Giouroglou, H. & Economides, A.A.: **An Implemented Theoretical Framework for a Common European Foreign Language Adaptive Assessment.** *Proceedings ICODL 2005, 3rd International Conference on Open and Distance Learning 'Applications of Pedagogy and Technology'*, Greek Open University, Patra, Greece, 2005.

An Implemented Theoretical Framework for a Common European Foreign Language Adaptive Assessment

Giouroglou, H.
PhD. Candidate, University of Macedonia
Thessaloniki, Greece
hara@uom.gr

Economides, A.
Associate Professor, University of Macedonia
Thessaloniki, Greece
economid@uom.gr

Abstract

The promotion of multilingualism, communication, mobility and cross-cultural awareness among EU member-states has created the demand for easily-administered, self-paced, time-effective, multilingual, and internationally accredited foreign language assessments (FLA). The Common European Framework of Reference (CEF) has paved the way by setting internationally certified standards for formal as well as self-assessment in all European languages and describing in detail the productive and receptive skills needed to attain a specific level of competence. This paper will analyze the rationale for a common European language assessment based on the CEF and Computer Adaptive Testing (CAT). Finally, it will briefly describe the development of a computer adaptive placement test for mixed-ability students that can measure both the breath and depth of foreign language awareness in little time.

Introduction

New Trends in Foreign Language Assessment

One of the milestones of the EU is the maintenance of the linguistic and cultural diversity for every member-state. This resolution implies that instead of linguistic and cultural accumulation, the EU fosters linguistic and cultural dissemination. In order to improve communication and mobility and fight off cultural or racial intolerance among member-states, the EU encourages multilingualism. The realization of this prospect presupposes co-operation and the establishment of common European standards in language education, training and assessment. To this end, the EU and the Council of Europe has issued the Common European Framework of Reference (CEF) which is being applied to all member-states and concerns the adaptation of common language syllabuses, curricula, and examinations across Europe. CEF is based on the communicative, action-oriented and skill-based approach to language learning which is the essence of linguistic competency [Council of Europe, 2001]. Language is a tool for communication, and mental reasoning in a social context and foreign language awareness is the ability of the individual to use different communicative and reasoning "tools" to achieve a goal. Each individual is a different social agent with divergent cognitive skills, empirical or academic knowledge, and social variations. Therefore, new approaches in cross-

European education and assessment should be aimed at mixed-ability students. In a stratified and comprehensive way, CEF describes the linguistic and cultural skills and the knowledge a foreign language (FL) learner needs to possess in order to achieve a desired level of competence in the target language. As such, CEF is applicable both to traditional education and open, life-long or autonomous learning.

In terms of foreign language assessment (FLA), CEF has created six clearly defined proficiency levels (A1,2/B1,2/C1,2) in all European languages for both formal and self-assessment purposes. The levels describe what receptive and productive skills the examinee needs to possess in order to attain the desired level of competence [Council of Europe, 2001,2002]. More specifically, each reference level provides analytical information regarding the quantitative and qualitative competence of the two primary language activities involved in communicative language use (reception and production), and the two secondary language activities (interaction and mediation) in which reception and production overlap. CEF also describes the quantitative and qualitative procedures to ensure assessment validity and standardisation [Council of Europe, 2004]. The CEF standards are internationally accredited and all major FL testing organizations [fig.1] have adapted their examinations to the six common reference levels enhancing international co-operation in FLA.

Fig.1: Adapted from Cambridge ESOL exams linked to CEF.

Mixed-Abilities and FLA

Cross-cultural education has created diverse, mixed-ability students. The one-to-many, teacher-centered tutoring model that was used in traditional education up to date and in traditional education and the previous generation of Computer Aided Instruction (CAI) is no longer applicable to distance education due to students' heterogeneity. Modern learning environments acknowledge the fact that there is no average student model with predetermined behavior and are adapted to students' diverse educational and socio-economic background, age, nationality, motivation and time and place accessibility.

Moreover, the rapidly evolving Information Society demands constant retraining of the global workforce [Twigg, 1994]. Thus, nowadays students of all age groups participate in training and lifelong-learning programmes matching both their occupational and personal needs. In personalized learning and assessment emphasis is given in students' differing ability, interests, motivation, learning needs, and achievement. To this end, modern education needs to provide the new student society with the tools to construct their own knowledge with their own pace, ability, individual learner characteristics and aptitude [Schunk, 1996].

