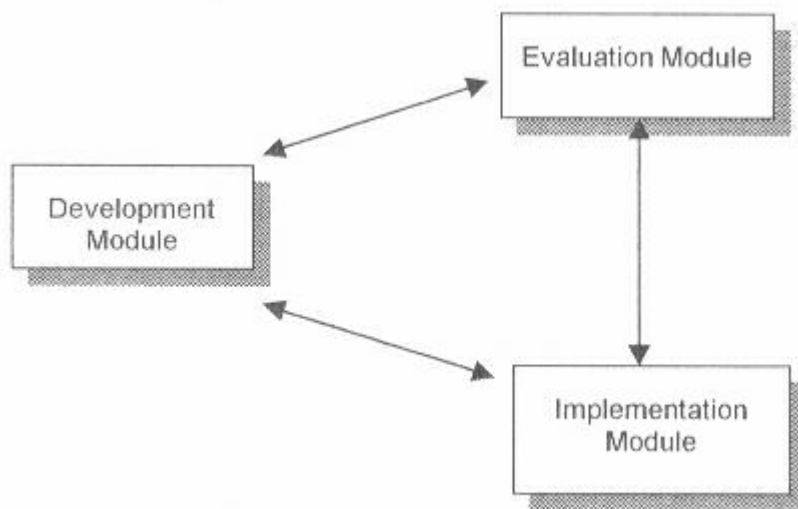
In his "Network" approach, Hubbard argues for a framework in which the development scheme of software is laid out in the same fashion as a curriculum and instructional scheme. In this way, he argues that software development, implementation, and evaluation can be explicitly related to instructional principles and their related components.

4.3 The case of new and multiple literacies

As we are at the turning point of changing understandings of literacy - particularly its status in regard to language education, reading, and writing - traditional approaches to software evaluation will probably lag behind emerging knowledge and theory about literacy and its relationships to language learning. New learning media affect the instructional objectives of literacy and language classrooms, which in turn requires evaluating those technologies accordingly. *Visual literacy*, or the ability to use the graphic elements of multimedia effectively for constructing and conveying meaning, for example, brings about new instructional objectives, which in turn requires new features in software development and, consequently, new criteria for evaluating those features. Perspective changes of this sort requires software developers, including teachers, to rethink the design of educational software to accommodate new objectives emanating from those perspectives. In addition to visual literacy, a descriptively adequate approach to software evaluation will often need to deal with dimensions of *cultural literacy* and *critical literacy* (Edelsky, 1994), *inter alia*.

4.4 The intercultural divide in language learning software design and evaluation

While the development of new information and communication technologies has steadily and rapidly accelerated over recent years, relatively little attention has been paid to the actual use of those technologies by different user cultures (social cultures, gender cultures, learner cultures, national cultures). It is widely assumed that the standardisation of the equipment automatically leads to a standardisation in its application. Research has shown that ways of communicating cross-culturally vary greatly even among related languages (Kramsch, 1993; Roche, 2001a, 2001b). However, little is yet known about cultural attitudes towards information and communication technologies (Roche 2000). As a result, issues of intercultural communication and sociocultural variation amongst user groups have not been adequately reflected in software assessment. In fact, with a few exceptions they have not even been considered in the production of language learning software. For example, this lack of cultural awareness can hardly be better illustrated than by the different versions of *Who is Oscar Lake?* The program has simply been translated into different languages keeping the same name, except for the question word, and exactly the same setting (e.g., the railway station). Only a few items are labelled differently, e.g., *gare* or *Bahnhof* for *railway station*.

However, the significance of intercultural issues is not limited to mere translation of linguistic items. Rather, intercultural mediation involves a complex set of parameters reaching from the linguistic code to cultural values and patterns of viewing and using media. Sometimes, the reluctance to implement cultural awareness in software design is even due to different goals and expectations of the software developers. For example, a recent case study of the struggles of language educators to incorporate awareness of intercultural communication and cultural stereotyping issues into the work of a joint university-industry language software development team is documented in Beckett, McGivern, Reeder & Semenov (1999) in their account of the development of *Edubba*, multimedia software for the enhancement of academic writing in English (Reeder, Hooper, Roche & Shaddock, 2000; Reeder, this volume).

