Proceedings 2nd International Conference on Intelligent Networking and Collaborative Systems, pp. 92-99,
Thessaloniki, Greece, 24-26 November 2010, ISBN: 978-0-7695-4278-2

Text based computer-mediated communication: A system for automated interpretation of discussion threads in distance education fora by using formal languages

Kiriakos Patriarcheas, Spyridon Papaloukas, Michalis Xenos
School of Sciences and Technology, Computer Science
Hellenic Open University
Patras, Greece
k.patriac@eap.gr, s.papaluk@eap.gr, xenos@eap.gr

Abstract— **Electronic discussion fora are increasingly becoming part of the distance education process and are an evolving field which needs to be constantly updated and redefined. This paper presents a system development for automated interpretation of messages in distance education fora by using a modelling in formal language.**

Keywords- distance education; text-based comunication; electronic fora; formal languages; AdaBoost; Naive Bayes; 1-Nearest Neighnor; WINNOW.

I. INTRODUCTION

The quality of distance education relies on the quality of communication (in a broad sense) between the student and the tutor [1]. Electronic asynchronous discussion fora (hereinafter "fora") is a key tool, supporting a large part of the distance education process. An important issue which, in recent years, has concerned researchers in the field, designers, coordinators and tutors is how they might gain, at any given moment, an overall view of the situation from a number of discussion threads in a distance education forum, not only on the quantitative level of participation but also on the level of what is discussed and where the discussion is focused on.

Given that many systems for analyzing text based Computer-Mediated Communication (CMC) messages are too tedious and time-consuming to serve as practical assessment tools [2, 3], this research aims to cover this gap through the development of a system that will automatically classify messages per content, according to a modelling developed for this.

II. THEORETICAL FRAMEWORK

According to Harasim [4] communication via text-based messages provides a high level of interactivity, which encourages collaboration and influences the learning process. The asynchronous capabilities of text-based CMC allow for more thought, reflection and processing of information [5]. These two factors indicate that electronic messages are potentially a rich source of data for researchers. Text-based messages commonly used in CMC have unique characteristics [6]. A large part of the complexity related to the analysis of human communication through the exchange of electronic messages in asynchronous discussion fora is due to the fact that while they are written texts they do not share the same features as traditional written communication and contain more characteristics of spoken communication [7]. Similarly, Kern [8] argues that the CMC is somewhere on the continuum between paper-based writing and speech. According to McCreary [9] the written word demands an exactness and coherence of thought, indicating that text-based communication results in more well planned and structured interactions. In more detail, Kol and Schcolnik [10] argue that the language complexity focuses on lexical or syntactic factors. Specifically, lexical complexity is reflected in two dimensions: *range* (lexical variation) and *size* (lexical sophistication) [11]. Syntactic complexity reflects elements such as sentence length, amount of embedding, and range and sophistication of structures [12]. This complexity leads to the question of Kol and Schcolnik [10] "how can forum discussions be analyzed?" who note that many studies focus on the nature of student messages or their length, depth and purpose.

An important issue in this framework is the content analysis, a technique frequently used in the approach of issues concerning asynchronous computer mediated discussion groups in distance education. Over the last years, numerous efforts to approach this issue have been made, stemming from different theoretical backgrounds. Indicatively, Barrett and Lally [13] have used content analysis to investigate learning and socio-emotional behaviour in learning community from the Gender differences view. Henri [7] uses the point of Cognitive and metacognitive knowledge, while Newman et al. [14] and Bullen [15] the point of Critical thinking. Many, though, start from social constructivism using different variations. Indicatively, some [16-19] utilize the approach of social constructivism in combination with knowledge construction; Jarvela and Hakkinen [20] in combination with perspective taking, while Lockhorst et al. [21] in combination with learning strategies. Rourke et al. [22] utilize the approach of inquiry community from the point of social presence, rather than [23] from the point of cognitive presence or [24] the teaching presence. In relation to Cognitive aproach, Henri [7] proposed a model for analyzing CMC messages tapping five aspects of the learning process as reflected in the messages: participative, interactive, social, cognitive, and metacognitive. This model focused on process rather than product and has been used by other researchers of CMC [5, 25] as a basis for their work or the development of their own models. Oskoz [26] and McLoughlin and Luca [27] have also developed process-based systems for the analysis of online interactions. The latter traces knowledge construction as it moves through five phases from knowledge sharing to knowledge building. In relation to Social constructivism, this emphasizes the negotiation of meaning and construction of shared understanding through dialogue [28-32]. Vygotsky [33] views learning as a social process that occurs within the zone of proximal development (ZPD), which positions dialogue as crucial to the development of thought and behavior. Consequently, dialogue becomes the focal point for understanding learning.