In terms of foreign language assessment (FLA), research in neuropsychology and especially psycholinguistics, that examine the inner processes of the human mind that lead to linguistic proficiency and language acquisition, has revealed that individuals process language differently, according to their overall intelligence, brain dominance, sex, inherent traits and cognitive skills [Akmajian, et al. 1998]. There is an apparent relationship between language, thought and cognition as Chomsky and Piaget have advocated from different points of view [Chomsky, 1997]. Accordingly, educationalists acknowledge the fact that there are mixed-intelligence students, meaning that learners can attain new knowledge using different learning strategies and paths, suited to their individual intelligence [Gardner, 1993, Armstrong, 1999]. Finally, research in first language acquisition has revealed serious findings in human language development through the study of certain phenomena, such as hesitations, speech errors and language disorders that can also be applied in second/foreign language acquisition.

Cross-cultural FL learners in particular also have another important differentiation. They do not have a common first language or mother tongue. Thus, they do not have the same experiences and cognition. Apart from that, during FLA, multi-lingual examinees might have different achievement, not due to FL ignorance but due to misunderstanding. As such, FLA environments should also be multilingual so that each examinee's adequate comprehension of the items and tasks is established.

In order to foster learners' success, we need to adapt FLA environments to accommodate learners' diversity accordingly. Any assessment in foreign language that does not adapt to the aforementioned mixed student abilities cannot be considered reliable and valid. Mixed abilities create mixed needs which result in mixed implementations in all educational settings.

Traditional Assessment versus CAT in FLA (CALT)

CAT provides personalized testing and more accurate results for every individual examinee. Test items are categorized in terms of levels of difficulty. The test starts with an item of average difficulty that corresponds to the level of the average student. Based on an algorithm, the computer can update the estimate of the examinee's ability after each item and select the next item on the basis of the new ability estimate. On the basis of the examinee's previous answer, the system selects the next item which is of greater or less difficulty in accordance to the examinee's previous response. The test proceeds in the same pattern, until the stopping parameter comes. The test score derives from the average level of difficulty of the items answered correctly.

The major advantage of CAT systems is that they are student-centered as, in contrast to their paper-and-pencil (P&P) counterpart, they can be tailored to the ability and level of each examinee by updating the estimate of the examinee's ability, called User Profile, and adapting the subsequent items to the individual ability of each examinee. Item adaptation results in reduced standard errors and improved accuracy of scores for both high and low ability test takers. Tailored item selection also leads in avoidance of examinees' boredom from answering too easy questions and of frustration from answering too difficult questions. Thus, CAT is said to have increased efficiency, greater precision with less items, and time-effectiveness, since only a few tailored items are needed to achieve accuracy. CAT systems offer also greater test security and longer duration [Wainer et al, 2000] than traditional P&P tests, as they are comprised by large item pools with controlled item exposure, rendering examinees incapable of knowing the items in advance. CAT shares all advantages of CBT, such as immediate feedback and self-pacing. Many problems associated with P&P tests such as ambiguous answers or physical problems with the answer sheets [Wainer et al, 2000] are being solved. Web-based CAT exploits the capacities of the net, such as on-line, immediate scoring, easily downloadable software, and low cost software and item pool update.

Yet, CAT is not applicable to all subjects and skills, as it is based on the Item Response Theory model (IRT), which is not applicable to all item types. To achieve accuracy, IRT requires careful item calibration, excluding items that cannot be easily calibrated, such as open-ended questions [Lord, 1980]. Another crucial drawback is that the examinees are not permitted to go back and change answers, as the program selects next item on the basis of the previously answered item(s). This renders reviewing implausible, and in many cases examinees that sat both P&P and CAT failed or achieved low marks in computerized testing. Studies show that only when both P&P and Computer Based Testing (CBT) had the same test-taking flexibility, test results were equivalent [Sawaki, 2001]. CAT philosophy, however, prohibits reviewing.

