Evaluation in a project life-cycle:

the hypermedia CAMILLE project

Thierry Chanier

Université de Clermont II and Université de Paris V

In the CAL literature, the issue of integrating evaluation into the life-cycle of a project has often been recommended but less frequently reported, at least for large-scale hypermedia environments. Indeed, CAL developers face a difficult problem because effective evaluation needs to satisfy the potentially conflicting demands of a variety of audiences (teachers, administrators, the research community, sponsors, etc.). This paper first examines some of the various forms of evaluation adopted by different kinds of audiences. It then reports on evaluations, formative as well as summative, set up by the European CAMILLE project teams in four countries during a large-scale courseware development project. It stresses the advantages, despite drawbacks and pitfalls, for CAL developers to systematically undertake evaluation. Lastly, it points out some general outcomes concerning learning issues of interest to teachers, trainers and educational advisers. These include topics such as the impact of multimedia, of learner variability and learner autonomy on the effectiveness of learning with respect to language skills.

Introduction

This paper reports on a series of evaluations undertaken in the countries which participated in the CAMILLE project.[1] The principal aim of this European project has been the development and delivery of hypermedia courseware in Dutch, Spanish and French. The courseware encompasses the training of general linguistic competencies for beginners (Dutch and Spanish) as well as competencies related to the use of language for specific purposes (French). The target audience includes students in science or business, and technicians or engineers from SMEs (Small and Medium Enterprises – small businesses). This report may be of interest to two kinds of reader of this journal, as follows.

• Each one of our packages exploit the full range of hypertextual and multimedia facilities currently provided by standard computing platforms. Furthermore, each package offers learners a large-scale learning environment capable of supporting autonomous study. Consequently, these preliminary outcomes relating to the way

CAMILLE has been practically used by learners and to its effectiveness are of potential interest to teachers, trainers and educational advisers.

• The various experiments conducted by the teams and integrated into the process of software development will be of interest to Computer-Aided Learning (CAL) developers in general. Indeed, within the CAL literature, the issue of integrating evaluation into the life of a project, i.e. either in the course of the development or at the end of it, has often been recommended but much less frequently reported, at least for this type of environment. The paper discusses the constraints, advantages and drawbacks of actually adopting such a procedure.

In order to make clear both the nature of the experiments undertaken within CAMILLE and the significance of the results obtained, a brief preamble on evaluation is necessary. The term *evaluation* is widely used by various groups connected with Computer-Assisted Language Learning (CALL) but frequently approached from very different perspectives, and this can leave the reporting of results open to misinterpretation. At one end of the spectrum there is an increasing pressure on researchers and developers to adopt more methodological and scientific procedures, and, at the other end, educational advisers and executives constantly require concrete and positive results before extending their support to CALL. CAMILLE is one project, among an increasing number of others, which has had to try to make these potentially contradictory viewpoints coexist.

Below I describe various aspects of evaluation in language learning and in CAL. After this, I set out the initial requirements and achievements of the CAMILLE project and introduce the common features of the different experiments. This is followed by detailed evaluations made in two countries, and a report of the main general outcomes, summing up our experience of managing evaluation as an integral part of a project life-cycle.

Preamble on evaluation

In order to delimit the framework adopted in this research, this section presents the principal functions of evaluation, the initial questions in the design process, the overlapping forms of evaluation, and the evaluation procedure.

Functions of evaluation
For almost thirty years, a distinction has been frequently made between two principal functions of evaluation: formative evaluation and summative evaluation. This distinction exists in language teaching (Lussier, 1992) as well as in CAL (Knussen *et al*, 1991; Demaizière and Dubuisson, 1992; Mark and Greer, 1993), but they are differently interpreted.

In language teaching, formative evaluation consists in regularly diagnosing the learner's state of knowledge, abilities, attitudes. It is undertaken for learners in order to let them know their current position with respect to their final goals; and for teachers to gain information that may lead them to adjust and adapt their teaching before the end of the course. In CAL, formative evaluation also occurs before the end of the implementation phase. It is intended to help the designers review their progress towards achieving the goals of an educational innovation. It is set up by designers, and involves a few learners who are carefully observed in order to assess whether they use the software as intended.

Such aspects as interface, human-machine interaction, learner strategies, hardware configuration and computing architecture are observed with rather informal methods. This process brings both detailed and general information, which may lead to surface changes (correction of bugs) or more profound changes in the design and the development. It also provides insights into the way the courseware will perform when integrated into a real-life learning situation.