III. Methodological Framework

A. *Hellenic Open University & fora*

Hellenic Open University (HOU) is the main distance education institution in Greece. The HOU has currently 25,418 students (16,066 graduate, 9,301 post-graduate, and 51 PhD students) and 1,485 professors (27 of which are tenured and the rest are external tutors-consultants).

An educational unit at HOU is a course module; today, 184 course modules are offered at HOU. The tutor and all students of a module may participate in the discussion threads of each module. As far as Computer Science students are concerned, at the time of this survey, 6,067 discussion threads with 26,246 messages had been created in the 16 Computer Science modules (at graduate level) offered by HOU. With reference to the evolution of the use of HOU fora, by way of illustration, in the last academic years there has been a large increase in the number of messages in the module Introduction to Information Technology (INF10), with a (relatively) invariable number of discussion threads, as shown in Table I.

TABLE I. The evolution of the number of messages for the academic years 2005-6, 2006-7 and 2007-8 in the module INF10

Year	2005-6	2006-7	2007-8
Threads	237	236	219
Messages	982	1205	1942
Messages/Thread	4.14	5.11	8.87

After an extensive study that explored the behaviour of students of the HOU [16, 34-40] it was observed that the HOU fora make a significant contribution to the learning process as they help with organising the study of a module and the processing of and elaboration on what learners have already studied as follows:

i) for the organization of studies during the course module:

- to the communication between the teacher and the students (regularity of contacts, subject, resolution of "technical" problems etc.).
- to the organization of homework (method of use of the teaching material and the preparation of the activities, exploitation of the literature and the other sources, timetables, encountering problems related to it etc.)
- to the supply of information about the advisory meetings (their number, their duration, the timetables, the goals, their content and methodology applied, problems encountered in the ability to attend them etc.).
- to supply clarification about the procedure of preparation and evaluation of the written assignments (form, method of preparation, evaluation criteria, means of support by the teacher etc.).
- to inform about the procedure of final exams (student preparation, support by the teacher, marking criteria, means and time of examination etc.).

ii) for the elaboration and development of what the students have already studied, the HOU's fora may be exploited for:

- the presentation of consolidation exercises, short suggestions, presentation of examples, methodologies, literature etc.,
- the resolution of questions and the supply of clarifications about the teaching material.
- the interconnection between what has already been studied, subsequent chapters and the written assignment to follow.

Given the heavy flow of information carried through HOU fora, the designing and development of the system presented in this paper simulated the development of a formal language to interpret messages in the fora of HOU.

B. *Unit of analysis*

Given that the choice of a unit of analysis is dependent on the context and should be well-considered, because changes to the size of this unit will affect coding decisions and comparability of outcome between different models [41], as well as given the fact that Schrire [25] refers to a dynamic approach in which data is coded more than once and the grain size of the unit of analysis is set, depending on the purpose and the research question, it was decided to use, as unit of analysis, the category of message content.

On the observation of the discussion threads, it was realized that there are cases of messages which may comprise two (or/and more) content categories, e.g. a question about the next advisory meeting and a reply to a question concerning the study of the educational material (see below *modelling* subsection).

Thus, in such a case, the analysis at the message level used by some researchers [15-17, 22-24] is deficient in exploiting information that arises in order to reach educational conclusions, since more than one content categories may coexist in a given message.

C. Modelling of HOU distance education forum

Based on observations at HOU fora the following became evident: a) There are two categories of communication participants: Tutors and Students (for brevity, tutors will be symbolised with a *T* and students with an *E),*. b) As regards to message types, these are distinguished into questions and answers. Hereinafter, symbolised with *q* and *a* respectively, c) As to their content, messages are distinguished into those relating to (the respective symbols are given in brackets):

- the study of educational material (*M*)
- questions/answers for exercises – assignments (*X)*
- presentation of sample assignments by tutors (*P*)
- instructions (*I*)
- assignment comments, corrections (*F*)
- student comments on assignments (*D*)
- sending – receiving assignments (*J*)
- sending - receiving grade marks (*G*)
- notification of advisory meeting (*V*)
- pointless message (*L*)

Finally, the order in which above symbols will be written is: a) message carrier b) message type and c) the content of the category to which the message belongs.