Computer Adaptive Language Testing (CALT) uses adaptive technologies to assess foreign language (FL) competence. Most international FL testing organizations have started delivering their tests in self-paced CBT and CAT mode, making their tests available to even more people. CAT successfully assesses multiple-choice (MC) items in vocabulary, grammar, reading and listening, using IRT. One way to assess vocabulary is by categorizing words into levels of competence in order to assess the size of vocabulary each examinee has. Another way to assess vocabulary is by measuring the strength of vocabulary knowledge. Studies in this field have shown that adaptive tests measuring vocabulary knowledge in terms of size and strength have managed to assess examinee's level of vocabulary knowledge accurately [Laufer, et al, 2001]. Reading is a receptive skill and can be easily assessed in multiple-choice form that can be easily computed in an adaptive algorithm. The important issues in the development of adaptive reading items regard the reading construct validity, the IRT theory used and the measurement of the items. Adaptive listening items assess examinees' ability to understand a range of oral speech, from short utterances, such as single words to short monologues and dialogues and to longer discussions [Dunkel, 1997]. Up-to-date, CALT systems do not have adaptive components in oral proficiency and writing tasks. Open writing tasks can be marked by electronic marking with the aid of human markers. The Intelligent Essay Assessor (IEA) is a commercial grading software financed by the Army Research Institute and developed at the Knowledge Analysis Technologies. Research has shown that it can assess specific topic essays as accurately as a human examiner [Streeter et al, 2002]. However, scores of essays written by non-native speakers of English had a slightly bigger variation between the e-rater of GMAT and human readers, showing that there are some non-native syntactic and semantic structures not evaluated by the e-rater [Burstein, 1996, Burstein, and Chodorow, 1999]. Though promising, e-rater cannot still be used without the surveillance of a human reader, as studies have proven that it can be fooled by experts in writing [Powers, 2001].

To sum up, CAT systems have both merits and flaws, and they cannot specialize on every plausible item. Although IRT increases the validity and reliability of the test, it lacks the flexibility to cover a wide range of activities and abilities, including open answers, and productive language use.

The Problem

Modern FLA implementations fail to cater for mixed-ability students, as they are linear and targeted to the average student. Computer Adaptive Language Testing (CALT) technology can provide student-centered assessment, replacing traditional testing wherever possible.

However, CALT nowadays is based on solid programming that is collective rather than individualized and fails to include crucial cognitive parameters of student language competence and performance [Giouroglou and Economides, 2003]. Such systems cannot replace the human examiner without nasty consequences for its group of examinees. The new generation of assessment systems for cross-cultural examinees should not assess students horizontically as an equable lot but vertically as mixed-ability individuals with mixed-scoring options [fig. 2]. Moreover, the new generation of assessment should create different experiences that will motivate test-takers, as it is proven that the new technologies are profoundly preferable to students, whenever they make the lesson interesting, motivating and interactive [Ali, 2001].

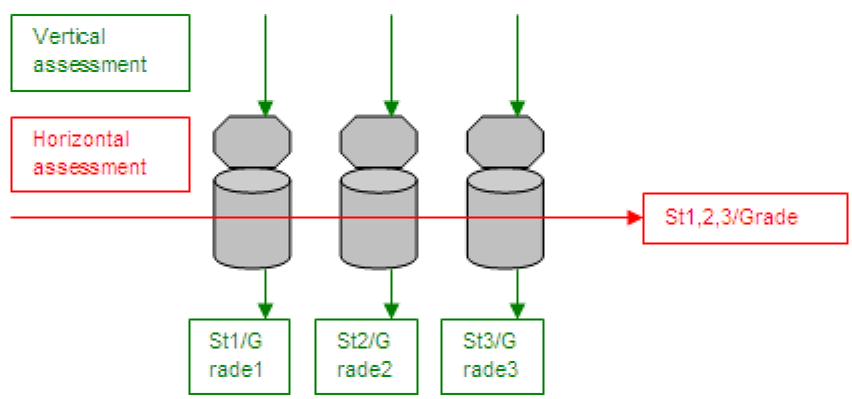

Fig. 2: Vertical versus horizontal assessment

As described above, the majority of CALT systems use MC, close-ended items to discriminate among proficient, good and weak learners. This is mainly due to the fact that MC items are easily programmed and calibrated in IRT. The program can easily identify correct and wrong answers and move on to easier or more difficult items. This technique is also reliable and valid as long as items are adequately pre-tested and correctly calibrated. However, MC items cannot allow active expression and language production. Examinees are passive viewers of the proposed answers and they only try to segregate the correct answer out of the distracters. This method is widely used by language testing organizations, such as the University of Michigan Certificates in English, while other organizations use a variety of MC and open-closed items, such as the Cambridge Syndicate and the State Examinations on Language Competence (KPG). Proficient learners answering MC items are not given the opportunity to discriminate themselves from good learners by openly typing the correct answer in case they know it. They are forced to choose among the four intended choices and receive the same mark as other learners who will purposefully or accidentally choose the correct item. This limitation does not allow the proficient learner discern from others by testifying active language production. Another problem is caused by the prohibition of item reviewing. In psycholinguistics there is a clear discrimination between errors, made due to ignorance, and mistakes, made due to negligence. Examinees are prone to mistakes not only out of ignorance but also out of misunderstanding, anxiety, confusion, distraction or other physical reasons. Since reviewing is impossible, adaptive systems may form false impressions and give low scores. To this end, CALT should become more "intelligent" and simulate the human examiner in order to be more accurate and precise in their scores.