In language teaching, summative evaluation comes at the conclusion of a course, or a programme, in order to measure the level of proficiency acquired by a learner with respect to normative goals explicitly fixed by the learning institution. It is a global measure which compares the performance of learners. It is intended to certify learners in order to give them credits, to recommend an orientation, or to check the effectiveness of the course or programme. In CAL, summative evaluation is concerned with the evaluation of completed systems. Its purpose is to measure the effectiveness of an innovation in terms of its stated aims. It is intended for trainers, centres and designers to assess the suitability of the software for certain tasks and users, or to compare it with other products already in use. In both cases, summative evaluation has to be undertaken in real-learning settings, and to involve a larger number of subjects than formative evaluation.

Since the central topic of this paper is the role of evaluation in developing a software package, I will adopt the CAL standpoint rather than that of language teaching. Moreover, since the computing environments developed in the CAMILLE project have been designed to support autonomous learning, some aspects of the language-teaching model would be inappropriate. However, beyond the discrepancies between CAL and language-teaching models of evaluation, there is a common feature which distinguishes them from the issue of assessment. Evaluation is not a judgemental but a decision-making process. Since outcomes may be interpreted by various audiences (e.g. designers, teachers, institutions) in order to make lasting changes, the framework for setting up an evaluation and its procedure will be examined hereafter.

Initial questions in the design of an evaluation
Evaluating a language program, or any piece of CALL software, is a complex process. There follow some key questions (taken from Nunan, 1992, chapter 9) that should be answered before starting any evaluation.

Objectives: What is the purpose and who is the audience of the evaluation (for whom is it made)?

Methodology: What principles of procedure should guide the evaluation? What tools, techniques, and instruments are appropriate?

Material constraints: Who should carry out the evaluation? When should it be carried out? What is the time-frame and budget?

Release: How should the evaluation be reported?

It may seem obvious that it is extremely important to clarify, from the beginning, the goals of the evaluation. However, it is not a straightforward task. Let us consider an innovation. Relationships are not clear at all between the original working hypotheses of designers, the actual achievement, and the selection of precise experimental variables: a

shift may have appeared between the starting and end points; an innovation may have unexpected effects (it may not raise the level of proficiency, but the learner's motivation); comparisons with other existing learning environments may be problematic simply because they are so different. For example, determining a scale for measuring effectiveness with respect to communicative goals and specific purposes was an expected outcome, in itself, in the CAMILLE project. Fixing the objective of an evaluation is again not always easy when the audience is diversified: designers, teachers, administrators, and funding bodies often have different perceptions.

If learning objectives need to be elicited, they also need to be associated with precise forms of evaluation which are themselves associated with different methodological approaches. Below, I extract overlapping forms of evaluation from one (Knussen *et al*, 1991) out of many possible presentations.

- *Experimental.* A limited number of clearly defined variables are scientifically measured, usually based on statistical inferences. The laboratory is the traditional setting for evaluations which generally have a formative function. If such a form is considered as more scientific, its relevance to real learning settings is problematic.

- *Research and developmental.* The purpose is to apply quasi-experimental methodologies, including pre- and post-tests, in situations closer to real learning settings. This form, which many evaluations of CAL systems try to adopt, also requires clear statements of measurable objectives. They may be easier to guarantee in scientific or industrial environments than in educational ones. They more often concern summative than formative functions.

- *Illuminative.* Isolating variables and associated parameters, as well as quantifying measures, is hard to achieve in real learning settings, especially if estimation of the impact of social factors and the participant's views on the meaning of educational innovations are at stake. Consequently, methods, essentially qualitative and usually based on observations and interviews, are applied to 'illuminate' important factors rather than to test hypotheses. Associated pitfalls here range from the risk of observers' obtrusiveness to findings which cannot be generalized to apply to other settings.

- *Teacher as researcher.* Since teachers play a prominent role in the integration of CALL systems into the curriculum, it seems natural to let them take charge of the evaluations. Biases (e.g. subjectivity, role-conflict, work overload) introduced by this form suggest that it should be used only in addition to other approaches.

- *Case studies.* Understanding the effects of situational and personal factors in the use of innovative software is generally based on the detailed study of a restricted number of learners. However, a generalization of the findings to other situations may be difficult.

In CAMILLE, two forms of evaluation have mainly been used: the research and developmental one, and the illuminative one.

Once the form of the evaluation has been determined, material constraints need to be appreciated before performing the evaluation procedure. The first step of the procedure consists in designing the whole evaluation. The initial task of the second step is the

construction of the instruments: materials for the tests, questionnaires and forms, and extra materials for the control group, if necessary. Data collection and analysis follow, according to the methodological approach chosen. The third step, drafting the report, learning and deciding from it, may not be the last one. In formative evaluation, immediate decisions may be taken, followed by changes which then will be measured a second time.