Thus, we have a language which contains:

a) Terminal symbols alphabet V_T, where *V_T = {T, E, q, a, n, M, X, P, I, F, D, J, G, V, L }*

b) Non terminals alphabet V_N, where *V_N = {u, r, y, c},* more specifically :

r: represents the message carrier (*T* for tutors and *E* for students)

u: represents a pair *yc* i.e. a message type *y* (whether it is a question *q* or an answer *a*) followed by its content category.

c) The grammar P

A set of rules of the form α → β, where α and β sequences containing terminal and non-terminal symbols and α is not an empty sequence, as follows:

(1) $S \rightarrow ruS$	(8) $y \rightarrow q$	(15) $c \rightarrow F$
(2) $S \rightarrow \varepsilon$	(9) $y \rightarrow a$	(16) $c \rightarrow D$
(3) $u \rightarrow uyc$	(10) $y \rightarrow \varepsilon$	(17) $c \rightarrow J$
(4) $u \rightarrow \varepsilon$	(11) $c \rightarrow M$	(18) $c \rightarrow G$
(5) $r \rightarrow T$	(12) $c \rightarrow X$	(19) $c \rightarrow V$
(6) $r \rightarrow E$	(13) $c \rightarrow P$	(20) $c \rightarrow L$
(7) $r \rightarrow \varepsilon$	(14) $c \rightarrow I$	(21) $c \rightarrow \varepsilon$

Where ε stands for an empty symbol

d) Symbol S where every sentence generated starts with this symbol.

A message concerning a student's question about an assignment is represented as: *EqX* (where *E* for student, *q* for question and *X* for the fact that this message is about an assignment).

An indicative example is presented that contains a series of messages represented by the sequence *EqMEqXTaMX,* which, according to the above, when it should be represented a message concerning a student's message, addressing a question about the study of the educational material, followed by another student's question about the following assignment and at the end of the thread there is the reply of the teacher both for the study of the material and for the following assignment, it will be represented as follows: *E* for the student's capacity, *q* for the question, *M* as it concerns the study of the educational material, *X* for the fact that the next message concerned an assignment, *T* for the teacher's capacity, *a* for the fact that it is an answer, *M* for the fact that this reply concerns the study of educational material and *X* for the fact that the second part of the message concerns an assignment. According to the above, the sequence *EqMEqXTaMX* constitutes a sentence of the *Language* because:

Rule:

$$S \xrightarrow{(1)} ruS \xrightarrow{(1)} ruruS \xrightarrow{(1)} rururuS \xrightarrow{(3)} ruycruycruycS$$

$$\xrightarrow{(4)(6)(8)(11)} EqMruycruycS \xrightarrow{(4)(6)(8)(11)} EqMEqXruycS$$

$$\xrightarrow{(3)} EqMEqXruycycS \xrightarrow{(2)(4)(5)(9)(10)(12)} EqMEqXTaMX$$

As it is obvious from the example, while one content category corresponds to the 1st and 2nd, messages (*M* and *X* respectively), in the 3rd message there are two content categories *MX.*

It is worthy to note here that this system incorporates the sense of time along with its association with each of the 10 categories of message content chosen as unit of analysis. Given that within a message more than one content may exist, the dates are recorded for each such case and not simply in each message. Consequently, time differences may automatically exist (in days, if from each current date, by content category, it is deduced the previous one) and thus there may arise another 10 respective stacks with the above date references. Of course, the length of these stacks is equal to the length of dates minus one (-1), i.e. apart from the initial message, which is considered to be the point zero (0), where the numbering of the time differences begins.

IV. System development

According to this approach, a system of automatic classification was developed, which comprised the following: a) Data filtering: where some web pages accommodating the discussion threads of a distance education forum of HOU (which include much data having no essential information concerning the educational procedure e.g. titles, images etc.) are considered as input and creates a temporary file with the "useful" part (User name, date, message's content) which may become a source of information for educational conclusions.

b) Storage of roots files: indicated the experimental results of the algorithm execution proved wrong the initial estimate that when a message contains the question mark (English "?" or Greek ";") then it is a question, seeing that there were messages comprising a question, although they did not include a question mark, yet other phrases such as "I would like to ask", etc.

Therefore, a dynamic method to store the information required to determine the type of message (whether it is a question or answer) should be designed. This effected the decision to create an algorithm (procedure "message_type") that takes pairs of information: a) word or phrase root or symbol, and b) terminal symbol *q* or *a* if it is a question or answer, and to create a records file of two respective fields containing the above pairs.