CALT Meets CEF

CAT is a research field that needs to evolve, as it is necessary due to the demand for life-long and open distance learning. Especially in FLA, CAT plays an essential role, providing wide audiences with easy, self-paced, time-effective, and individualized assessment, thus, enhancing multilingualism across Europe. Open Distance Learning (ODL), web-based learning and life-long learning can help Europe evolve into a "Knowledge Society" by simultaneously reaching its citizens regardless of time or place constraints. Nowadays, education is not only for the privileged, but for all European citizens of all age-groups, occupations, or background. This means that on the one hand there is abundance of students and on the other hand shortage of educators. This gap needs to be bridged by educational technology that needs to be evolved in order to simulate the human educator, by providing individualized learning and assessment.

In terms of FLA, CALT should incorporate the CEF standards in order to develop internationally accredited, valid and reliable assessments. FLA needs to adapt to individual student needs, abilities, backgrounds, strengths and weaknesses, giving emphasis to cognitive language skills, such as comprehension, production and use.

A CEF Placement CAT: AILA

The Rationale

As explained above, there is a perceived need for a new generation of FL tests, which should be adaptive and adaptable in nature, catering for diverse, mixed-ability students. Students' diverse needs and abilities pose the necessity for the development of flexible assessments that will suit the diverse student's cognitive skills. The Adaptive Item Language Assessment (AILA), developed by CONTA Lab at the University of Macedonia, is an adaptive placement test, based on CEF standards, that is both adaptive and adaptable in that examinees are given the choice to select how to answer each item presented. Examinees can choose between two options, the first in MC and the second in open-answer (OA) mode. The system adopts course content tailored to the student's needs, taking into account different difficulty levels as well as different knowledge levels.

System Architecture

It is important to create a system that is affordable, and easily maintained. To achieve this, it is required to create re-usable objects of previous existing educational content. The recommendation is a CPU Pentium III 800 MHz, with 2 GB RAM, and either the Apache or IIS web server. The software can run on Windows NT 4.0, Windows 2000 Server or Advanced Server. The system software includes the required MySQL database software.

For reasons of re-usability XML has been used to separate content from the way it is processed (i.e. presented) and which avoids to re-write the same content that needs to be displayed in different formats. The software used is Windows 2000, My SQL (free), PHP, VB script, Javascript, HTML, XML. The system has a modular, component-based architecture that makes it easy to create the adaptive testing system and to re-use data from different learning levels. It is an independent platform and avoids vendor lock-in.

System Description

AILA [fig. 3] measures the ability of non native speakers of English to use and understand English as a foreign Language for achievement and placement purposes. The test-takers who sit AILA can quickly assess their competence in English in the scale issued by the Common European Framework of Reference (CEF) for foreign language assessment. The test measures competence in four out of the six CEF levels, A2, B1, B2 and C1 each of them consisting of three item types:

- Grammar and Structure (Use of English) items. They measure the ability to recognize and/or produce grammatically English Language structures that are appropriate for each CEF level.
- Vocabulary (Use of English) items. They measure the ability to recognize and/or produce high or low frequency English words that are appropriate for each CEF level.
- Reading items. They measures the ability to understand and extract information from short passages that are appropriate for each CEF level.

Fig. 3: AILA User interface

AILA is a proficiency assessment [fig.4], in that it assesses what a test-taker does or knows in a context. It is criterion-referenced, in that it assesses the learner's ability and skills in relevant domains irrespective of the abilities of his/her peers. It also follows the continuun CR approach, as the individual ability is referenced to a defined continuum of all relevant degrees of ability in the area in question. It is a fixed-point assessment, in that grades are awarded according to the measured competence at a given time. It adopts some of the characteristics of formative assessment, as it gathers information on the extent and quality of the examinee's learning, which can be given as feedback. It provides indirect assessment of both receptive and productive skills. Is is a knowledge assessment, as the learner needs to answer questions of different item types in order to provide evidence of the extent of their linguistic knowledge. It provides objective assessment in that there is always one right answer; however, it allows subjectivity for mixed-ability students, as the system "recognizes" incorrect but very close to the correct answers as "acceptable" as well as common "slip of the key" errors. AILA also provides rating on a scale at a particular level, based on CEF. It uses guided judgement in that it sets defined criteria to distinguish between levels of competence. AILA uses a holistic rating strategy with three analytic scales of criteria (i.e. grids), receiving different grades: a correct "productive" answer, a correct "receptive" answer and a incorrect but acceptable answer. Finally, AILA can be used for self-assessment purposes by competent learners who will be able to recognize their strengths and weaknesses and improve their performance.