4.5 New media, new modes of language learning

Closely related to the effects of the new curricular assumptions about literacy on technology formation and evaluation is the instructional dichotomy of 'curricular learning' versus free interactive learning paradigms (Lemke, 1998). In contrast to 'curricular learning', in interactive learning paradigms the instructor as well as the institution's syllabus take on a less central or at least quite different roles in learning. This is especially evident with new technologies that have enabled collaborative learning environments in the form of synchronous or asynchronous communication. For example, studies (Heift & Caws, 2000; Wang & Teles, 1998; Kelm, 1992, 1996; Berge, 1995) in Computer-Mediated Communication (CMC) have shown that the instructor takes the less dominant role of a facilitator or mediator and that student participation increases (Kern, 1996). The interaction between emerging possibilities offered by new technologies and their effects on instructional paradigms are of main interest. Technological advances create learning situations different from those of mainly text based traditional classrooms, and these new learning opportunities require new instructional methodology to accommodate them. An efficient evaluative system for language learning software will be flexible enough to factor in these possibilities.

5. A Research Agenda for the Development of New Paradigms for the Evaluation of

New Media

As a first step in meeting the need for a concerted effort to extend the reach of CALL software evaluation to take into account the distinct nature of multimedia language learning software and the teaching and learning that can ensue from its uses, the authors propose a systematic agenda of research and development in this field. We suggest that such an agenda consist of the following four steps:

   a. Description: *constructing and pilot testing a research instrument to describe MMLL materials* in a systematic and consistent manner;

   b. Theory building: *developing a theoretical framework for evaluation* of new media in language learning;

   c. Instrumentation: *deriving from the theoretical framework a suite of new instruments and guidelines for evaluation* of MMLL materials of different types in different development and application contexts;

   d. Empirical studies: *testing the suite of instruments and guidelines* on current MM language learning materials in a representative range of instructional settings with a representative range of users.
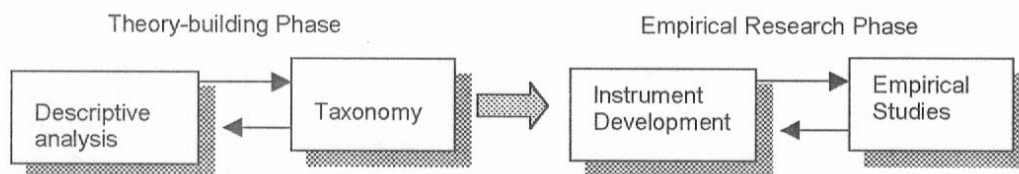
We discuss each of these steps in turn, in an effort to illustrate the sorts of research and development that might be possible in such a program of work. Because our work is in its first phase, we deal in more detail with the descriptive phase of the E/Valuating New Media for Language Development project, and necessarily in less detail with the remaining phases.

5.1 Describing multimedia language learning software

The common trait of the three different software types identified above (microcosm simulations, microethnographies and online programs) is their proximity to authentic or simulated immersion settings. They are all representatives of the most recent generation of programs. We turn to descriptions and recent examples of each main category, by way of exemplifying the sort of work that can be undertaken as a necessary preliminary to any effort to enhance the coherence, validity and usefulness of the educational evaluation of such learning media.

   a. Microcosm Simulations

Microcosm simulations are those programs that attempt to create fictional and non-fictional self-contained worlds reflecting segments of the target culture. These simulations often take the form of stories, ethnographic recordings, and target culture situations. The most common

representatives of this type of software are CD-ROMs containing short clips of target culture communications filmed in authentic or pseudo-authentic environments and subsequently adapted for classroom use. Such programming features seemingly present authentic communication patterns while also providing different help features for learners who need them (e.g., reduced speech rate, vocabulary explanations, reference materials).

