# Evaluating foreign language tutoring systems

Brian MacWhinney
Carnegie Mellon University

Key words: tutoring systems, foreign language instruction, learning theory, psycholinguistics, analogy, transfer, testing, evaluation, practice effects, critical periods, parsing, microworlds, feedback, phonology, human-computer interaction, stages, translation, schema theory, context, pretest-posttest design, order effects, control groups, and Hawthorne effect.

## Abstract

The design of computational systems to support foreign language instruction needs to be grounded on what we know about human learning, language processing, and human-computer interaction. Principles derived from these fields can be tested and quantified in the context of specific tutoring systems. Evaluation of the pedagogical impact of particular principles can best be achieved by comparing two tutoring systems that differ in controlled and manipulable ways.

## Introduction

For adults, the learning of a foreign language is often difficult, sometimes virtually impossible. Systems that promise a way of reducing this "language barrier" will interest millions of people, if they can be demonstrated to really work. It is not surprising that the enormous breakthroughs in personal computer technology over the last dozen years have led to a renewed interest in the development of computational systems for facilitating language learning. However, the development of these computer-assisted language learning (CALL) systems is dependent on many types of expertise and requires a great investment of time and energy. If we had a set of basic design principles that could guide the construction of these systems, we could avoid inadequate products and mismatched capabilities. If one looks at work in learning theory, psycholinguistics, and human-computer interaction, there are many signposts or principles that could direct this work. However, to provide a surer grounding for this effort, we will need to make direct tests of the pedagogical effects of these principles. Before we examine ways of making these tests, let us review some of the lessons that can be derived from previous work in psychology.

## Lessons from Experimental Psychology

On the face of it, one would think that Psychology would have a great deal to tell us about the ways in which human beings learn second languages. Ideally, experimental psychology should provide clear principles for designing computational aids and tutoring systems for foreign language learning. Indeed, Psychology is rich with theories and principles that have some general applicability to this question. However, few of these principles will strike any of us as surprising or new and some of them are stated in such a general way that their application to this problem is insufficiently explicit. Consider the following list of basic ideas from learning theory and experimental psychology:

1. **Practice makes perfect**. One of the most well-established principles of learning theory and cognitive science is that the more time one spends on a task, the better one gets at that task (Anderson, 1990; Ebbinghaus, 1885; Newell, 1990). This is effect is variously known as the "learning curve", the "power law of practice", or the effect of "time on task". We also know that is better to distribute practice across time, rather than massing it into concentrated "cram sessions". There is no doubt that the longer one studies a foreign language, the better one gets at that language -- at least up to a certain point. How much of the overall variance between learners is accounted for by this principle? Perhaps as much as 50% of the variance can be attributed to this single basic principle.
2. **Rewards work better than punishments**. A robust finding of decades of work with reinforcement training of rats and pigeons led workers such as Skinner (1957), Mowrer (1960), and Miller (1963) to conclude that learners prefer not to be shocked and beaten while they are trying to learn. It makes good sense to apply this same principle to foreign language tutoring systems.
3. **Success is rewarding**. A slightly more sophisticated variant of the second point was developed in the theory of intrinsic and extrinsic rewards. If the learning process itself can be perceived as rewarding, progress is even swifter. Further elaborating on this theme, McClelland (1961) noted the role of achievement motivation in learning. Most instructors and instructional systems undoubtedly follow these principles even without making a conscious effort.
4. **Feedback promotes learning**. Working within the framework of a hypothesis-testing theory of learning, Levine (1975) and others have demonstrated the importance of feedback in promoting learning. This principle has important potential consequences for the language classroom and computerized language tutors.
5. **Learning strategies can be learned.** Those of us who have learned two or more foreign languages realize that the process of language learning can itself be learned.

The notion of learning strategies and learning to learn is a familiar one to psychologists (Hayes, 1981) and can be directly applied to the task of foreign language learning (Atkinson, 1975).

6. **Learning is schema-based**. The recent surge of interest in contextualized bases for language learning matches up closely with an ongoing theme from cognitive psychology. This is the theme that bases all long-term learning on the construction of schemas (Bartlett, 1932; Bransford & Franks, 1971). This view sees the learning of new material as involving integration into old material (Potts, 1978) and emphasizes the role pictorial cues in information-integration (Larkin & Simon, 1987).

