Repetition and Rhythmicity Based Assessment Model for Chat Conversations

Costin-Gabriel Chiru[1], Valentin Cojocaru[1], Stefan Trausan-Matu[1,2], Traian Rebedea[1], and Dan Mihaila[1]

[1] "Politehnica" University of Bucharest, Department of Computer Science and Engineering,
313 Splaiul Independetei, Bucharest, Romania
[2] Research Institute for Artificial Intelligence of the Romanian Academy,
13 Calea 13 Septembrie, Bucharest, Romania
costin.chiru@cs.pub.ro, valentin.cojocaru@cti.pub.ro,
{stefan.trausan,traian.rebedea,dan.mihaila}@cs.pub.ro

Abstract. This paper presents a model and an application that can be used to assess chat conversations according to their content, which is related to a number of imposed topics, and to the personal involvement of the participants. The main theoretical ideas that stand behind this application are Bakhtin's polyphony theory and Tannen's ideas related to the use of repetitions. The results of the application are validated against the gold standard provided by two teachers from the Human-Computer Interaction evaluating the same chats and after that the verification is done using another teacher from the same domain. During the verification we also show that the model used for chat evaluation is dependent on the number of participants to that chat.

Keywords: Rhythmicity, Polyphony, Chat Analysis, Repetition, Involvement, Computer-Supported Collaborative Learning, Dialogism.

1 Introduction

Lately, one can see a tendency toward an increased use of collaborative technologies for both leisure and work. There is an intense use of instant messaging systems (chats), blogs, forums, etc. for informal talks in our spare time. Their usage is also encouraged by the learning paradigm of Computer-Supported Collaborative Learning (CSCL) that suggests these tools are also suitable for collaborative knowledge building: "many people prefer to view learning as becoming a participant in a certain discourse" [10, 13].

Unfortunately, these tools do not provide analysis facilities that keep up with the above mentioned tendency, and therefore nowadays there are a lot of collaborative conversations that cannot be assessed – one cannot say whether one such conversation was good/efficient or not and is also unable to evaluate the participation of every participant.

Most of the research done in conversations' analysis is limited to a model with two interlocutors where at all moments there is usually only one topic in focus. The

M. Kryszkiewicz et al. (Eds.): ISMIS 2011, LNAI 6804, pp. 513–522, 2011.

analysis is often based on speech acts, dialog acts or adjacency pairs [6]. Most of the time, the analysis is done to detect the topics discussed and to segment the conversation [1, 9] or to identify the dialogue acts [7].

However, there are situations when more than two participants are involved in a conversation. This claim is obvious for forums, but is also valid for chats allowing explicit referencing, like ConcertChat [5]. In such cases, some complications appear, because the conversation does not follow only one thread, multiple topics being discussed in parallel. Therefore, a new model is needed, which allows the understanding of the collaboration mechanisms and provides the means to measure the contributions of participants: the inter-animation and the polyphony theory identified by Bakhtin [2] which states that in any text there is a co-occurrence of several voices that gives birth to inter-animation and polyphony: "Any true understanding is dialogic in nature." [13]. The same idea is expressed in [8]: "knowledge is socially built through discourse and is preserved in linguistic artefacts whose meaning is co-constructed within group processes".

For the moment, there are very few systems that use the polyphony theory for the conversation's analysis, PolyCAFe [12] being one such example. This system analyzes the contribution of each user and provides abstraction and feedback services for supporting both learners and tutors. It uses Natural Language Processing techniques that allow the identification of the most important topics discussed (with TF-IDF and Latent Semantic Analysis), speech acts, adjacency pairs, Social Network Analysis in order to identify the conversation threads and the individual involvement of the participants.