- The British program *Business Challenges* (Addison Wesley Longman, 1996), for instance, presents short video clips of business-related situations and offers context-sensitive help features in several different languages.

- Similar in approach are *Travel Tur* and *Einfach Toll*, "interactive" programs for the teaching of Spanish and German respectively (Houghton Mifflin/Interactive Factory, 1996) which feature a number of conversations and settings involving young adults.

- *A la rencontre de Philippe*, developed by Gilberte Furstenberg and her Project Athena team at MIT and published by Yale UP (1994), is among the best known and most sophisticated specimens that use a fictional story to incorporate into language teaching. Filmed on location in Paris by a feature film production crew and scripted by screen writers rather than by educators, the back-story follows the misadventures of Philippe, a young man whose girlfriend has thrown him out of her apartment. The viewers/learners are immediately drawn into the action as they are asked to help Philippe find a new apartment in Paris. The ensuing branching storylines involve learners in a variety of search activities and discussions, some of which aim to resolve the conflict between Philippe and his girl friend. The learners must utilize the tools at their disposal, of which there are primarily two types: those necessary to solve given tasks (such as maps, directories, an answering machine, and a note pad), and those that facilitate learner comprehension (such as play, repeat and preview functions, search and reference functions, transcriptions, vocabulary glosses, cultural notes, alternative soundtracks, including the original colloquial speech of the actors, simplified versions and reduced speech rates).

---

b. Microethnographies

This fairly recent category of language learning software takes advantage of multimedia technology to bridge the gap between (or perhaps even redefine the boundaries between) the learning the language and the learning of culture and appreciation of historical and social contexts of a language and its speakers, often in an interdisciplinary fashion.

- *Berliner Sehen*, a hypermedia documentary under development by Ellen Crocker, Kurt Fendt and their MIT team combines CD-ROMs and other material collections stored on decentralized servers in an attempt to develop a more open learning environment. This program allows students to expand their own archives and to collaborate on the construction over networks of new collections which can then be made available to other users. Both elements of the program use authentic audio and video recordings taped in a Berlin neighbourhood in the late nineties, as well as documentary components from pre- and post-wall periods, and allow the user to reconstruct and experience the chosen microcosm in different ways. (http://web.mit.edu/fll/www /projects/BerlinerSehen.html).

- *Ucuchi* is a first-year course (two CD-ROM discs, software, textbook, Installation/Getting Started Guide, Reference Guide, VHS version of video) for Quechua, the language of the descendants of the Incas. It was filmed in Bolivia in the village of Ucuchi and the nearby city of Cochabamba as an ethnographic documentary. The film was edited to a two-hour set of 20 scenes of natural speech and interaction. It also contains simple word and suffix references specific to the film content and full transcripts. The transcripts can be displayed in either Quechua or its English translation, along with optional detailed glosses. (see Andersen, R. W., & Daza, J. L., 1994; Kramsch, C. & Andersen, R. W., 1999).

- *Star Festival*, a CD-ROM based fiction/documentary developed at MIT for the teaching of Japanese (beginners and intermediate), http://web.mit.edu/fll/www/projects /StarFestival.html.

c. Online Programs

A virtually infinite amount of material, including sites and links containing authentic audio and video sources in addition to synchronous and asynchronous communication could provide ideal opportunities for language learning and teaching. However, while probably not too difficult at advanced levels of language learning, it remains to be resolved how the largely unstructured and frequently overpowering abundance of information presented in a foreign language can be mediated for beginner or intermediate students.

- *PILOT*, a large-scale program for the teaching of German for scientific and technical purposes, combines a structured instructional design with the exploratory options of the Internet. While focussing on those discourse types and genres which are most relevant for advanced students (e.g., reading scholarly publications, listening to lectures, writing research papers, participating in scholarly discussions), the program also offers both basic and advanced assistance and practice on vocabulary and grammar. It contains a large number of exercises which are embedded in the thematic progression of the chapters and address the specific learning conditions of various learner cultures. An intelligent electronic tutor provides feedback on errors and collects acquisition data for research and program development. Online resources such as a news module, a complex communication module as well as help and reference modules complete the program.