These principles from experimental psychology contain certain core truths that help us understand the basic nature of learning. They suggest that computational aids for language learning should be designed to be easy to use, intrinsically motivating, and contextually rich. However, few of these ideas speak directly to the specific situation of foreign language instruction or computational aids for language learning.

## Lessons from Developmental Psycholinguistics

If we are to discover some finer-grained ideas about language learning, we need to turn to the study of first language learning or developmental psycholinguistics. Work in this area has yielded an additional set of findings and principles:

1. **Language is learned in context.** Research on language learning has tended to emphasize the role of the social context in facilitating and determining language learning (Ervin-Tripp, 1981; Ninio & Snow, 1988; Ochs & Schieffelin, 1979; Schieffelin & Ochs, 1987; Vygotsky, 1962). An extreme example of the importance of social context can be found in hearing children of deaf parents in isolated social situations who receive minimal language input. Although these children may hear many hours of speech on radio and TV, they will not learn to speak until they have the opportunity to use language in real social contexts. The parallel to the foreign language learner is clear. Children learn language by interacting with people, not passive machines.
2. **Words are the basis of language learning**. Work on the acquisition of grammar (MacWhinney, 1982; MacWhinney, 1988) and the development of early vocabulary has underscored the central role that word learning plays in all of language learning. These findings point to a need to emphasize word learning in syntactic contexts (Gleitman, 1990) in systems for foreign language instruction.
3. **Teaching is effective**. Although there is reason to believe that children learn fairly little from overt correction (Brown & Hanlon, 1970), there is also reason to believe that they learn a great deal from scaffolding (Scollon, 1976), modeling (Bohannon, MacWhinney & Snow, 1990), and recasting (Nelson, 1982).
4. **Errors reflect learning**. A simple, but important finding of language acquisition theory is that errors reflect the overapplication of general patterns (Jespersen, 1922). Some particularly useful examples of this overgeneralization process can be found in connectionist models of language learning (MacWhinney, Leinbach, Taraban & McDonald, 1989; Plunkett & Marchman, 1991). These models stress the extent to which learning is based on generalization, analogy (MacWhinney, 1989), and transfer.
5. **Cue conflicts are crucial for learning**. Work on the learning of both L1 and L2 within the framework of the Competition Model (MacWhinney & Bates, 1989) has shown that higher levels of learning require exposure to large numbers of those relatively rare sentences in which cues conflict. Increased exposure to these conflict cases allows learners to properly tune the weights for conflicting cues.
6. **Auditory form scaffolds the learning of articulatory form.** Recent models of the learning of the articulatory forms of words make use of the notion of an auditory template (Houghton, 1990) that can guide the construction of an articulatory plan. This dynamic view of the relation between word comprehension and articulatory

production can have important consequences for the design of sound-based language tutoring systems, particularly for those that focus on the lexicon and phonology.

7. **Language learning progresses through a fixed series of stages**. All children move through a one-word and a two-word stage (Brown, 1973). Early sentences always lack auxiliary inversion and display uniform negative positioning. Many of these stages are inevitable consequences of the limited productive abilities of the child, but others reflect consistencies in language learning strategies.

## Lessons from SLA Research

These findings from first language acquisition research are bolstered in many ways by parallel findings from second language acquisition (SLA) research. The notion that communication is primary has been supported by studies of second language learning in a variety of real contexts (Broeder, 1991; Klein, 1984; Klein & Perdue, 1989). SLA research also indicates the importance of movement through a series of stages much like those of L1 learning, although there is also evidence transfer from L1 to L2 (Kilborn, 1989; MacWhinney, 1987; McDonald, 1987; Odlin, 1989; Ringbom, 1987), although this transfer is seldom based on direct translation strategies. The other major finding of this research is that, past a certain critical period, learning of a second language becomes progressively more difficult (Johnson & Newport, 1989; Johnson & Newport, 1991), particularly in regards to the learning of phonology (Oyama, 1976).