The variety of tasks and their co-ordination create genuine obstacles for the successful completion of a project. Who will the evaluators be? What is the time-frame and the budget? This may explain why many CALL developers seem reluctant to include an evaluation procedure as part of their project.

The last issue, raised in the initial questions, refers to the release: how is the evaluation to be reported? On one hand, evaluation is often described as a public act which should be open to inspection. On the other, unsatisfactory findings, and/or disagreements between participants, may impede the publication of a final report. Alternatively, interesting findings may be over-generalized if the final conclusions are not clearly delimited.

General aspects of the CAMILLE evaluation

The European LINGUA CAMILLE project started in 1993 and will finish this year. Descriptions of the project and of its theoretical standpoints can be found in Ingraham *et al* (1994); Chanier (1996); and Pothier (in press). In this section I recall only its initial requirements and its main achievements. From there, I examine the purposes and audiences of the various evaluations, the common features shared by the different experiments, and details of evaluations.

Initial requirements of the whole project
The CAMILLE project is aimed at conducting a large-scale experiment touching on issues arising from both pedagogical and software-engineering viewpoints.

From a pedagogical viewpoint, hypermedia technologies are often presented as an opportunity to enhance language learning. Although these factors are often assumed to play an important role in the acquisition process, as yet there have been no large-scale experiments based on their use. CAMILLE was thus seen as an opportunity to undertake such an examination in a multi-cultural environment. The objective was the construction of an environment that would provide learners with all the tools and information, short of a live teacher, that they might need to undertake a specific level of course in the target language. One consequence was the integration of books/resources (on lexicon, culture, function, grammar) with the textbook (the course proper) on the same desktop. Another consequence was the mode of its use and of its integration into a whole curriculum: CAMILLE was designed to be used by well-motivated adults, who may or may not be engaged in formal education or training and who may or may not have access to a tutor. Thus the emphasis was on autonomy.

From a software-engineering viewpoint, hypermedia programming tools are often recommended as an opportunity to speed up courseware development, and therefore make CAL a realistic complement for training learners in and out of the academic world. But production of courseware in hypermedia also dramatically increases the number of skills required, and up to now our experience in reusing modules of software, or shared knowledge

for large-scale software, is still very limited. The CAMILLE project was supposed to help to gain a clearer understanding of trans-national courseware development. As a starting point, it was decided to use a common template for development, a template which consisted of a software and hardware platform, created by our British partners in 1991/92. The effectiveness of the software-engineering viewpoint was enforced by the decision to launch a commercial release at the end of the project, i.e. at the beginning of 1996.

Main achievements
A few months before the conclusion of the project, the main courses finalized or near completion are as follows.

*   *Español Interactivo, Interactif Nederlands,* and *France Interactive.* These three packages are respectively designed for the training of general linguistic competencies for beginners in Spanish, Dutch, and French, and developed in Spain, The Netherlands, and the UK.

*   *Travailler en France.* This package has been designed for the training of competencies related to the use of Language for Specific Purposes (LSP) for intermediate-advanced level French, and developed in France.

Each package includes two CDs which run on a standard, basic hypermedia PC platform, namely the international standard MPC2. This has the minimal equipment to play full-motion video, and offers good quality for recording and playing sounds. Each disc gathers approximately 30 minutes of original video, plus other oral, graphical and textual data, on top of which are built resources, and several dozen activities, which offer more than 20 hours of study to the learner.

While, at present, debugging and some coding processes are still under way, CAMILLE partners are fixing the legal aspects in order to start the commercial release of the most advanced courses.

Common features of the evaluations
Following the general framework discussed above, I review here the common features of the evaluations undertaken by all the CAMILLE partners.

Three kinds of audiences with their respective purposes can be distinguished.

The first encompasses the European Union and publishers, as external (to the CALL community) actors which intervene in the project life. The former (the EU) partially funded CAMILLE (actually for less than a quarter of the total budget), added its own requirements, and annually examined achievements before deciding any extension of funding. The latter (the publishers) have recently undergone internal restructuring in order to be prepared to release multimedia software. Most limit the major risks linked to innovations by expecting developments to be supported by small, recent private ventures. Furthermore, they are not accustomed to dealing with academic institutions. For them all, evaluations were intended to assert our reliability, by proving that learners could turn their hand to our courseware in real settings, by convincing them that academics could challenge private companies and be more transparent when performing evaluations as public acts open to inspection.