A comprehensive program of research on MMLL software evaluation would, of course, need to move beyond detailed descriptions of program features like those noted above to systematic development of relevant categories for the general description of the whole class of MMLL software with which we are concerned, building upon such examples as Hubbard's CALL software selection guide (1992) and incorporating relevant categories such as learning and teaching assumptions and distinctive characteristics of the new media. The goal would be to develop *a descriptively-adequate taxonomy* sufficient to characterize as many of the educationally important features as possible for any example of MMLL software. This would

be the first step toward a theory of evaluation of MMLL software.

5.2. Theory building for evaluation of new media in language learning

Our group suggests that the research agenda needs to include the knowledge bases from applied linguistics, language pedagogy and second language learning at a minimum, if theoretical progress is to continue in the field of language learning software evaluation. To take one specific example of a necessary knowledge base, recent views in language acquisition will serve to illustrate the possibility of interdisciplinary illumination of the field. Current thinking about language acquisition as cognitive construction or "meaning making" (Wells, 1986; Wells & Chang-Wells, 1992) implies that for learning media to be effective, they must initiate and support students' active processes of knowledge acquisition, problem solving and meaning construction. To what extent does the new generation of MMLL software meet these criteria? A rigourous research-based evaluation might reveal that many current programs are insufficiently interactive (e.g., in any open turn-taking activity), leaving inadequate room for learners' creative construction of the second or foreign language. Even an over-generous supply of intelligent 'help' features may pose problems as learners take shortcuts, limiting their opportunities for productive practice and rendering themselves passive. While contemporary pedagogy argues that learners' energies should be directed to actively involve them in the learning process as opposed to simply clicking 'buttons' or 'hotspots' on the computer screen (Davey, D., Gade J. & Fox, J., 1995, 42), it remains to be seen - empirically - whether the latest generation of MMLL software represents a significant improvement in enhancing learning according to such theoretical definitions.

5.3 Developing new tools for the empirical study of new media for language learning

The descriptive-taxonomic phase of a program of E/Valuating New Media for Language Learning would be followed by a theoretical phase, during which relevant categories not only for the description but also for the evaluation of MMLL software would be generated (for an early example of such a theoretical model for MMLL software evaluation, see Tabyanian & Reeder, in preparation). That theoretical work would in turn be followed by an instrumentation phase in which prototype software evaluation tools would be generated from the theoretical model and tested in field settings to determine their validity, reliability, and utility for language teaching.

5.3.1. Tracking systems for assessing learning processes and outcomes associated with MMLL software

In evaluating MMLL software, one of our major goals would be to assess the language learning process as well as the learning outcomes. Tracking systems in the form of computer logs allow researchers to collect accurate information on student-computer interaction and student progress. Moreover, the data will assist researchers to achieve a higher standardization of measurement in software evaluation.

A computer's access log generally contains just the bare details of timing, path and input-response. A visit to a website may be recorded in the server log with the date and time of the request, the originating Internet address, and the system response. Depending upon the

system and task at hand, additional information is usually available and relevant and therefore worth storing. These logs, particularly when supplemented by analysis and information from a system's components, represent a rich data source for determining the validity and efficacy of pedagogical decisions implicit within the system design and content.

A detailed computer log is necessary for researchers to assess the learning process because studies have shown that learners do not always use every option available in the software although - from an instructor's or software designer's point of view - some features might be very valuable and effective. For instance, Cobb and Stevens (1996) discovered that students did not make use of help options although they knew that such use could improve their learning outcome (see also Steinberg 1977, 1989; Chapelle, Jamieson and Park, 1996; Bland, Noblitt, Armington and Gray, 1996). Moreover, Heift (2001, 2002) found that students showed distinct interaction patterns with the software depending on their language skill level. For example, lower performers made more use of system help options than mid and high performers. Given the outcomes of these studies, an instructor's or designer's judgement of the software cannot measure the learning process as accurately as a computer log.