## The Need for an Interdisciplinary Approach

These ideas from experimental psychology, developmental psychology, and SLA research can provide us with some reasonable guideposts in designing foreign language tutoring systems and other computational aids. However, it would be a mistake to think that these principles are specific enough to fully determine the correct shape of foreign language systems. It is crucial for us to conduct new empirical research that evaluates the role of each of these principles within the context of foreign language tutoring systems. Most importantly, the process of conducting this evaluation can itself teach us important new facts about language learning.

Designing, engineering, and evaluating effective FL Tutoring Systems is no mean task. At a minimum, this task requires a high level of understanding of computers, programming languages, parser technology, multimedia design, linguistic theory, SLA theory, expert system design, psycholinguistic theory, research methods, and statistical analysis, as well as a through working knowledge of the target languages involved. It is unlikely that any one individual would possess full competence in all of these areas. In addition, one can divide the types of linguistic competence targeted by a particular tutor into the standard areas of phonetics, prosody, lexicon, grammar, and discourse. Different systems may rely relatively more or less on techniques for constructing parsers, microworlds, speech generators, speech recognition, or error analysis programs. Although it is tempting to break up the problem in these various ways, a system designed to imbue a full level of language competence will have to integrate these various diverse components into a general, synchronized approach. In the end, this problem can only be successfully addressed by an interaction between linguists, psycholinguists, language teachers, ontologists, statisticians, and students of human-computer interaction.

## Basic Questions to the Answered

We are standing at the very beginning of a challenging and fascinating journey. To guide us in the construction of new systems, there are some very basic empirical questions for which we will have to get solid answers. Previous research in psychology will have to be repeated within this new applied context. Up to this point, system design has proceeded on the basis of a series of hunches and guesses. For us to put foreign language tutor

design on a firmer basis, we will need to have real tests of these hunches. Some of the issues that we will need to investigate include the following:

1. **Error diagnosis**. We need to know whether the various error diagnosis or feedback features built into systems such as Herr Kommissar (DeSmedt, this volume) or CALLE (this volume) are pedagogically effective. It makes good sense to imagine that feedback is useful, but we need specific tests.
2. **Parser technology**. Several major projects (Evans and Levin, this volume; Weinberg, this volume) place a strong emphasis on parser technology as the core of a foreign language tutor. But there are many ways in which one can deliver foreign language instruction without relying on a full parser. How much does the introduction of a parser improve instruction? Are there practical and attractive alternatives that do not rely on the construction of parsing systems. For example, instead of constructing a complete parsing system to analyze a learner's sentence such as "*Die Schloss Eisenbach stehen in die Stadt Eisenbach", we could just give the right answer "Das Schloss Eisenbach steht in der Stadt Eisenbach." Or perhaps learners can provide the error diagnosis themselves. Indeed, it might be the case that learning is *more* effective when the learner provides a self-diagnosis.
3. **Microworlds**. Several projects (Yazdani, this volume; Hamburger, this volume; Tomlin, this volume) assume that microworlds provide an effective framework for contextualized language learning. But how effective are microworlds in facilitating acquisition of basic skills such as lexicon, morphology, and syntax?
4. **Learner models**. Is it important for an intelligent tutoring system (Singley & Anderson, 1989) to have a model of the learner? Does an intelligent system also need a model of the world? Is there any evidence that these models can be used to facilitate learning?
5. **Naturalness**. Current approaches to foreign language tutoring systems make the default assumption that systems that most closely resemble the "real" or "natural" language learning context will be most effective. Is there any evidence that this is true? For example, the stated goal of Herr Kommissar is "a system which supports meaningful and engaging dialogue with the second-language learner, focused on tasks and interaction which make the exercise and improvement of communicative skills a means, rather than an end in itself." This goal seems entirely reasonable. However, we do not know whether systems designed with such goals in mind are more effective than systems that do not have these goals.
6. **Discourse context**. Frederiksen et al. (this volume) assume that the learning of a foreign language can be strongly facilitated by embedding the learner inside a full rhetorical or narrative context. Is there direct evidence for positive effects of full rhetorical context? What kinds of tests could we use to examine this hypothesis? Do learners need to receive specific instruction regarding the shape of these discourse structures or can tutoring systems simply use these structures as supports for the acquisition of lower-level knowledge?
7. **Exploration and interface**. Do students learn better if they are allowed to explore a language through a Hypertext-type interface? Do they make effective use of help systems or on-line grammars? Can they use a dictionary to explore the conceptual structure of the vocabulary (Swartz, this volume).
8. **L1 and transfer**. Should L2 systems be designed differently for learners with different first languages? Can the first language be used to promote learning of the second language? Should we encourage or discourage transfer and how?