The second kind of audience is the CALL community, which includes teachers and

researchers. The pedagogical perspectives outlined above needed to be made explicit. The purpose here is twofold: firstly, measuring what kinds of language skills multimedia technologies can help practise, what sorts of learning strategies are performed in hypermedia environments, and how effective autonomous learning is in various settings; secondly, scaling effectiveness with respect to communicative goals and specific purposes. The latter point refers to the problem of finding criteria by which educational objectives can be measured: how can we assess the learner's ability to master knowledge and skills, mobilized around the specific purposes of each piece of courseware, to transfer them and create new pieces of discourse (cf. de Landsheere's trilogy, 1984)?

Developers constitute the third kind of audience. The purpose here was to appreciate to what extent formative evaluation is necessary for adapting and debugging the learning environment, to perform summative evaluation to clarify the software goals (i.e. exactly identify what can be measured) and to determine the constraints and overheads brought upon the whole project.

The different CAMILLE research teams set up evaluations, located either in their own institutions or in neighbouring ones. This happened over 14 months (1994-1995) during implementation, or at the end of large parts of it. No extra budget, nor extra human resources, were available. The results of these evaluations are being reported in three different ways: the final report to the European Union, conferences such as Eurocall (Emery *et al*, in press) and academic papers.

Details of the evaluations

In the CAMILLE project, objectives and methodologies varied from one research team to another. As an illustration, in this section I detail evaluations undertaken on *Interactif Nederlands* and *Travailler en France*. Results drawn from *France Interactive* and *Español Interactivo* are included in the general outcomes presented in the next section.

Evaluation at HEBO (De Haagse Hogeschool, The Netherlands)
The Dutch CAMILLE team performed both formative and summative evaluations. The formative side of the evaluation was designed as a two-round experiment. As soon as data was analysed, changes were made and new experiments were based on the modified software. The aim of the summative evaluation was to compare the software with local classroom learning. This second side directly interested the HEBO managers and the local teachers. The school supervised more than a thousand university Dutch or foreign students who needed intensive training in several languages for professional purposes (legal or business). It offered a strong integration of CALL into the curriculum: nearly 50% of the students' work time in language learning was organized around free access to computers. Heads of the school consider the familiarization with the Dutch language and society by foreign students as an important factor of integration into a country where they are spending several years. Of course, Dutch is not a 'survival' language (learners can easily talk English and be understood by anyone in everyday life), but attendance in Dutch classrooms is strongly encouraged, though not mandatory, and learners' credits can easily be transferred. It was thus decided that learners who learned Dutch only through *Interactif Nederlands* would take the same oral examination as the other learners of Dutch who attended classroom sessions.

The experiment took place at the HEBO in the multimedia, free-access room. Evaluators used a network version of the software on computers with the recommended hardware configuration. Sixty local students were involved, on a voluntary basis. They were true beginners in Dutch, but experienced language learners (Dutch often being their third language) had a low motivation for learning Dutch, and only basic experience with computers. Learning tasks were organized around half of the software, which represented 30 hours of work, distributed over 10 weeks, with free-access conditions. Learners had to fill in questionnaires and were interviewed at the end of each session. Evaluators also made non-systematic observations. Data from 14 students was analysed for the formative part of the evaluation. This analysis will not be detailed here, but the lessons evaluators learned from this experiment will be mixed with the other general outcomes in the next section. The final examination was organized by the usual Dutch teachers, not by the evaluators. Marks and teachers' comments on the CALL group showed that results were neither better nor worse than usual. Since the timing and the assessment procedures were the same as for the live course, the software would appear to be efficient in this sort of situation and with these types of learners.

Evaluations in CAVILAM (France)

Formative and summative evaluations of Module 1 of *Travailler en France* were organized at two different stages of the project: the formative evaluation in October 1994, at the very end of the development of the prototype of Module 1, and the summative one in June 1995, after changes and debugging had been finished on Module 1 and while Module 2 was under development. Before considering the details of these evaluations, it will be helpful to consider certain common features

Local students, who were following full-time language training periods of 1 to 6 months in length, participated in the evaluation. They were between 21 and 47 years of age, with an average age of 25, coming from various continents and cultures. All were intermediate (200 hours) or advanced (400 hours) learners of French, with French often being their third language. They had good professional motivation for learning, either because they already had a job, or were seeking work where the mastering of specific skills in French was important, or because they wanted to attend French universities. They had a mixed experience with computers, some being almost computer-illiterate as they came from countries where computers are not part of the work or study environment.