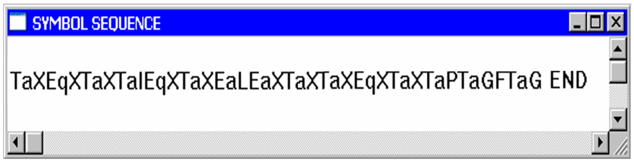
The same reasoning (procedure "content_category") was used to store information needed to determine the content category of a message, namely if it refers to a study, assignment, comment, etc., or a combination thereof (e.g. when a message refers to both a study and assignment). Therefore, an algorithm was created that would input pairs of information as follows: a) word or phrase root, and b) terminal symbol of content category (*M, X, P, I, F, D, J, G, V, L*). Thus, it is possible to add more content categories if needed.

As for the roots file creation on the message content category the basic syntax rule in Greek Language that the endings are created by combinations smaller and simpler endings was followed. The rules that are used are in form: A1 → A2 (conditions) with the significance of replacement of ending A1 from the A2 if the letters that remain from the A1 satisfy the condition. At the first step of algorithm becomes handling of plurals and aorists (pasts). This step is separated in three sub-steps. The first handles plurals (e.g. in English language: caresses → caress). The second removes and/or changes the endings, if this is required (e.g. in English: ed and ing). The process is continued and, if the ending has been removed, the root that remains is converted (e.g. in English language: conflated → conflate, motoring → motor, agreed → agree). The third (sub-step) converts the final letter (e.g. in English language: happy → happi). The steps deal mainly with the different sequence in the ending groups.

The second stage materialises an algorithm of roots export of words that produces the result with one parsing and removes the endings based on the *Quick Fitting (QF)* principle. The algorithm includes two sub stages. First, is the sub stage of roots export (stemming phase) where the endings are removed and the application checks if (by any chance) coincidentally there are exceptions between the steps. The second stage uses rules for the reconstruction of words from the endings.

c) strings' production: receiving as input the temporary file with the "useful" information (User name, date, message's content) and the files with the couples of roots words/ phrases/ symbols and terminal symbols and presents (and stores) the respective strings with the relative extensible file, so as the results to be kept for further exploitation.

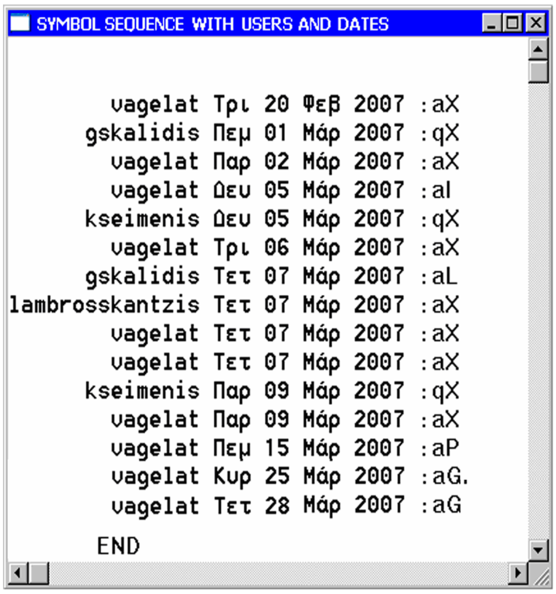
The execution of "symbol_sequence" results in the automated output of strings (see in Fig. 1) where each one represents the messages of the respective discussion thread and finishes with the word END, as follows:



```
TaXEqXTaXTaIEqXTaXEaLEaXTaXTaXEqXTaXTaPTaGFTaG END
```

Figure 1. Strings output according to the model.

Figure 2 shows the same results after the execution of "symbol_sequence_with_users_and_dates" where the symbols of the message carrier have been replaced by User names and date per message is also displayed.



```
     vagelat Τρι 20 Φεβ 2007 :aX
   gskalidis Πεμ 01 Μάρ 2007 :qX
     vagelat Παρ 02 Μάρ 2007 :aX
     vagelat Δευ 05 Μάρ 2007 :al
   kseimenis Δευ 05 Μάρ 2007 :qX
     vagelat Τρι 06 Μάρ 2007 :aX
   gskalidis Τετ 07 Μάρ 2007 :aL
lambrosskantzis Τετ 07 Μάρ 2007 :aX
     vagelat Τετ 07 Μάρ 2007 :aX
     vagelat Τετ 07 Μάρ 2007 :aX
   kseimenis Παρ 09 Μάρ 2007 :qX
     vagelat Παρ 09 Μάρ 2007 :aX
     vagelat Πεμ 15 Μάρ 2007 :aP
     vagelat Κυρ 25 Μάρ 2007 :aG.
     vagelat Τετ 28 Μάρ 2007 :aG

    END
```

Figure 2. Results after the replacement of symbols T and E by the respective user names and display, at the same time, of message publication dates (Days: Κυρ=Sun, Δευ=Mon, Τρι=Tue, Τετ=Wed, Πεμ=Thu, Παρ=Fri, Σαβ=Sat, Months:Φεβ=Feb, Μάρ=Mar).