1	Achievement assessment	Proficiency assessment
2	Norm-referencing (NR)	Criterion-referencing (CR)
3	Mastery learning CR	Continuum CR
4	Continuous assessment	Fixed assessment points
5	Formative assessment	Summative assessment
6	Direct assessment	Indirect assessment
7	Performance assessment	Knowledge assessment
8	Subjective assessment	Objective assessment
9	Checklist rating	Performance rating
10	Impression	Guided judgement
11	Holistic assessment	Analytic assessment
12	Series assessment	Category assessment
13	Assessment by others	Self-assessment

Fig. 4: FL assessments, adapted from CEF

Content Adaptation

AILA measures English language proficiency in use of English and reading and is computer based, using adaptive technology in item selection. The system increases student motivation, by providing tailored content adapted to his/her needs and level of competence. This will also result to a reduction of time spent with maximum benefit. To create an open and flexible testing environment, the system should be focused on the student; the student shall define his/her educational goals by choosing his knowledge level according to which the information will be displayed in a predefined way.

As the test is administered, examinees are given the freedom to choose between two options [Chart 1]. They can either type their answer (OA) or choose the correct answer in MC mode. If the examinee knows the answer and is able to produce it, then he/she can type the answer in the OA mode and has the opportunity to demonstrate his/her advanced knowledge. A correct OA response receives a bonus in the total score (grade+ 0.25) and updates the User Profile of the examinee. Then, the item selection algorithm proceeds to the next item of increased difficulty. A wrong OA response does not affect the final score and it immediately directs the examinee to the MC mode of the same item. When the MC mode appears, the examinee cannot go back to the OA mode.

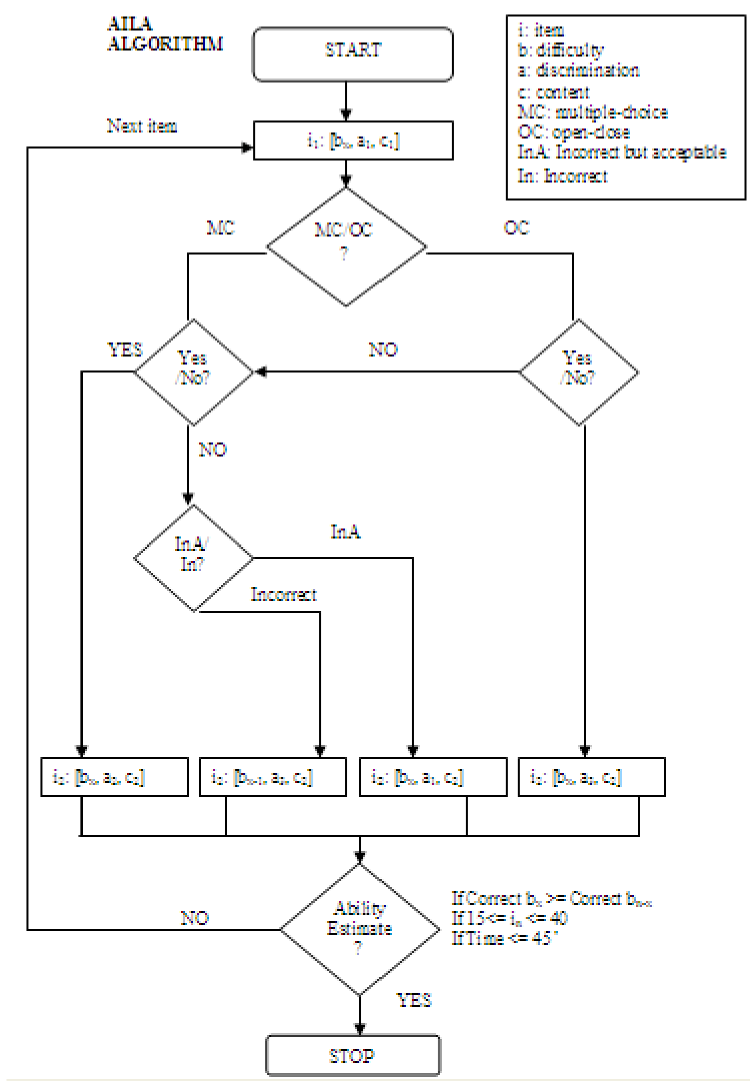

Chart 1: AILA flowchart

The immediate selection of the MC mode does not have a negative effect on the score, as correct MC choices receive the highest mark (1), and the adaptive algorithm immediately proceeds to the next, increased difficulty item. Wrong choices receive no mark and the next item is easier.