In this paper, we present a system that also starts from Bakhtin's polyphony theory [2, 3], where by voice we understand either a participant to the chat, or an idea (a thread of words that are present throughout the chat in order to represent something). This larger view of the notion of "voice" was inspired by Tannen's ideas [11] related to the use of repetitions as a measure of involvement. The purpose of the system is to evaluate the quality of the whole conversation from the point of view of participants' involvement in the conversation and by the effectiveness of the conversation from some given key-concepts point of view.

The paper continues with the presentation of the functions of repetitions and the information that we have extracted from chat conversations considering these functions. After that, we present the results of the application's validation and what we have undertaken for its verification. The paper concludes with our final remarks.

2 Functions of Repetition

Deborah Tannen identified four major functions of repetitions in conversations: production, comprehension, connection and interaction. She also pointed out that these functions taken together provide another one – the establishment of coherence as interpersonal involvement [11].

Repetition "facilitates the production of more language, more fluently" [11]. People are supposed to think about the things that they are about to utter and using repetition, the dead times that could appear during this time are avoided, and thus the fluency of the talk is increased.

The comprehension benefits from the use of repetitions in two ways. First of all, the information is not so dense when using repetitions and the one receiving it has enough time to understand it. Secondly, repetition is also useful for comprehension because usually only the important concepts are repeated, which signals what is the real message of the conversation, or what does it emphasize.

The repetition also serves as a linking mechanism for connecting the phrases from the text. Through repetition, the transition between ideas is softer, and the topics seem to be better connected. Repetition “serves a referential and tying function” [4].

In the same time, repetition has a role in connecting the participants also, because the author is able to present his opinion on the spoken subjects, emphasizing the facts that he/she believes are of greater importance and trying to induce the same feelings in the audience. Therefore, the repetition also has an interactional role by bonding the “participants to the discourse to each other, linking individual speakers in a conversation and in relationships” [11].

According to Tannen [11], the combination of all the previous functions leads to a fifth purpose – the creation of interpersonal involvement. Repeating the words of the other speakers, one shows his/her response according to what previous speakers said, along with his/her attitude by presenting their own facts and therefore keeping the conversation open to new interventions.

Tannen considers that “dialogue combines with repetition to create rhythm. Dialogue is liminal between repetitions and images: like repetition is strongly sonorous” [11].

3 Extracted Information

Considering the above ideas, we have built an application that tracks the repetitions from a conversation and evaluates the contribution of the users in terms of their involvement and the quality of the conversation in terms of some given key concepts that needed to be debated. In this analysis, we did not consider repetition only as exact apparition of the same word, but in the broader sense of repetition of a concept determined using lexical chains built using WordNet (http://wordnet.princeton.edu).

The information that we collected was both qualitative and quantitative:

- how *interesting* is the conversation for the users - counted as the number of a user’s replies, since once a conversation is interesting for a user, it is more likely that he/she will be interested in participating and therefore will utter more replies than if he/she is not interested in the subject debated in the conversation;
- *persistence* of the users -the total number of the user’s consecutive replies;
- *explicit connections* between the users’ words - considered as the explicit references made by the participants (facility provided by ConcertChat environment);
- *activity* of a user - the average number of uttered characters per reply for that user. This information is needed in addition to the number of uttered replies because we desire that the answers to be as elaborate as possible, thus giving a higher probability to the apparition of important concepts;

- *absence* of a user from the conversation - determined as the average time between a user's consecutive replies;
- *on topic* - a qualitative measure of the conversation, showing to what degree the participants used concepts related to the ones imposed for debating. This measure is intended to penalize the off-topic debate;
- *repetition* - how often a participant repeats the concepts introduced by others, showing the interaction between users and the degree of attention devoted by one participant to the words of the others;
- *usefulness* of a user - how often the concepts launched by a user have been used by a different participant;
- *topic rhythmicity* - the number of replies between two consecutive occurrences of the same topic. This measure is also intended to eliminate off-topic talk.

Once we decided what information will be extracted, we needed to determine the threshold values that allow us to consider a chat to be useful or not from the debated concepts and the participants' involvement points of view.