V. EXPERIMENTS

During the development of the system, we followed the experimental control process. At first, experiments were carried out by using 80 discussion threads of the INF10 module of the academic year 2007-8. Given that 219 threads with 1,942 messages had been created throughout the year, there was the ratio of approximately 9 messages (in particular 8.87) per discussion thread. Therefore, out of the 80 selected threads, we tried to use those containing 8 or 9 messages for the purpose of experimental control. Thus, we finally chose 80 discussion threads with 712 messages in total (average 8.90 messages/thread).

At the first experimental operation, the word root files in relation to both the type (question/answer) and (mainly) the content category of message contained 18 and 92 entries respectively. Under these conditions, we ended up having 58 discussion threads with no errors and 16 threads with only one wrong symbol (compared to what was expected). Namely, out of (approximately) 9 messages (of each of the 16 threads), 8 of them were correct and one message was wrong because it did not contain not even one of the 92 provided word roots. Respectively, there were 5 threads with two errors and 1 thread with more errors (this thread was created before Christmas holidays and its messages contained mainly wishes). We should note here that there has been no error regarding the type of messages (question/answer), only in terms of determining the content category. Following the observation/study of messages in the 21 threads that contained 1 or 2 errors, we recorded 49 additional word roots (concerning the content category) and we decided to enter them in the root file. The experimental operation performed in the same 80 threads had, clearly, better results, with total success in 70 threads, one wrong symbol in 8 threads, two errors in 1 thread, and 1 thread that did not actually refer to educational content (Table II).

At this point it should be clarified that the control of the results produced by the system in this phase, was conducted with manual comparison of all the messages in the discussion threads that were used in order to control system reliability at the first degree.

Subsequently, terms and concepts were extracted as features (word roots) from the messages in the training and test corpus. The feature extraction process consisted of the stages described in system development section (Roots files storage subsection) using the stemming algorithms (stage 1 and 2), resulting in a total number of 92 (in 1st experimental operation) and 141 (in 2nd) distinct term features (word roots).

TABLE II. EXPERIMENTAL OPERATION - PHASE A

	1st Exp. operation	2nd Exp. operation
Threads	80	80
Messages	712	712
Messages/Thread	8.9	8.9
Success (threads with no errors)	58	70
Threads with one error	16	8
Threads with two errors	5	1
Threads with more errors	1	1
Correct interpretation (in message level)	677	693
Wrong interpretation (in message level)	35	19

Given that the 8 discussion threads with one error were found not to have any common word root feature that would adequately correspond, we decided to initiate the second experimental phase (B'). Classification was performed according to international literature [42-48], using the algorithms indicated for this purpose: Naive Bayes (NB), 1-Nearest Neighnor (1-NN), WINNOW and discrete AdaBoost (in the form generalized by Nock & Nielsen [49] based on Freund and Schapire [50]).

During this phase every algorithm was formed using the data collected from the academic year 2007-8. Subsequently, a group of data for two other academic years (2005-6 and 2006-7) was also collected. The results show that the discrete AdaBoost algorithm produced the greatest accuracy. This result agrees with Bloehdorn & Hotho [48] who used the discrete AdaBoost algorithm in a similar experiment. The accuracy is denoted in the Table III.

TABLE III. ACCURACY OF ALGORITHMS FOR THE ACADEMIC YEARS 2005-6, 2006-7 AND 2007-8 – PHASE B.

	2005-6	2006-7	2007-8	Average accuracy (1)	Average accuracy (2)
In thread level					
AdaBoost	75.11	80.08	87.21	80.64	80.80
Naive Bayes	72.47	77.83	86.18	78.66	78.82
1-Nearest Neighnor	73.45	76.66	83.65	77.77	77.92
WINNOW	70.13	73.24	83.10	75.34	75.49
In message level					
AdaBoost	92.36	95.19	97.89	94.96	95.15
Naive Bayes	89.11	92.51	96.73	92.59	92.78
1-Nearest Neighnor	90.31	91.13	93.89	91.60	91.78
WINNOW	86.23	87.06	93.27	88.67	88.85

The average accuracy (1) corresponds to the total number of threads and messages, while in (2) the years have an equal participation (1/3) in the total average.