Both evaluations were undertaken with Module 1, where the specific purpose is to learn how to apply for a job in France. This makes a noticeable difference from other CAMILLE courses which are for general purposes (Chanier, 1996). The module is built around one main task: making a job application. Knowledge bases and activities allow learners to fulfil the task and immerse them in a socio-cultural context which determines the architecture of the software. The story-line of the module presents two characters who are very different in nature and who encounter a series of representative situations, for example how to find appropriate information and acquire experience in the employment market; how to write a letter of application and a CV in the French way; how to make an appointment on the telephone; how to handle an interview. Linguistic knowledge and activities have been designed from the task context, but do not have top priority. The learning tasks require a total amount of 20 to 25 hours' work over three weeks. The

learners used the software during the time usually allocated for practical work in their training, and had further opportunities for free access.

Formative evaluation

One purpose of the formative evaluation was to measure the performance of the courseware. The second one focused on how effectively the kinds of activities and resources available matched the learners' strategies and interests. The sample population was limited to five volunteers because we wanted one of our observers always to be present. They could work alone or in a group. The observer, who acted in a non-obtrusive way, either video- or audio-recorded all the sessions, and took detailed notes on the learners' moves, selections and timing. Learners filled in pre- and post-questionnaires and had a form to fill in at the end of every session, followed by a short interview.

Through this procedure we were able to collect detailed information about the learners' behaviour and reactions as well as their (positive) comments. All this helped us to make subsequent adaptations. Details are discussed below, but one point is worth mentioning at this stage because it relates to the LSP aspect of our software. Even when learners were not directly, personally concerned with seeking a job, they all (even subjects of the summative evaluation) indicated that the experience provided important discoveries concerning socio-cultural aspects of the target-language country, and of its everyday native language. Apprehending variations in the target language and links between language and complex situations encountered daily by natives is an efficient way of raising language awareness; as such, it is an important aspect of second-language learning.

Summative evaluation

The purposes of the summative evaluation were threefold:

- assessment of the suitability of the first LSP courseware with respect to the local learners,

- comparison with autonomous (audio + paper) learning,

- measurement of the impact of hypermedia CALL on vocabulary learning.

For this second experiment, the audience was not limited to the project team. The CAVILAM staff were also interested in the outcomes, and took over the supervision of the learning task, acting as counsellors. The project team only handled the various tests.

Subjects were divided into two groups on a voluntary basis: group 1 (G1), the paper-and audio-based group, comprised six people; group 2 (G2), the CALL one, seven. For G1 we extracted large parts of textual data contained in the software activities and resources, and all the sounds of the dialogues. They then had a document and audio-cassettes to work with. They also had access to paper-based dictionaries available in the language laboratory.

We prepared pre- and post-questionnaires, the post-questionnaire contents being different for G1 and G2. We also translated into French and administrated the SILL (Strategy Inventory for Language Learning test: Ehrman and Oxford, 1990) which allows learners clearly to indicate which sort of strategies they usually apply when learning a language generally. Results show to what extent they use (and are aware of using) appropriate strategies for remembering more effectively, using mental processes, compensating for missing knowledge, organizing and evaluating their learning, managing emotions, and

learning with others. Subjects also passed a pre- and a post-test on vocabulary (pre- and post-tests were identical) and a post-test to assess communicative competence in the same domain. For the latter one, called the main post-test, we created original aural and textual materials. Subjects had to write their answers and essays. The main post-test had three parts: an aural comprehension of an interview which included subjective appreciation of the applicant's situation; a comprehension and a written production of part of the exchanges in a dialogue on the telephone; and the writing of a letter of application for a post-profile described in an advertisement. This test was not ready when the experiment started, so we could only use it as a post-test.

The two groups appeared not to be equally balanced. Analysis of subjects' answers in G2, the computer-based group, showed that they used more varied strategies and were more self-conscious of the way they usually learned. They proved to have better lexical knowledge than G1 in the pre-test. Both groups progressed in this domain, G1 slightly more than G2. This may not be very surprising since the lexical test was difficult (the emphasis was put upon the relationship between words and phrases, and collocations; semantic relationships, grammatical structures and relational constraints of lexical phrases were required to be understood). Within this context, progression in subjects with lower-level knowledge is easier. As regards the main post-test, there was not much difference between G1 and G2. This result is not easy to explain since samples were limited in both groups. However, we noticed that G1 behaved as if they were competing against G2. The learners in G1 did, however, have to find by themselves extra resources which were easily available in the software: for example, we observed that G1 learners frequently used dictionaries. G1 strongly protested against their learning materials which they found boring, while G2 found much interest in the software. Learning may have been a harder process for G1, but both groups satisfactorily learned, and passed their exams (vocabulary and main post-tests), which was what we were expecting.