A tracking system can also provide accurate and relevant information on the learning outcome. A detailed computer log on student input will allow researchers to create a student profile over time. This information can then be used to analyze student language skill level.

   5.3.2. Complementary and qualitative analyses of learning processes and outcomes associated with MMLL software

Partly to safeguard against the sorts of dangers we identified in Section 3 above (reductionism, limitations to generality of findings), we also propose that a comprehensive approach to the evaluation of MMLL software will include research methods of a more naturalistic, observational nature that entail analyses of a "softer", more qualitative nature. Such methods include observational approaches that are structured to a greater or lesser degree.

- *Video observations of users interacting with the program and with one another.* These methods are particularly appropriate for those programs that actively promote interaction not only with the program but with a learning partner either in dyadic or small group arrangements. One of the best recent examples of this type of research method applied to MMLL software is Murray's (1998) ethnographic analysis of dyads using the French language program *A la rencontre de Philippe*. Other video-observational methods of the more highly structured type can track users' eye movements, visual lines of regard, or other elements of facial expression or body language in order to provide data on the users' cognitive and affective engagements with a program. Such visual means of research and evaluation seem to us particularly appropriate for programs that contain a high proportion of graphic material.

- *Audio recordings of users*. Hart (2000) studied in detail the ways in which partners collaborated with each other to develop simulated news reports on an environmental issue that was the subject of an animated program, *Edubba* (Reeder, et al, 2000). Hart tracked all of the "writing partners'" conversations using the audio-recording capacity of

language lab workstations, analyzing these by means of a qualitative discourse analysis approach, and generating a taxonomy of collaborative styles that emerged, dissolved or advanced throughout the partnerships' development as writers and editors. Those collaborative styles were not of mere sociological interest, for they bore critically on the degree of efficiency with which the learners engaged in the assigned task.

- *Structured and semi-structured interviews.* Such methods, while subject to many of the weaknesses we have already identified above, can complement the precision afforded by logging/tracking observations and even capture elements that video and audio observations will miss. Their specific advantage is that they offer opportunities to evaluate in a fairly direct manner specific elements of the learners' intellectual, attitudinal, emotional or aesthetic responses to their experiences using MMLL software. The major technical challenge for such approaches remains the reliability of measurement, and safeguards such as multiple ratings can ensure that data analyses are reasonably consistent from case to case, for example.

5.4 Striking a balance: general considerations in designing research and evaluation methods for new media

Our research group believes that a widely applicable approach to evaluation design must be flexible enough to encompass different overall aims of instruction defined by the authors of any given instructional program. This breadth can include a wide range of dominant learning issues and approaches to language learning, e.g., grammatical correctness versus communicative and/or intercultural orientation. Learning approaches may either aim at a more or less perfect imitation of a native speaker or at a competent intercultural interlocutor. Consequently an assessment of learning outcomes will have to take into consideration varied instructional objectives.

Just like most scholars in the field of language teaching and learning today, we do not regard the mere reproduction of facts the central or exclusive aim of instruction. It follows that evaluation materials should not be too closely related in content to the learning materials being assessed. In addition, if the assessment materials are not too closely related to specific teaching materials they can then be applied to the evaluation of a wider *range* of course types and hence be of more general utility to the profession. An orientation by global instructional aims would preserve a desirable distance from individual courses, but still be appropriate for use in materials with somewhat different detailed learning objectives. Our suggestion, therefore, is to use overall educational objectives formulated in individual language learning curricula as the basis for evaluating the efficiency of any particular language learning software. In cases where those broad objectives or global aims are not explicitly formulated, as in *Berliner Sehen*, they may be re-constructed from the materials itself. In *Berliner Sehen*, for example, listening comprehension and some degree of *Fremdverstehen* may be re-constructed as implicit global aims.