This method does not affect the final score of the test or punish a wrong OA answer. Instead, it gives the opportunity to the examinees to demonstrate productive FL use and active FL extraction from their long-term memory.

In order to cope with students' divergent cognitive strengths and weaknesses, AILA tries to discern between errors and mistakes, using a simple method. The MC option that bears a close resemblance to the correct option is be regarded "acceptable" and the item selection algorithm proceeds to an item of equal difficulty. The reason for doing this is the fact that in most MC questions at least one destructor is so close in meaning or in grammatical resemblance to the correct answer that may sometimes puzzle even examiners. Bearing in mind the fact that language is a flexible, ever-changing, living entity used to communicate meaning and retrieve information, we should create CATs that will accept answers that have a slight deviation from the standard form. It is

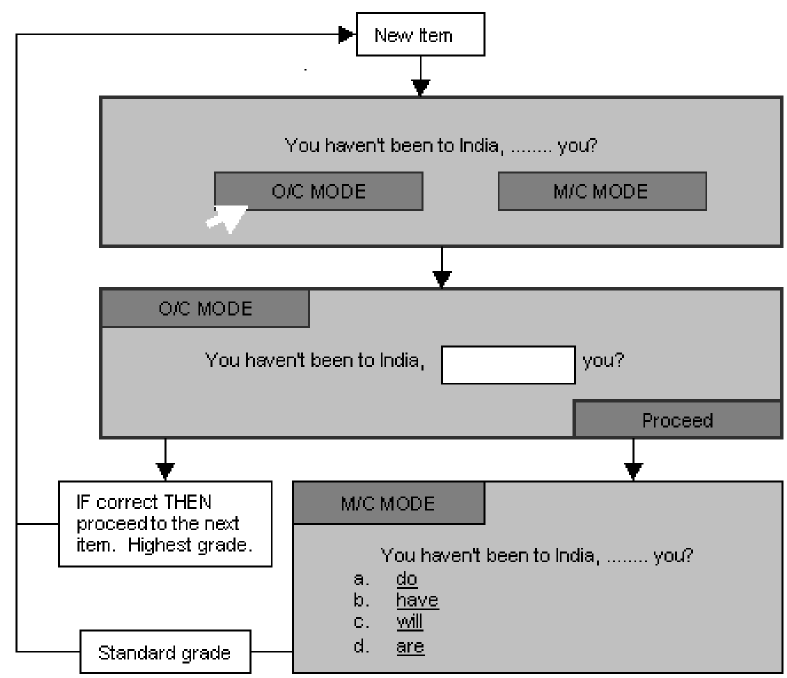

Fig. 5: Item representation in OA (O/C) - MC (M/C) mode

also a fact that while native speakers of every language tend to do mistakes in oral and written language production, they are still fluent and proficient speakers of their mother tongue.

The Learner Model

The LM reflects specific characteristics of the learner and thus it is used as the main source of the adaptive behavior of AILA. The information held is divided into domain dependent information and domain independent information. As far as the domain dependent information is concerned, the LM keeps information about: (i) the learner's knowledge level (qualitative – which levels of competence – and quantitative – how many items are correct – estimation) with respect to the average level of the items answered correctly, (ii) the learner's errors, and (iii) the learner's behaviour during his/her interaction with the tool in terms of the frequency of errors made, time of response, etc. As far as the domain independent information is concerned, the LM keeps general information about the learner such as username, age, sex, learner's right or left-handedness, last time/date the learner logged on/off. The LM is dynamically updated during the learner's interaction with AILA in order to keep track of the learner's "current state".