To some of us, the answers to some of these questions may seem to require nothing more than common sense. However, one of the most important lessons we have learned in psychological research is that common sense makes poor empirical predictions. The only way to properly evaluate these various common-sense-based hunches is by detailed evaluation of the instructional effectiveness of the principles being proposed.

## Evaluation

The easiest way to evaluate a computational system is to ask students if they like it. If they do, the system receives a high "smile coefficient". Unfortunately, the correlation between evaluations based on the "smile coefficient" and objective evaluations of the effectiveness of educational systems is not very high. Students might enjoy a particular computer program because it includes a challenging game or nice sound and graphics, not because it does a good job of teaching Arabic or German. It is a good idea to compute a system's smile coefficient, but we shouldn't confuse this coefficient with real measured effectiveness.

Another level of evaluation for computer systems is that conducted by research in human-computer interaction. If a system is full of bugs, design flaws, and cumbersome options, one can often see easy ways of fixing the system that are sure to improve its effectiveness. If a system is too costly or too low in portability, one can often ways of addressing these limitations. However, once these obvious flaws are fixed, a further evaluation from the viewpoint of human-computer interaction will require the use of a serious experimental design.

The standard design for measuring educational effectiveness is the pretest-posttest comparison of an experimental and a control group. For example, one can randomly assign students to either an experimental group or a control group. The experimental group works with a language tutoring system and the control group spends time reading material in a foreign language and discussing it with a peer. At the beginning of the treatment, all students take a pretest. After several weeks in the two opposing conditions, all students take a posttest that has items similar in type to those on the pretest. The group that has the largest positive differences on posttest scores minus pretest scores is declared "the winner" and the educational treatment given to that group is considered the better educational treatment.

Applying this conventional design to the study of foreign language tutoring systems raises a lot of problems:

1. What is the correct control group? It is not clear that one really wants to compare a language tutor with a read-and-discuss method. Perhaps this is a comparison of apples and oranges. Perhaps the best comparison is one in which some students stay in conventional language classrooms and others do not. But is this a fair comparison if computer-assisted instruction is considered enrichment, rather than part of the core curriculum? And how can the assignment of students to groups be handled in practical terms?
2. Is there a process-neutral posttest? If either the pretest or the posttest are biased toward inclusion of items from either the textbook or the computer system, the results of the evaluation will be biased. There is also the danger that either or both of the instructional formats will start to "teach to the test".
3. If the group receiving computational instruction perceives that they have been singled out for special treatment, they may perform better simply in accord with the "Hawthorne effect".
4. If the computational system has bugs or design flaws, the basic effectiveness of the method may be underestimated.

Fortunately, there are clear solutions that can be offered to address these problems.

1. Rather than comparing apples and oranges, one can compare apples and apples. In particular, one can compare one foreign language tutoring system with another. The closer the two systems are, the more illuminating positive results will be.
2. The problem of bias in test construction can be minimized by orienting half of the items on both the pretest and posttest toward system 1 and half toward system 2.
3. If both treatments involve computational systems, the Hawthorne effect is minimized, since both groups will believe that they have been singled out for special treatment.
4. If each system contains internal ways of checking for design flaws, it will be possible to separate out bugs from the real effects of educational designs.