In addition, we propose that the process of designing evaluation instruments be a dialectic process. This, on the one hand, takes into account the rich and varied nature of MMLL learning materials and is, on the other hand, a necessary consequence of post-behaviorist theories of language learning which do not assume an omniscient or even omnipresent

instructor. In order to illustrate this disposition, let us choose a global aim such as 'understanding foreign cultures' as an example. Improvements in ability to do so could be tested by confronting learners with video material from an intercultural encounter that contains some covert or overt intercultural difficulty. The learners, then, could be asked to analyze the encounter, which would elicit relatively little pre-structured feedback. As is common in social research, the actual criteria which are taken as indicators of learning results or progress may be developed in dialectic processes on the basis of learners' productions. The basis of the evaluation and comparison of software, then, is formed by sufficiently abstract and empirically constructed categories. However, it is important to keep in mind that these categories remain an open class in order to provide a suite of evaluation tools flexible enough to deal with new generations of learning software and hypertext learning (e.g., not only naming learning deficits in a comparison with linear texts but also being able to qualify improvements).

An open, minimally pre-structured approach to the empirical evaluation of MMLL software is compatible with a constructivist theory of learning. As noted earlier, such an evaluation approach may be complemented by additional assessment modules - perhaps of a more immediately quantifiable nature which focus on specific skills (such as grammar competencies) - and make use of the learners' mother tongue as a means of determining comprehension (e.g., using the foreign language to understand and the mother tongue to produce a text which is based on the understanding in the foreign language). A useful and effective balance between the rigourous quantification of language learning behaviour and qualitative assessment tools that allow us to capture subtle learning processes, dispositions and outcomes can be struck, in our view, by adopting such a dialectical, and inclusive approach to the design of empirical tools for new media software evaluation.

6. Conclusion

Our research group hopes to carry such a four-phase research and development agenda forward on a large scale in an international project currently in the advanced stages of planning. The authors, representing four institutions in Canada and Germany, will be joined by colleagues at MIT (Cambridge, USA) and Ritsumeikan University (Kyoto, Japan) in a four-nation study of the theory of evaluation of multimedia language learning software. We anticipate that our project will illuminate not only the distinct characteristics of the new genre for language learning software we have described here and discover more about the novel kinds of learning it promotes, but also that it will apply very current understandings of language learning and language teaching. We hope that our diverse backgrounds, ranging from computational and applied linguistics to specialists in EFL will assist us to achieve that eclectic, inclusive approach to evaluation that we aspire to.

One of the outcomes we hope to produce is an online, readily-updated technical manual of MMLL software reviews in particular, and what we are calling "E/Valuation" techniques which we hope will be of considerable assistance to the profession. A second outcome we hope to achieve is the founding of an International Institute for Language Learning Software Evaluation (IILLSE). Third, we hope that our project, by building upon the pioneering work of Hubbard and many others cited here, will not only extend our scholarly understandings of

learning and teaching with a remarkable new generation of software, but also extend and disseminate professional knowledge about best practices for software integration into our classrooms and labs. Perhaps we will all be in a better position to answer our earlier question, "Is this multimedia language software effective?" by addressing first the more fundamental question "How do we go about learning about the effectiveness of the new generation of language learning software?"

References

Andersen, R. W. & Daza, J. L. (1994). *Ucuchi: Quechua live and in color!* [CD-ROM]. Los Angeles: UCLA, Authentic Discourse Research and Development Project. A PDF download of a description of *Ucuchi* is available at http://www.humnet.ucla.edu/humnet/AL/CLRL/media.html (April 2001).

Beckett, G., McGivern, L., Reeder, K., & Semenov, D. (1999). Dilemmas in designing multimedia software for language learners: Action research. *Journal of Research in Computers in Education, 32 (2)*, 287-297.

Berge, Z.L. (1995). Facilitating computer conferencing: Recommendations from the field. *Educational Technology*, *35 (1),* 22-30.