Item Bank and Stopping Rule

The item bank [fig.6] consists of 600 items divided in the four CEF levels of competence (A2, B1, B2, C1), signifying item difficulty ($b_{1\text{-}4}$). In each broad level of competence, items are sub-divided in three discrimination levels ($a_{1\text{-}3}$). The first discrimination level (a_1) contains items that are expected to be answered correctly by all examinees having the given competence, the second (a_2) contains items that can be answered correctly by the average examinee, while the items in the third level (a_3) can only be answered by the most competent students in this level. Finally, each discrimination level is seperated in 5 content areas ($c_{1\text{-}5}$), in order to ensure that examinees will answer a wide variety of language items.

The test starts with a given difficulty specified by the test-taker (b_x), low discrimination (a1), first content area (c1), and random item selection. If the test-taker answers in MC mode correctly, then the next item is of the same difficulty (b_x), medium discrimination (a2), second content area (c2), and random item selection, otherwise the next item is one difficulty level lower. If the examinee answers in OC mode correctly, then the next item is of higher difficulty (b_{x+1}), high discrimination (a3), fourth content area (c4), and random item selection. With this stratified way, we ensure that examinees will gradually attain their level of competence, by answering different item types.

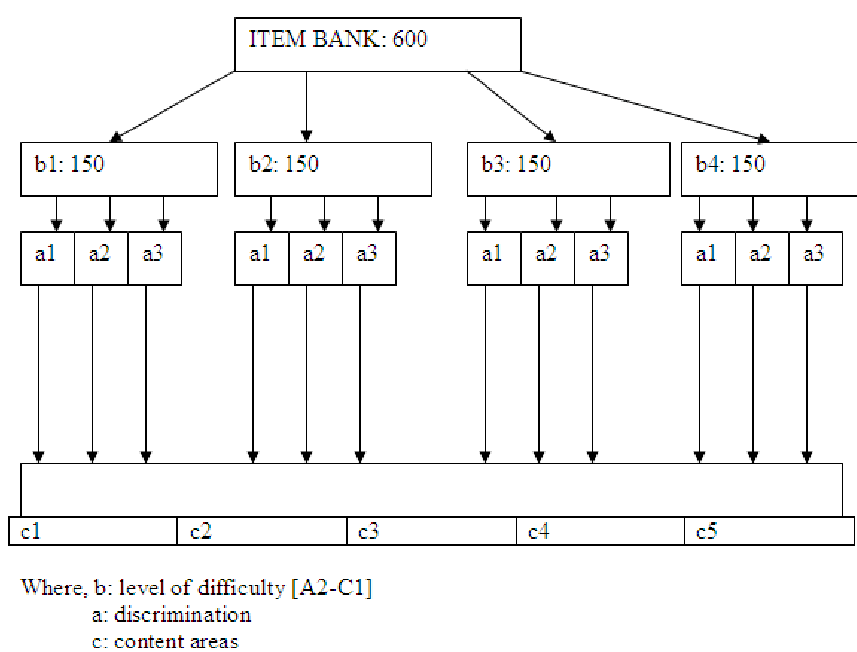

Fig.6: Item bank architecture

The stopping criterion could be time, number of items administered, change in ability estimate, content coverage, a precision indicator such as the standard error, or a combination of factors. In a variable-length adaptive test, the number of items administered to each examinee differs depending on the number of correct/incorrect responses given by him or her to the items presented. A variable-length stopping rule terminates a test once a pre-specified level of measurement precision has been reached, based on the standard error associated with a given ability. The advantage of implementing variable length stopping rules is that all examinees' ability estimates have the same measure of precision [Thissen and Mislevy, 2000]. However, a non fixed-length stopping rule has the potential to produce adaptive tests that are much shorter than P&P tests and this may have a negative effect on examinee reactions and scores. Therefore, AILA algorithm has a compulsory minimum number of 15 required items [fig.7]. Thus, the minimum test length is 15 items and the maximum is 40 items. The test stops when the examinee answers at least 15 items, having shown competence at one level of difficulty. There are no time limits per item; however, the maximum test time is 45 minutes.

[R= right responses, W = wrong responses, D = item difficulty according to level]
[IF 15 items administered]
THEN
IF R>W AND $R_{Dx} > R_{Dy-x}$
THEN
STOP
ELSE
NEXT ITEM

Fig.7: The stopping criterion

Outcomes and Conclusion

In the dawn of the European "Knowledge Society", the Council of Europe promotes multilingualism and establishes the CEF of educational standards that apply to all member-states. However, the EU is a melting-pot of civilizations and, as a result, its learning society consists of mixed-ability and mixed-intelligence students. FLA tools need to adapt to this challenge, using adaptive technologies to provide personalized, self-paced and time-effective assessments. CAT can introduce a new, student-based era in FLA that will be personalized, flexible, and sensitive to human cognition, language processing and error correction. To this end, we developed AILA, an adaptive placement test that measures competence in EFL in terms of CEF levels, giving students the choice to show productive and receptive language use. The system also tries to discern errors from mistakes by evaluating students' answers. Thus, proficient learners will be able to excel, showing active language production. All in all, CALT needs to adapt to the new social conditions, adopting a new test theory [Mislevy, 1996], gathering information from various disciplines and assimilating this information in the new assessment medium.