One can push the notion of a direct comparison between two different tutoring systems even further. Instead of comparing two separate groups in terms of the conventional pretest-posttest design, one can construct a series of variants of individual exercises or components of a larger tutoring system. For each component, the experimenter can design the program so students are randomly assigned to one of two or more conditions right when they log on to the computer. In this way, students will be varied across tests in a Latin-square type of design. Some of the types of modules that can be varied in this way include: multiple choice exercises, pointing to a location for an answer, filling in the blanks, sorting items to categories, short answers, pronunciation tutors, morphology tutors, and so on. Here are a few examples of the types of fine-grained comparisons that can be made by comparing slight variations in individual tutoring systems:

1. In a system such as CALLE (this volume), one can compare a teaching module that focuses primarily on the less frequent, but "core" or "prototypical" meaning of Spanish "se" with a module that focuses on its most frequent meaning.
2. When teaching phonology, one can compare a system that teaches phonology without accompanying orthography with one that includes orthography. Of course, this treatment may have different effects for languages with varying degrees of regularity in their orthographic systems.
3. One can compare a system that promotes transfer by illustrating structures in terms of similar L1 patterns with one that avoids or even blocks transfer.
4. One can evaluate the relative difficulty of teaching different structures by spending a fixed amount of time on Structure 1 in one module and on Structure 2 in another and then looking at how much is learned in the two cases.
5. One can evaluate the value of a Hypertext system by allowing student navigation and option selection in one tutor and blocking it in another.

## Conclusion

True experimental evaluation of foreign language tutoring systems has not yet been attempted. It is difficult to predict at this point how easy it will be to conduct this evaluation. However, the manipulable nature of computer systems makes them ideal test beds for the evaluation of particular micro components of instructional design. Given the great potential commercial market for tutoring systems, it will be important to distinguish between established educational effectiveness, smile coefficients, and simple market acceptance. If these new systems are evaluated carefully, we will learn more not only about computational systems, but also about core processes in language learning.

## References


Anderson, J. (1990). The adaptive character of thought. Hillsdale, NJ: Lawrence Erlbaum Associates.

Atkinson, R. (1975). Mnemotechnics in second-language learning. American Psychologist, 30, 821-828.

Bartlett, F. C. (1932). Remembering: A study in experimental and social psychology. Cambridge: Cambridge University Press.

Bohannon, N., MacWhinney, B., & Snow, C. (1990). No negative evidence revisited: Beyond learnability or who has to prove what to whom. Developmental Psychology, 26, 221-226.

Bransford, J. D., & Franks, J. J. (1971). The abstraction of linguistic ideas. Cognitive Psychology, 2, 331-350.

Broeder, P. (1991). Talking about people: A multiple case study on adult language acquisition. Amsterdam: Swets and Zeitlinger.

Brown, R. (1973). A first language: The early stages. Cambridge, MA: Harvard.

Brown, R., & Hanlon, C. (1970). Derivational complexity and order of acquisition in child speech. In J. R. Hayes (Ed.), Cognition and the development of language, (pp. 11-54). New York: Wiley.

Ebbinghaus, H. (1885). Über das Gedächtnis. Leipzig: Duncker.

Ervin-Tripp, S. (1981). Social process in first and second language learning. In H. Winitz (Ed.), Native language and foreign language acquisition, . New York, N. Y.: The New York Academy of Sciences.

Gleitman, L. (1990). The structural sources of verb meanings. Language Acquisition, 1, 3-55.

Hayes, J. (1981). The complete problem solver. Philadelphia: The Franklin Institute Press.

Houghton, G. (1990). The problem of serial order: A neural network model of sequence learning and recall. In R. Dale, C. Mellish, & M. Zock (Eds.), Current research in natural language generation, (pp. 287-319). London: Academic.

Jespersen, O. (1922). Language: Its nature, development, and origin. London: George Allen and Unwin.

Johnson, J., & Newport, E. (1989). Critical period effects in second language learning: The influence of maturational state on the acquisition of English as a second language. Cognitive Psychology, 21, 60-99.

Johnson, J. S., & Newport, E. L. (1991). Critical period effects on universal properties of language: the status of subjacency in the acquisition of a second language. Cognition, 39, 215-258.

Kilborn, K. (1989). Sentence processing in a second language: The timing of transfer. Language and Speech, 32, 1-23.

Klein, W. (1984). Zweitspracherwerb. Kónigstein: Athenäum.