VI. CONCLUSION

A big part of the research presented in the international literature concerning distant education's fora, refer to the content analysis, which principally aims to reveal information invisible at first sight. There is a variety of approaches, varying both in the level of details and in the type of categories of analysis used stemming from different theoretical backgrounds.

The development of this system was stimulated by the heavy flow of information in HOU's distance education fora, and it aspires to cover a gap in the interpretation of messages in an asynchronous discussion forum for distance education, by creating a system that automatically classifies messages according to a modelling built to this effect. This system uses the content category as unit of analysis for the messages' interpretation.

The creation of this system makes an important contribution to the decoding of discussions in fora, and aims at summary identification of discussions which do not develop in the desired way. Therefore this approach can be used as a tool which may assist in "intelligent" coordination, in order to limit potential malfunctions, and it could ultimately be interpreted as a step towards a procedure for formulating quality indicators for the educational value of a forum in distance education.

Even though the approach presented may apply to other distance education institutes that use fora, there are limitations. It is obvious that satisfactory results in the operation of the system are based on the fact that they concern specific subjects with a defined field of knowledge, and therefore more standardized dialogues compared to similar systems of text classification that refer to more open forms of discussion. Furthermore, the system that is presented was designed for students who are attending courses in the Greek language; therefore the results may be different in other languages. Another parameter is that in the HOU Forum environment, after an initial agreement between tutors and students, (Greek) words are unabbreviated, therefore the satisfactory results of the system may have been different if abbreviations or greeklish (Greek words in the Latin alphabet) were used, as used profusely in other forms of communication (e.g. SMS, mobile learning). An important parameter which could also be a future goal is the use of the system in the case where postings are not signed, as they are now, but they are anonymous, something which would not be possible now under HOU's legislative framework of operation, but only after obtaining the relevant permission from the Greek authorities.

REFERENCES

[1] T. Plagemann and V. Goebel, "Analysis of quality-of-service in a wide-area interactive distance learning system." Telecommunication Systems Modeling, Analysis, Design and Management, vol. 11, pp. 139–160, March 1999.

[2] L.P. Dringus and T.J. Ellis, "Building the SCAFFOLD for evaluating threaded discussion forum activity: Describing and categorizing contributions", Proc. 34th ASEE/IEEE Frontiers in Education Conference, Savannah, GA, Oct. 2004.

[3] S. Ho, "Evaluating students. participation in on-line discussions. Proc. 23rd annual ascilite conference: Who's learning? Whose technology?", Proc. (AUSWEB 2002), Southern Cross University, Queensland, Australia, Jul. 2002.

[4] L. Harasim, "Online Education: An environment for co llaboration and intellectual amplification", In (ed. L.M.Harasim.), Online Education: Perspectives on a new environment, pp. 39-64. Praeger, New York, 1990.

[5] N. Hara, C. Bonk, and C. Angeli, “Content analysis of on-line discussion in an applied educational psychology course”, Instructional Science, vol. 28, pp. 115-152, March 2000.

[6] J.A. Duncan-Howell, “eCAF: A new tool for the conversational analysis of electronic communication”, Proc. British Educational Research Association (BERA 2008) Annual Conference, Herriot-Watt University, Edinburgh, Sep. 2008.

[7] F. Henri, “Computer conferencing and content analysis”, In A. R. Kaye (Ed.), Collaborative learning through computer conferencing, The Najadan Papers, pp. 117–136. Springer-Verlag, London, 1992.

[8] R. Kern, “Perspectives on technology in learning and teaching languages”, Teachers of English to Speakers of Other Languages (TESOL) Quarterly, vol. 40, pp. 183-210, March 2006.

[9] E.K. McCreary, “Three behavioral models for computer-mediated communication”, In (ed. L.M. Harasim), Online Education: Perspectives on a new environment, pp. 117-130, Praeger, New York, 1990.

[10] S. Kol and M. Schcolnik, “Asynchronous forums in EAP: assessment issues”, Language Learning & Technology, vol. 12, pp. 49-70, June 2008.

[11] K. Wolf-Quintero, S. Inagaki, and H-Y. Kim, “Second Language Development in Writing: measures of fluency, accuracy and complexity”, Honolulu: Second language teaching and curriculum center, University of Hawaii, 1998.