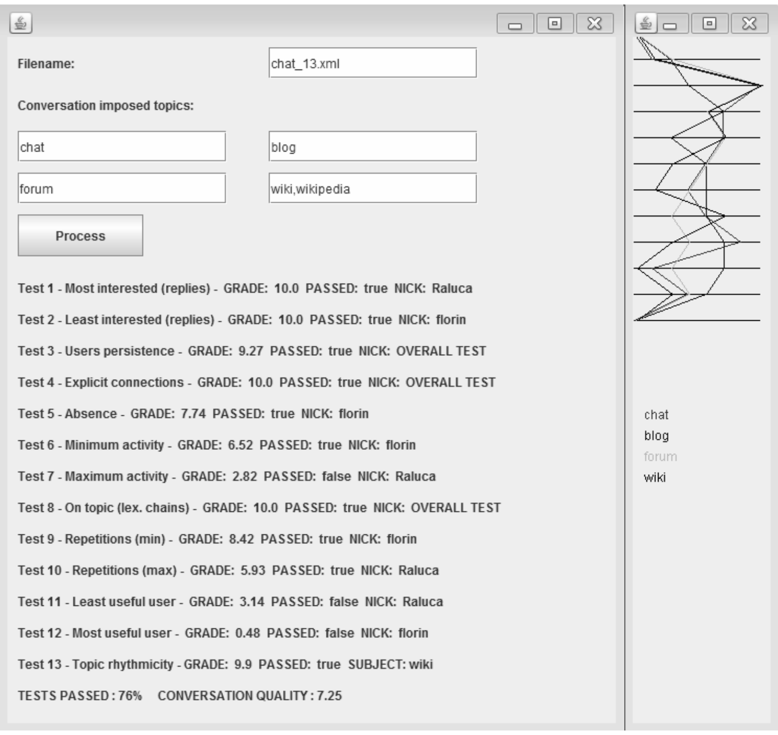
In order to determine these values, we considered 6 chats consisting of 1886 replies – ranging from 176 to 559 replies – that have been created by undergraduate students in the senior year involved in the Human-Computer Interaction (HCI) course using the ConcertChat environment [5]. They were divided in small groups of 4-5 students, every participant having to present a web-collaboration platform (chat, forum, blog, wiki) and prove its advantages over the ones chosen by the other participants.

The purpose of the chats was to facilitate the understanding of the pros and cons of all the given platforms and to find the best (combination of) communication and collaboration technologies to be used by a company to support team work.

These chats were automatically analyzed using the application that we have developed. A couple of tests have been developed starting from the collected information. For each of these tests, the application gives a grade from 0 to 10, specifies if that test have been passed or not and what was the cause of that test (a person, a topic or an overall criterion). See Figure 1 for an output of the application.

Based on the obtained values we identified the thresholds and the tendencies to be desired for a chat. All these values are presented in Table 1.

As it can be seen, we usually want high values: we want both the most and least interested person in the chat to be as active as possible (test 1 and 2) because it means they had a reason to discuss more – either for presenting more information or for debating more the given topics; we want the explicit connections between users to be as high as possible (thus showing the involvement in the discussion – test 4); the minimum/maximum activity (reflected as the average number of words/characters per utterance) should be high as well, because we desire to have elaborated sentences and not just key words appearing sporadically (test 6 and 7); the chat should contain as many words as possible related to the subjects given as an input, therefore discouraging spamming and off-topic discussions (test 8); we also should look for high values in repetitions, for they tell us that the users were paying attention to other users' words (tests 9 and 10); and finally, we want users to say important things that can be useful for the other participants to better understand the debated subjects and that can help them build their own ideas on those users' words (tests 11 and 12).