References

Akmajian, A. et al. (1998) *Linguistics. An Introduction to Language and Communication.* Third Edition. The MIT Press.

Ali, A. (2001) "Technology Integration and Classroom Dynamics" In *Proceedings of AACE Webnet 2001 World Conference on the WWW and the Internet,* October 23-27, Orlando, Florida, USA, pp. 7-8.

Armstrong, T. (1999) *Seven Kinds of Smart : Identifying and Developing Your Multiple Intelligences*. New American Library.

Bailly, S. et al. (2002) *Common European Framework of Reference for Languages:Learning, teaching, assessment. A Guide for Users.* Council of Europe. Language Policy Division, Strasbourg.

Burstein, J. and Chodorow, M. "Automated essay scoring for nonnative English speakers." *In Proceedings of the ACL99 Workshop on Computer-Mediated Language Assessment and Evaluation of Natural Language Processing*. June 1999, College Park, MD.

Burstein, J. et al. "Using lexical semantic techniques to classify free-responses." In *Proceedings of the ACL SIGLEX Workshop on Breadth and Depth of Semantic Lexicons.* June 1996, Santa Cruz, CA: University of California, Santa Cruz.

Chomsky, N. (1997) Powers and Prospects. Reflections on Human Language and the Social Order. Second Edition. Pluto Press.

Council of Europe, (2001) *Common European Framework of Reference for Languages***.** Cambridge University Press.

Council of Europe, (2004) *Reference Supplement to the Preliminary Pilot version of the Manual for Relating Language examinations to the Common European Framework of Reference for Languages: learning, teaching, assessment.* December 2004. Language Policy Division, Strasbourg.

Dunkel, P. "Computer-Adaptive Testing of Listening Comprehension: A Blueprint for CAT Development" *The Language Teacher Online*, October 1997.

Gardner, H. (1993) *Frames of Mind: The Theory of Multiple Intelligences.* Basic Books.

Giouroglou, H. and Economides, A. (2003), "Cognitive CAT in Foreign Language Assessment" *Eleventh International PEG Conference "Powerful ICT Tools for Learning and Teaching", PEG'2003*, 28 June-1 July, St. Petersburg, Russia.

Laufer, B. and Y. Yano. 2001. Understanding unfamiliar words in a text: do L2 learners understand how much they don't understand. *Reading in a Foreign Language* 13: 549-566.

Lord, F.M. (1980). *Applications of item response theory to practical testing problems*. Hillsdale, NJ: Lawrence Erlbaum Associates.

Mislevy, R.J. (1996), "Test Theory Reconceived." *Journal of Educational Measurement*, 33 (4), 379-416.

Powers, D. E., et al. (2001). *Stumping e-rater:® Challenging the validity of automated essay scoring* (GRE No. 98-08bP, ETS RR-01-03). Princeton, NJ: Educational Testing Service.

Sawaki, Y. (2001). *How examinees take conventional versus web-based Japanese reading tests.* Work in progress session presented at the 23rd Annual Language Testing Reserach Colloquium, March, 2001, St. Louis, MO.

Schunk, D. (1996), *Learining Theories: An Educational Perspective*. Pentice Hall.

Streeter, L. et al. (2002), "The Credible Grading Machine: Automated Essay Scoring in the DoD" Paper presented at the Interservice/Industry, Simulation and Education Conference (I/ITSEC). December 2-5, 2002. Orlando, FL.

Thissen, D., & Mislevy, R. J. (2000). Testing algorithms. H. Wainer (Ed.), *Computer Adaptive Testing: A Primer* (2nd ed., pp. 101-133). Mahwah, NJ: Lawrence Erlbaum Associates, Inc.

Twigg, C. (1994), "The Need for a National Learning Infrastructure", *Educom Review*, 29, Nos. 4,5 and 6.

Wainer, H. Et al. (2000). *Computer Adaptive Testing: A Primer* (2nd ed.). Mahwah, NJ: Lawrence Erlbaum Associates, Inc.