Conclusions

General outcomes

Multimedia
The activities which the learners rated most highly were the video-based and the audio-based activities, in order of preference. When asked to evaluate the activities upon quality alone, this order of preference was reversed. When learners found quality of sound unsatisfactory, they expressed their view strongly, although they never complained about the definition or size of the video material. This result supports our original decision to use the basic MPC2 standard, since limitations to the quality of video are less important given the range of functions we assigned to video. In CAMILLE, as in other CALL environments, video is primarily used:

• to put language into context, thus to raise motivation;

• to support the interpretation of the linguistic contents of utterances: in simulation activities, looking at the speaker's face may bring information on the pragmatic contents of the message (happiness, irony, discontent, etc.), and when pronunciation activities are essential, as in the first lessons of the Dutch course, focusing on the speaker's lip movements facilitates comprehension and production of phonemes.

For such functions, the video supports sound. This means that when use of video is suppressed (for example, in some telephone-based activities where we wanted to increase the level of difficulty), the linguistic content is still comprehensible, provided that the quality of the sound is very good.

However, learning a language is not reducible to purely linguistic knowledge. Kinesics and proxemics are also very important (Feldman and Rivé, 1991). In real communication settings, the hearer not only interprets the speaker's message from its linguistic content, but also from his/her gestures, location, etc. In situations such as interviews or negotiations, the issue not only relies on what is said but also fundamentally on the predisposition of the various parties – a predisposition which will be interpreted according to a protocol of behaviour and gestures. In foreign-language learning, these aspects are never neglected in live courses. If we want to do as well in CALL, we need to study other functional uses of video. In one module of the French for Business courseware, we have designed three activities on gestures, which can either support the verbal message or completely replace it. However, no experiments have yet been made concerning this new type of activity because its development was completed only after the evaluation phase.

As regards sound, the results of our experiments showed that we had underestimated the potential of simple technologies which allow recording and producing sounds of high quality. In CAMILLE, it is possible for the learner to record him/herself in almost every activity. In some of them, self-recording is an accessory, but in others (like simulation activities) it is essential. Experiments showed that even if all learners regarded it as important to have self-recording facilities, there was a large discrepancy between the way they claimed to use these resources and the extent to which they actually did. This can be explained by learners' lack of self-assurance, and by the lack of explicit stress in the first versions of our software on this important and preliminary step for the support of oral production skills. We have now switched to simple solutions such as adding signposts and interactive comments in relevant activities, and in the general learners' follow up. In fact, the CAL environment must indicate to every user the importance of adopting effective, interactive strategies, such as re-recording oneself several times and making a (subjective) comparison with the model (as we observed some learners actually doing).

Learner variability

In all the questionnaires, learners almost unanimously expressed their preference for interactive activities compared to more passive ones, but they disagreed about which ones they considered better (with the exception of simulations, which were always highly rated). Learners also often stressed the fact that even if they found communicative activities attractive, basic linguistic activities, on grammar or vocabulary, should not be forgotten. In the case of *Interactif Nederlands*, for example, this led to an adjustment of the balance between both types of activities by adding new, more linguistically oriented, tasks. The learners' reaction was not necessarily a plea for activities of a 'traditional' nature. Linguistic activities can be designed in new ways. Thus learners found our presentation of vocabulary knowledge in lexical networks in *Travailler en France* very appealing.

Learner variability appeared not only in opinions but also in ways of working with the

courseware. Learners followed very different routes in the scheduling of their overall work: some undertook activities strictly in the order suggested by our presentation; others took a quick overview of the whole contents and of the various kinds of activities (which were signposted), then started with the ones they preferred. Learners also performed activities in very different ways, some trying to finish them quickly without paying much attention to instructions or without looking at the associated resources (they generally then got stuck and had to restart), others self-monitoring their task by first carefully considering in which order to proceed, looking at the cues and available resources. Some were systematic, relying on repetitions of self-recording and exhausting the various possible alternatives. Some systematically took notes before actually performing any activity. Some verbalized their thoughts and reactions, whereas others were almost completely silent. When group-work occurred, and when skills were complementary inside the group, effective collaboration took place with one taking over the interaction with the system, while a second controlled the planning, or negotiated the knowledge.