Filename: chat_13.xml

Conversation imposed topics:

chat | blog

forum | wiki,wikipedia

```
Test 1 - Most interested (replies) -  GRADE:  10.0  PASSED:  true  NICK:  Raluca
Test 2 - Least interested (replies) -  GRADE:  10.0  PASSED:  true  NICK:  florin
Test 3 - Users persistence -  GRADE:  9.27  PASSED:  true  NICK:  OVERALL TEST
Test 4 - Explicit connections -  GRADE:  10.0  PASSED:  true  NICK:  OVERALL TEST
Test 5 - Absence -  GRADE:  7.74  PASSED:  true  NICK:  florin
Test 6 - Minimum activity -  GRADE:  6.52  PASSED:  true  NICK:  florin
Test 7 - Maximum activity -  GRADE:  2.82  PASSED:  false  NICK:  Raluca
Test 8 - On topic (lex. chains) -  GRADE:  10.0  PASSED:  true  NICK:  OVERALL TEST
Test 9 - Repetitions (min) -  GRADE:  8.42  PASSED:  true  NICK:  florin
Test 10 - Repetitions (max) -  GRADE:  5.93  PASSED:  true  NICK:  Raluca
Test 11 - Least useful user -  GRADE:  3.14  PASSED:  false  NICK:  Raluca
Test 12 - Most useful user -  GRADE:  0.48  PASSED:  false  NICK:  florin
Test 13 - Topic rhythmicity - GRADE:  9.9  PASSED:  true  SUBJECT: wiki
TESTS PASSED : 76%    CONVERSATION QUALITY : 7.25
```

Fig. 1. Application output for a given chat

Table 1. The values obtained for the chats with 4-5 participants involved and the desired tendencies for the tests

Name of test	Tend ency	Chat 1	Chat 2	Chat 3	Chat 4	Chat 5	Chat 6	Min	Max
0. Utterance number	high	183	291	377	176	559	300	176	559
1. Most interested	high	0.38	0.4	0.3	0.34	0.38	0.32	0.3	0.4
2. Least interested	high	0.11	0.18	0.15	0.16	0.12	0.08	0.08	0.18
3. Users persistence	low	0.31	0.08	0.16	0.07	0.17	0.17	0.07	0.31
4. Explicit connections	high	0.43	0.79	0.27	0.4	0.19	0.48	0.19	0.79
5. Absence	low	0.047	0.019	0.016	0.033	0.011	0.037	0.011	0.047
6. Minimum activity	high	8	36	27	14	52	6	6	52
7. Maximum activity	high	59	132	93	48	345	48	48	345
8. On topic - lex. chain	high	0.23	0.28	0.21	0.19	0.2	0.25	0.19	0.28
9. Repetitions (min)	high	0.1	0.13	0.1	0.15	0.08	0.11	0.08	0.15
10. Repetitions (max)	high	0.14	0.15	0.13	0.18	0.12	0.15	0.12	0.18
11. Least useful user	high	1.84	2.03	2.1	1.73	2.1	2.69	1.73	2.69
12. Most useful user	high	2.12	2.16	2.16	2.16	2.36	2.95	2.12	2.95
13. Topic rhythmicity	low	1.36	0.76	1.51	1.69	0.92	1.22	0.76	1.69
14. Passed tests (%)	high	15	76	23	30	46	53	15	76
15. Quality	high	2.72	7.25	3.44	4.1	5.08	4.62	2.72	7.25

Now we shall focus our attention on the tests where small values are required. The first test that shows such a characteristic (test 3) is related to the users' persistence in the chat expressed as the number of consecutive replies uttered by a user without the intervention of the other participants, and based on our results we want them to be as few as possible. The idea behind this is the fact that too many consecutive replies of the same user show that the other participants had nothing to add or comment and that is a sign of not being involved, not paying attention to what that user had to say. More than that, when a user utters too much content that is not interesting for the other participants, they tend to get bored and they lose the interest in the conversation as a whole, which results in even less intervention from their part and a poor quality conversation. The second test that requires small values to show a high involvement of the participants is test number 5, which measures the maximum time between two consecutive replies of a user. If a user is taking too long to respond then he/she is not actively participating in that chat (the user is considered to be missing that part of the conversation). The last test needing small values is test number 13, which basically states that we need a small number of replies between two consecutive occurrences of a specific topic – the given topics should have high frequencies in the conversation. We desire a constant deliberation on all topics and not just users speaking in turns about the topics that were provided in order to be debated. This test also has a graphical representation of the provided topics' rhythmicity, therefore it is easier to understand what we measure, based on its graphical depiction (see Figure 2).