[12] L. Ortega, “Syntactic complexity measures and their relationship to L2 proficiency: A research synthesis of college-level L2 writing”, Applied Linguistics, vol. 24, 492–518, December 2003.

[13] E. Barrett and V. Lally, “Gender differences in an on-line learning Environment”, Journal of Computer Assisted Learning, vol. 15, pp. 48–60, March 1999.

[14] F.W. Newman, “The prospects for classroom thoughtfulness in high school social studies”, In (eds C. Collins & J.N. Mangieri), Teaching thinking: An agenda for the 21st century, pp. 105 – 132, 1992.

[15] M. Bullen, “A case study of participation and critical thinking in a university-level course delivered by computer conferencing”, University of British Columbia, Vancouver, Canada, 1997.

[16] K. Patriacheas and M. Xenos, “Collaborative learning: Reasons that influence the participation of students in distance education fora”, Article accepted as full paper in Social Applications for Lifelong Learning, 2010.

[17] A. Veerman and E. Veldhuis-Diermanse, “Collaborative learning through computer-mediated communication in academic education”, In Euro CSCL, pp. 625–632. University of Maastricht, McLuhan institute, Maastricht, 2001.

[18] J.B. Pena-Shaff and C. Nicholls, “Analyzing student interactions and meaning construction in computer bulletin board discussions”, Computers & Education, vol .42, pp. 243–265, April 2004.

[19] A. Weinberger and F. A Fischer, “A framework to analyze argumentative knowledge construction in computer-supported collaborative learning”, Computers & Education, vol. 46, pp. 71-95, January 2006.

[20] S. Jarvela and P. Hakkinen, “Web-based cases in teaching and learning: The quality of discussions and a stage of perspective taking in asynchronous communication”, Interactive Learning Environments, vol. 10, 1, pp. 1–22, 2002.

[21] D. Lockhorst, W. Admiraal, A. Pilot, and W. Veen, “Analysis of electronic communication using 5 different perspectives”, Proc. symposium conducted at the 30th Onderwijs Research Dagen (ORD), Kerkrade, Netherlands, 2003.

[22] L. Rourke, T. Anderson, D.R. Garrison, and W. Archer, “Assessing social presence in asynchronous text-based computer conferencing”, Journal of Distance Education, vol. 14, 2, pp.51–70, 1999.

[23] D.R. Garrison, T. Anderson, and W. Archer, “Critical thinking, cognitive presence, and computer conferencing in distance education”, American Journal of Distance Education, vol. 15, 1, pp. 7–23, 2001.

[24] T. Anderson, L. Rourke, D.R. Garrison, and W. Archer, “Assessing teaching presence in a computer conference context”, Journal of Asynchronous Learning Networks, vol. 5, 1-17, September 2001.

[25] S. Schrire, “Knowledge building in asynchronous discussion groups: going beyond quantitative analysis”, Computers & Education, vol. 46, pp. 495-70, January 2006.

[26] A. Oskoz, “Students’ dynamic assessment via online chat”, CALICO Journal , vol. 22, pp. 512–536, May 2005.

[27] C. McLoughlin and J. Luca, “Cognitive engagement and higher order thinking through computer conferencing: We know why but do we know how?”, In (eds. A. Herrmann & M. M. Kulski) Flexible futures in tertiary teaching. Proc. 9th Annual Teaching Learning Forum, Curtin University of Technology, Perth, 2000.

[28] D. Jonassen, M. Davidson, M. Collins, J. Campbell and B.B. Haag, “Constructivism and computer-mediated communication in distance education”, The American Journal of Distance Education, vol. 9, 2, pp. 7-26, 1995.

[29] C.J. Bonk and D. Cunningham, “Searching for learner-centered, constructivist, and sociocultural components of collaborative educational learning tools”, In (eds. C.J. Bonk and K.S.King), Electronic Collaborators: Learner-centered Technologies for Literacy, Apprenticeship, and Discourse (pp. 25-50). Lawrence Erlbaum, Mahwah, NJ., 1998.

[30] C.J. Bonk and K.A. Kim, “Extending sociocultural theory to adult learning”, In (eds. C.M. Smith and T. Pourchot), Adult Learning and Development: Perspectives from Educational Psychology (pp. 67-88). Lawrence Erlbaum, Mahwah, NJ. (1998).

[31] T. Paulus, “CMC modes for learning tasks at a distance”, Journal of Computer-Mediated Communication, vol. 12, article 9, July 2007.