This learner variability is an important positive outcome. Disagreement on attractiveness of activities showed that everyone found their own interest. Variations in the way of using the software happened according to learners' personal characteristics. Whatever our wishes may be in expecting learners to follow a particular route, individual variability remains the rule in language learning (Ellis, 1994). One of the advantages of multimedia learning environments is the support they can give to these individual variations by offering different types of activity, practice of different linguistic skills, flexible navigation, access to resources of various kinds, and note-taking.

Autonomy

CAMILLE has been designed for an audience of learners who are typically clients, professionals with clear demands and for whom flexibility and swiftness are essential criteria. Sample populations involved in our evaluations mostly corresponded to this profile. Furthermore, nearly all were experienced learners, either advanced learners of a second language, or beginners in a third. They were aware of their own preferred learning strategies, and used software in an autonomous way, evaluators and teachers, when present, being merely observers.

From the learners' answers, and from our own observations, it is possible to underline the points which follow.

- When we developed our software, we recognized the need to distinguish between activities and resources, but also the need for resources to be tightly linked to activities in order to make essential extra knowledge readily available within self-contained courses (Chanier, 1996). The fact that learners did use these extra resources suggests that significant time and energy should be allocated to their development in such hypermedia environments.

- Software can be self-contained, but learners will still be looking for discussion and feedback with experts. It is still an open question whether these experts should be teachers acting as guides or counsellors, or native speakers.

- Self-access has been, as far as possible, the rule. Learners have made it a basic requirement. Insufficient provision of equipment and flexible access time within institutions may jeopardize the whole learning procedure.

- Autonomous learning situations have been explored only in training institutions. Learners said they were willing to work alone, and to work at home. We have yet to investigate how this might affect learning outcomes. Experimenting with access at work is yet another possible approach that should be considered.

Not surprisingly, the types of learners with whom we were concerned reacted very positively. They appeared to master the essential three domains for managing one's own learning (Holec, 1990): methodological aspects, linguistic aspects and cultural background. We have collected no data for generalizing these outcomes to other types of learners. The experiment undertaken in Teesside with true second-language beginners, lacking the self-assurance and motivation, was not conclusive. Blin (in press) has also remarked that an insufficient level of confidence in using computers for language-learning purposes (which never appeared to be a problem with our experienced learners) may represent a major element in the learner's decisions to under-use computers as opposed to other materials in self-access centres.

Effectiveness and language skills

It is now time to come back to the question of the effectiveness of such hypermedia software with respect to the four language skills. As pointed out above, the technology we relied on is more adequate for practising aural (listening) and written comprehension than aural and written productions. Learners passed two summative tests, as described earlier, in HEBO and in the CAVILAM. The former test was completely based on aural skills and thus included aural production. The latter test encompassed aural comprehension, written comprehension and written production.

In order to appreciate these results correctly, it should be remembered that the evaluations involved small samples, related to specific types of learners, and in both places quality of results was not much better than that of more traditional approaches, live courses or audio-cassette methods. This quality is satisfactory because we were not expecting computing-learning environments to be much more efficient, but to represent an effective alternative which can be taken into account in autonomous learning situations, an alternative which possesses other advantages discussed in the previous sections. Another open question is whether or not our results can be generalized to all experienced learners.

Evaluation as part of the development process

The main goal of formative evaluation is to measure the performance of the courseware. It is a necessity for adapting the software, for debugging it, and for collecting essential information on timing etc., information which can then also help in preparing the user-manual delivered with the software. The procedure must involve real learners belonging to the target audience, and should be set up long before the end of the development. The elapsed time between the final release of the software and the evaluation phase is often as long as the duration of the development of the first version of the software which served in the evaluation. In general, a reduced protocol is sufficient, but if research questions are at stake, an extended protocol is necessary for setting up case studies. The whole evaluation procedure then becomes much more complex.

The purpose of summative evaluation is to measure the effectiveness of the courseware in terms of its stated aims. We have pointed out several caveats: the summative evaluation is

time-consuming; it requires adequate means for achieving it; many partners are often involved; and its results or its abandonment may be used against the project. Since it represents a real risk for the whole project, the first question which should be answered before making any decision is: who are the audience? Who really wants to know the outcomes?

Nevertheless, the organization of summative evaluations by project teams should happen more frequently. They are important for the research and pedagogical communities for deontological reasons, as follows.

• They help to clarify the functional differences between the various sorts of software reports and the evaluation reports. For most software, the only accessible reports are commercial reports, written by publishers, or software reviews, written by external teachers or researchers. These reports may bring useful information, but they have the disadvantage of often being labelled as 'evaluations'. Confusion with reports based on experiments involving real learners and following a methodological procedure should be avoided.