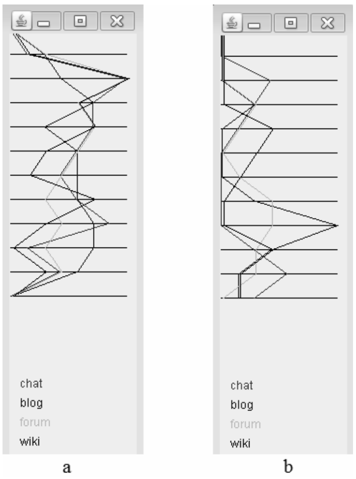


Fig. 2. Graphical representation of topic rhythmicity. a) chat with high rhythmicity for the debated topics; b) chat with poor rhythmicity for some of the topics.

In the above figure, there are two examples of rhythmicity in chats. The chats have been divided in equal shares (separated by the horizontal lines) and, in each of them, the topics to be debated are represented by a different line. The more a topic is debated in a share of a conversation, the closer is the line representing that topic to the right side of that share of the chat graphical representation. Figure 2.a. shows a chat with high rhythmicity for all topics – these were debated in parallel as it can be seen by the lack of flat lines near the left side of the representation. The other figure (2.b.) shows the opposite: it has flat lines on the left side of the graphic showing that the topic that they represent has not been debated in those parts of the chat. The conversation starts with a discussion about blogs, while the other topics are ignored. As time passes, these topics get into focus in the detriment of chat, which seems to be forgotten for a while (it is absent in three of the eleven shares of the given chat). The end of the conversation finds all the given topics in focus, as it is desirable, but having long periods of one-topic debate – the topics have been debated in turns which means the participants did not compare them and therefore did not achieve one of the purposes of the conversation.

Test 14 shows the percentage of the passed tests (tests where the obtained grade was above 5) considering the min and max as inferior and superior thresholds, while test 15 represents the average grade obtained by the chat for the 13 tests.

4 Validation

First of all we needed to validate the results obtained with the application and therefore we asked two HCI teachers to evaluate the chats having in mind two main criteria: the quality of the content related to the given concepts and the participants' involvement. Their grades, along with the average values and the scores provided by our application are presented in Table 2.

Table 2. The gold standard values provided for the 6 chats along with the scores provided by our application and with the revised values

Chat	Reviewer 1	Reviewer 2	Average	Application score	Modified app. Score
Chat 1	7.8	7.74	7.77	2.72	7.08
Chat 2	10	9.3	9.65	7.25	10
Chat 3	9	8.9	8.95	3.44	7.8
Chat 4	8.4	8.6	8.5	4.1	8.46
Chat 5	10	9	9.5	5.08	9.44
Chat 6	9	9	9	4.62	8.98

As it can be easily seen, the application's grades are much smaller than the ones provided by the reviewers and therefore we increased these grades by the average of the difference between the reviewers' grades and the scores provided by the application (4.36). The new values are presented in Figure 3 below.

Before modifying the application scores, we had to see how trustworthy were the grades provided by the reviewers and therefore we computed their correlation. This value was 0.8829, which shows that their values are very similar and being domain experts and having experience in teaching, we decided we can trust the values provided. We have also computed the correlation between the reviewers' average grades and the scores provided by the application. The value was 0.8389, very close to the correlation between the reviewers, showing a strong correlation between the application's grades and the real value of the chats.