[32] H. Kanuka and T. Anderson, “Online social interchange, discord, and knowledge construction”, Journal of Distance Education vol. 13, 1, pp. 57-74, 1998.

[33] L.S. Vygotsky, Mind in Society, Harvard University Press, Cambridge, MA, 1978.

[34] S. Kotsiantis, K. Patriarcheas, and M. Xenos, “A combinational incremental ensemble of classifiers as a technique for predicting students' performance in distance education”, Knowledge Based Systems, vol. 23, pp. 529–535, August 2010.

[35] K. Patriarcheas and M. Xenos, “Modelling of distance education forum: Formal languages as interpretation methodology of messages in asynchronous text-based discussion”, Computers & Education, vol. 52, pp. 438–448, February 2009.

[36] K. Patriarcheas and M. Xenos, “Asynchronous distance education forum - Brainstorming vs. Snowballing: a case study for teaching in programming didactics”, Proc. 8th International Conference on Web-based Learning, (ICWL 2009), Springer-Verlag, Aug. 2009. pp. 322-331.

[37] K. Patriarcheas, S. Papaloukas and M. Xenos, “Suitable asynchronous distance education fora size for working groups in informatics teachers training”, Proc. 4th IEEE Balkan Conference in Informatics (BCI 2009) IEEE Computer Society Press, Sep. 2009, pp. 157-162.

[38] K. Patriarcheas and M. Xenos, “Educational techniques comparative study by using combined environment via computer and mobile devices in asynchronous discussion forum”, Proc. 9th IEEE International Conference on Mobile Business and 9th Global Mobility Roundtable (IEEE ICMB/GMR 2010) IEEE Computer Society Press, Jun. 2010, pp. 297-304.

[39] K. Patriarcheas and M. Xenos, “The message content category as analysis unit for discussions study in asynchronous distance education fora”, Proc. 4th AIS Mediterranean Conference on Information Systems (AIS MCIS 2009) AIS, Sep. 2009, pp. 75-84.

[40] K. Patriarcheas and M. Xenos, “The Asynchronous Distance Education by means of Internet: Factors that Influence its Effectiveness. The Case of Hellenic Open University”, Proc. 12th Panhellenic Conference on Informatics (PCI 2008), IEEE Computer Society Press, Aug. 2008, pp. 204 - 208.

[41] D. Cook and J. Ralston, “Sharpening the focus: methodological issues in analysing online conferences”, Technology, Pedagogy and Education, vol. 12, 3, pp. 361–376, 2003.

[42] C. Aggarwal, S. Gates, and P. Yu, “On the merits of building categorization systems by supervised clustering”, Proc. 5th ACM SIGKDD international conference on Knowledge discovery and data mining, ACM Press, Aug. 1999, pp. 352-356.

[43] Y. Yang and X. Liu, “A re-examination of text categorization methods”, Proc. 22nd annual international ACM SIGIR conference on Research and development in information retrieval, ACM Press. Aug. 1999, pp. 42-49.

[44] S. Dumais & H. Chen, “Hierarchical Classification of Web Content”, 23nd annual international ACM SIGIR Conference on Research and Development in Information Retrieval, ACM Press, July 2000, pp. 256-263.

[45] M. Kongovi, J.C. Guzman, and V. Dasigi, “Text Categorization: An Experiment Using Phrases”, Lecture Notes In Computer Science, Vol. 2291, pp. 213-228. Springer-Verlag, 2002.

[46] F. Sebastiani, “Machine Learning in Automated Text Categorization”, ACM Computing Surveys,vol. 34, pp. 1-47, March 2002.

[47] F. Sebastiani Classification of Text, Automatic. Encyclopedia of Language & Linguistics, 2nd Edition, Section: Applications of natural language processing, Vol. 14. Elsevier Science Publishers, 2006.

[48] S. Bloehdorn and A. Hotho, “Boosting for Text Classification with Semantic Features”, Advances in Web Mining and Web Usage Analysis. Lecture Notes In Computer Science, Vol. 3932, pp. 149-166. Springer-Verlag, 2006.

[49] R. Nock and F. Nielsen, “A Real Generalization of discrete AdaBoost”, Artificial Intelligence, vol. 171, pp. 25-41, January 2007.

[50] Y. Freund and R.E. Schapire A Decision-Theoretic generalization of on-line learning and an application to Boosting. Journal of Computerand System Sciences, vol. 55, pp. 119–139, August 1997.