Klein, W., & Perdue, C. (1989). The learner's problem of arranging words. In B. MacWhinney & E. Bates (Eds.), The crosslinguistic study of sentence processing, (pp. 292-327). New York: Cambridge University Press.

Larkin, J. H., & Simon, H. A. (1987). Why a diagram is (sometimes) worth ten thousand words. Cognitive Science, 11, 65-99.

Levine, M. (1975). A cognitive theory of learning. Hillsdale, NJ: Lawrence Erlbaum Associates.

MacWhinney, B. (1982). Basic syntactic processes. In S. Kuczaj (Ed.), Language acquisition: Vol. 1. Syntax and semantics, (pp. 73-136). Hillsdale, NJ: Lawrence Erlbaum.

MacWhinney, B. (1987). Applying the competition model to bilingualism. Applied Psycholinguistics, 8, 315-327.

MacWhinney, B. (1988). Competition and teachability. In R. Schiefelbusch & M. Rice (Eds.), The teachability of language, (pp. 63-104). New York: Cambridge University Press.
MacWhinney, B. (1989). Competition and lexical categorization. In R. Corrigan, F. Eckman, & M. Noonan (Eds.), Linguistic categorization, (pp. 195-242). New York: Benjamins.
MacWhinney, B., & Bates, E. (Eds.). (1989). The crosslinguistic study of sentence processing. New York: Cambridge University Press.
MacWhinney, B. J., Leinbach, J., Taraban, R., & McDonald, J. L. (1989). Language learning: Cues or rules? Journal of Memory and Language, 28, 255-277.
McClelland, D. C. (1961). The achieving society. New York: Van Nostrand.
McDonald, J. L. (1987). Sentence interpretation in bilingual speakers of English and Dutch. Applied Psycholinguistics, 8, 379-414.
Miller, N. (1963). Some reflections on the law of effect produce a new alternative to drive reduction. In M. R. Jones (Ed.), Nebraska Symposium on Motivation, (pp. 65-107). Lincoln: University of Nebraska Press.
Mowrer, O. (1960). Learning theory and the symbolic processes. New York: Wiley.
Nelson, K. (1982). Experimental gambits in the service of language acquisition theory. In S. Kuczaj (Ed.), Language development: Syntax and Semantics, . Hillsdale, N.J.: Lawrence Erlbaum.
Newell, A. (1990). A unified theory of cognition. Cambridge, MA.: Harvard University Press.
Ninio, A., & Snow, C. (1988). Language acquisition through language use: The functional sources of children's early utterances. In Y. Levy, I. Schlesinger, & M. Braine (Eds.), Categories and processes in language acquisition, (pp. 11-30). Hillsdale, NJ: Lawrence Erlbaum.
Ochs, E., & Schieffelin, B. (Eds.). (1979). Developmental pragmatics. New York: Academic Press.
Odlin, T. (1989). Language transfer: Cross-linguistic influence in language learning. New York: Cambridge University Press.
Oyama, S. (1976). A sensitive period for the acquisition of a nonative phonological system. Journal of Psycholinguistic Research, 5(3), 261-283.
Plunkett, K., & Marchman, V. (1991). U-shaped learning and frequency effects in a multi-layered perceptron: Implications for child language acquisition. Cognition, 38, 43-102.
Potts, G. (1978). The role of inference in memory for real and artificial information. In R. Revlin & R. Mayer (Eds.), Human reasoning, . New York: Halsted Press.
Ringbom, H. (1987). The role of the first language in foreign language learning. Clevedon: Multilingual Matters.
Schieffelin, B., & Ochs, E. (1987). Language acquisition across cultures. New York: Cambridge.
Scollon, R. (1976). Conversations with a one year old: A case study of the developmental foundation of syntax. Honolulu: University Press of Hawaii.
Singley, K., & Anderson, J. (1989). The transfer of cognitive skills. Cambridge, MA: Harvard University Press.
Skinner, B. F. (1957). Verbal behavior. New York: Appleton-Century-Crofts.
Vygotsky, L. (1962). Thought and language. Cambridge: MIT Press.