Berliner Sehen. Hypermedia documentary for foreign language learning. Available: http://web.mit.edu/fll/www/projects/BerlinerSehen.html (April 2001).

Bland, S., Noblitt, J., Armington, S. & Gay, G. (1990). The naive lexical hypothesis: Evidence from computer-assisted language learning. *Modern Language Journal, 74 (4)*, 440-450.

Chapelle, C., Jamieson, J., & Park Y. (1996). Second language classroom traditions: How does CALL fit? In M. Pennington. (Ed.). *The power of CALL.* Athelstan Publications.

Child, J. (1997/1998). Assessing the impact of computer-assisted instruction (CAI) in undergraduate Latin American studies courses. *Computers and the Humanities, 31 (5)*, 389-407

Cobb, T. & Stevens, V. (1996). A principled consideration of computers and reading in a second language. In Pennington, M. (Ed.). *The power of CALL.* Athelstan Publications.

Davey, D., Gade J. & Fox, J. (1995). Multimedia for language learning: Some course design issues. *Computer Assisted Language Learning, 8 (1)*, 31-44.

Edelsky, C. (1994). Education for democracy. *Language Arts, 71 (4)*, 252-257.

Furstenberg, G. (1994). *A la rencontre de Philippe* [computer software]. M.I.T. Athena Language Learning Project. Yale University Press. Available: http://web.mit.edu/fll/www

/projects/Philippe.html

Gros, B. & Spector, J. M. (1994). Evaluating automated instructional design systems: A complex problem. *Educational Technology, 42,* 37-46.

Hart, G. L. (2000). *Collaborative writing strategies of students using multi-media software.* Unpublished master's thesis, University of British Columbia. Vancouver, Canada.

Heift, T. & Caws, C. (2000). Peer feedback in synchronous writing environments: A case study in French. *Journal of Educational Technology & Society, 3 (3)*, 208-214.

Heift, T. (2001). Error-specific and individualized feedback in a web-based language tutoring system: Do they read it? *ReCALL, 13 (2)*, 129-142.

Heift, T. (2002). Learner control and error correction in ICALL: Browsers, peekers and adamants. *CALICO Journal, 19 (3),* in press.

Hubbard, P. (1987). Language teaching approaches, the evaluation of CALL software, and design implications . In W. Flint Smith. (Ed.). *Modern media in foreign language education: Theory and implementation.* Lincolnwood, IL: National Textbook Company.

Hubbard, P. (1992). A methodological framework for CALL courseware development. In D. Sharp (Series Ed.), M. C. Pennington & V. Stevens. (Vol. Eds.). *Computers in applied linguistics: An international perspective* (pp. 39-65). Clevedon, Avon: Multilingual Matters.

Kelm, O. (1992). The use of synchronous computer networks in second language instruction: A preliminary report. *Foreign Language Annals, 25 (5)*, 441-454.

Kelm, O. (1996). The application of computer networking in foreign language education: Focusing on principles of second language acquisition. In M. Warschauer. (Ed.). *Telecollaboration in foreign language learning* (pp. 19-28). Hawaii: Second Language Teaching and Curriculum Centre.

Kern, R. (1996). Computer-mediated communication: Using e-mail exchanges to explore personal histories in two cultures. In M. Warschauer. (Ed.). *Telecollaboration in foreign language learning* (pp. 105-119). Hawaii: Second Language Teaching and Curriculum Centre.

Knowles, S. (1992). Evaluations of CALL software: A checklist of criteria for evaluation. *ON-CALL, 6 (2)*, 9-20.

Kramsch, C. & Andersen, R. W. (1999). Teaching text and context through multimedia. *Language Learning & Technology. 2 (2)*, 31-42. Available: http://polyglot.cal.msu.edu /llt/vol2num2/article1/index.html (April 2001) [Der Link existiert leider nicht mehr. Stand: 3. Januar 2005].

Kramsch, C. (1993). *Context and culture in language teaching.* Oxford: Oxford University Press.