• They minimize over-generalizations, either pro or con. An evaluation has specific aims. Results can be interpreted only with respect to the restricted parameters which have been tested. Unfortunately, papers are still published which either present evaluations as being aimed at definitely stating the superiority of CAL over other learning methods, without defining parameters such as types of learners, types of skills, levels of proficiency, nor actual learning situations, or which, when they make explicit their restricted purposes, do not incorporate any detailed information. It is then impossible for readers correctly to interpret their results, and to undertake other evaluations in order to verify them.

• They reinforce the idea that, in an evaluation, not only the software may be tested, but also the learning situations. A limited piece of software can be very useful, and on the contrary, a wonderful language package can be misused, depending on its integration into the curriculum, its access conditions, hardware configurations, etc.

• They may offer instruments for measuring various aspects of so-called communicative competence. Such references support the dialogue between designers and the Second Language Acquisition community.

Potentially, a summative evaluation is also of direct interest for the project team itself. It represents an efficient way of clarifying the final aims of the software, and of estimating the inevitable shift between the initial hypotheses and the reality of the achievements. Measuring tools make it possible to elicit how and on what grounds designers want their innovation to be estimated. Since evaluation is a cumulative process, it forms a starting point from which other researchers are able to set up new experiments in order to extend the initial measures. Tests can also be adequately joined to the software delivery in order to let learners evaluate themselves at the end of their training period.

Note
1 CAMILLE (which stands for Computer-Aided Multimedia Interactive Language Learning Environment) has partly been financed by the European LINGUA Programme. Members of the CAMILLE Consortium are: The University of Teesside; Université Blaise Pascal and Université d'Auvergne, Clermont-Ferrand; De Haagse Hogeschool, The Hague; and Universidad Politécnica, Valencia.

Acknowledgements

References

Blin, F. (in press), 'Integrating CALL in the negotiated learner-centred curriculum: a case study', Eurocall '95 Conference, Valencia, Spain, September, 1995.

Chanier, T. (1996), 'Learning a second language for specific purposes within a hypermedia framework', *Computer-Assisted Language Learning*, 9 (1), 3–43.

de Landsheere, V. and G. (1984), *Définir les objectifs de l'éducation*, Paris: PUF.

Demaizière, F. and Dubuisson, C. (1992), *De l'EAO aux NTF: utiliser l'ordinateur pour la formation*, Paris: Ophrys.

Ellis, R. (1994), *The Study of Second Language Acquisition*, Oxford: OUP.

Emery, C., Ingraham, B., Chanier, T. and Brouwer, J. (in press), 'Creating interactive multimedia CALLware: the CAMILLE experience', Eurocall '95 conference, Valencia, Spain, September, 1995.

Ehrman, M. and Oxford, R. (1990), 'Adult language-learning styles and strategies in an intensive training setting', *Modern Language Journal*, 74 (3), 311–27.

Feldman, R. S. and Rimé, B. (1991), *Fundamentals of Non-verbal Behaviour*, Cambridge and Paris: CUP and Editions de la Maison des Sciences de l'Homme.

Holec, H. (1991), 'Autonomie et apprentissage auto-dirigé: quelques sujets de reflexion', in *Les Auto-apprentissages, Actes des 6èmes Rencontres de l'ASDIFLE*, Paris: Les Cahiers de l'ASDIFLE, 2, pp. 23–33.

Ingraham, B., Chanier, T. and Emery, C. (1994), 'CAMILLE: a European project to develop language training for different purposes, in various languages on a common hypermedia framework', *Computers and Education*, 23 (1/2), 107–15.

Knussen, C., Tanner, G. R. and Kibby, M. R. (1991), 'An approach to the evaluation of hypermedia', *Computers and Education*, 17 (1), 13–24.

Lussier, D. (1992), *Valuer les apprentisages dans une approche communicative*, Paris: Hachette.

Mark, M. A. and Greer, J. E. (1993). 'Evaluation methodologies for intelligent tutoring systems', *Journal of Artificial Intelligence in Education*, 4 (2/3), 129–53.

Nunan, D. (1992), *Research Methods in Language Learning*, Cambridge: CUP.

O'Malley, J. M. and Chamot, A. U. (1990), *Learning Strategies in Second Language Acquisition*, Cambridge: CUP.

Pothier, M. (in press), 'Travailler en France: un environnement informatique hypermedia pour l'auto-apprentissage sur objectifs spécifiques', *Revue de Phonétique Appliquée*.