Lemke, J. (1998). Metamedia literacy: Transforming meanings and media. In D. Reinking, M. C. McKenna, L. D. Labbo, & R. D. Kieffer. (Eds.). *Handbook of literacy and technology. Transformations in a post-typographic world* (pp. 283-301). Mahwah, NJ: Erlbaum.

Leu, J. D., Hillinger, M., Loseby, P. H., Balcom, M. L., & Dinkin, J. (1998). Grounding the design of new technologies for literacy and learning in teachers' instructional needs. In D. Reinking, M. C. McKenna, L. D. Labbo, & R. D. Kieffer. (Eds.). *Handbook of literacy and technology. Transformations in a post-typographic world* (pp. 203-220). Mahwah, NJ: Erlbaum.

Murray, G. (1999a). Autonomy and language learning in a simulated environment. *System, 27 (3)*, 295-308.

Murray, G. (1999b). Exploring learners' CALL experiences: A reflection on method. *Computer Assisted Language Learning, 12*, 179-195.

Murray, G.L. (1998) *Bodies in cyberspace: Language learning in a simulated environment.* Unpublished doctoral dissertation, University of British Columbia. Vancouver, Canada.

Reeder, K. (2001). *Edubba:* Multimedia support for academic writing in English. *ZIFU, this volume*.

Reeder, K., Hooper, H., Roche, J., & Shaddock, P. (2000). *Edubba: The language arts program.* Unpublished CD-ROM. Lunny Communications Group and the University of British Columbia. Vancouver, BC. Available: http://www.lunny.com/edubba/ [Der Link http://www.lunny.com/edubba/ existiert leider nicht mehr. Stand: 1. Januar 2004]

Roche, J. (2000). Lerntechnologie und Spracherwerb - Grundrisse einer medienadäquaten, interkulturellen Sprachdidaktik. *Deutsch als Fremdsprache, 2*, 136-143.

Roche, J. (2001a). *Interkulturelle Sprachdidaktik - Eine Einführung.* Tübingen: Gunter Narr Verlag.

Roche, J. (2001b). Vom Medienmangel in der Lehre und dem Lehrmangel in den Medien. Wissenschaft und Philosophie. In Venanz Schubert. (Ed.). *Wissenschaft und Philosophie. Interdisziplinäre Studien, Band 23: Die Geisteswissenschaften in der Informationsgesellschaft.* St. Ottilien, in press.

Schmueckler, L. M. & Shuell, T. J. (1989). Comparison of software evaluation forms. *Journal of Educational Computing Research, 5 (1)*, 17-33.

Scholfield, P.J. (2000). *"Evaluation of CALL software". Work in progress.* Department of Language and Linguistics, University of Essex UK. Available: http://privatewww.essex.ac.uk/~scholp/calleval.htm (April 2001).

Steinberg, E.R. (1977). "Review of student control in computer-assisted instruction." *Journal of Computer Based Instruction, 3,* 84-90.

Steinberg, E.R. (1989). Cognition and learner control: A literature review, 1977-1988.

Journal of Computer Based Instruction. 16 (4), 117-121.

Wang, X. & Teles, L. (1998). Online collaboration and the role of the instructor in two university credit courses. In T. W. Chan, A. Collins, & J. Lin. (Eds.). Global education on the Net. *Proceedings of the Sixth International Conference on Computers in Education, Vol. 1.* (pp. 154-161). Beijing and Heidelberg: China High Education Press and Springer-Verlag.

Warschauer, M. (1999). *Electronic literacies: Language, culture, and power in online education.* Mahwah, NJ: Erlbaum.

Wells, C.G. & Chang-Wells, G.L. (1992). *Constructing knowledge together: Classrooms as centres of inquiry and literacy.* Portsmouth, NH: Heinemann.

Wells, C.G. (1986). *The meaning makers.* Portsmouth, NH: Heinemann.

Who is Oscar Lake? Interactive language software. Available: http://www.languagepub.com/oscar/oscar.html (April 2001).