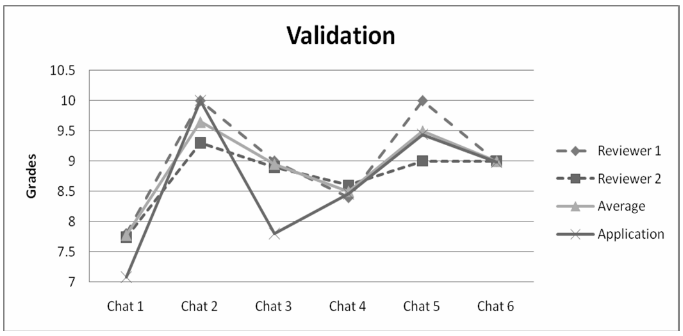


Fig. 3. Application's validation

5 Verification of the Model

We considered two different verification methods for our application. The first one was meant to demonstrate that the model used for a chat conversation depends very much on the number of participants. Therefore, we considered 4 chats consisting of 1250 replies that had between 6 and 8 participants. These chats had the same focus and objectives as the ones used for the application's validation. These chats have been automatically analyzed using our application in order to see whether the model for 4-5 participants could have been also applied for them. The values obtained, along with the thresholds for these chats and for the chats having 4-5 participants are presented in Table 3. The results clearly show that the model for 4-5 participants (represented by the used thresholds) is not adequate for chats with 6-8 participants.

Table 3. Differences between chats with 4-5 participants (chats 1-6) and chats with 6-8 participants

Name of test	Chat 7	Chat 8	Chat 9	Chat 10	Min 6-8	Max 6-8	Min 4-5	Max 4-5
0. Utterance number	138	380	473	259	138	473	176	559
1. Most interested	0.31	0.27	0.21	0.19	0.19	0.31	0.3	0.4
2. Least interested	0.06	0.04	0.07	0.02	0.02	0.07	0.08	0.18
3. Users persistence	0.3	0.02	0.004	0.003	0.003	0.3	0.07	0.31
4. Explicit connections	1.04	1.01	1.01	1.03	1.01	1.04	0.19	0.79
5. Absence	0.12	0.058	0.028	0.134	0.134	0.028	0.011	0.047
6. Minimum activity	1	2	16	1	1	16	6	52
7. Maximum activity	21	137	90	46	21	137	48	345
8. On topic - lex. chain	0.2	0.19	0.29	0.34	0.19	0.34	0.19	0.28
9. Repetitions (min)	0.07	0.03	0.06	0.06	0.03	0.07	0.08	0.15
10. Repetitions (max)	0.13	0.09	0.09	0.13	0.09	0.13	0.12	0.18
11. Least useful user	2.77	3.53	4.24	4.09	2.77	4.09	1.73	2.69
12. Most useful user	3.39	4.96	5.1	4.81	3.39	5.1	2.12	2.95
13. Topic rhythmicity	1.54	1.46	0.51	0.69	0.51	1.46	0.76	1.69

After we have seen that the thresholds do not match, we wanted to verify that we have the same type of chats as previously presented. Consequently, we have modified these chats by considering not the physical participants, but the point of view - "the voice" - that they represent. Therefore, we considered the persons debating the same topics as being a single participant and thus we ended up having again chats with 4 participants debating the same topics as before. These chats have been automatically evaluated and the results showed that they fit well enough in the model with only 4-5 participants, as it can be seen in Table 4. In conclusion, the chats were not different from what we have seen already, but the thresholds were not adequate for them.

The second, and maybe the most important verification method, was tested on three different chats from the same set with the ones used for learning and validation (4 participants debating about chat, forum, blog and wiki), consisting of 911 replies, and asked another teacher of the HCI class to evaluate them in the same fashion as for validation. After that, the chats have been automatically evaluated using our application and the correlation between the reviewer and the application's grades has been computed. The correlation was 0.7933. The values are presented in Table 5.

Table 4. The values obtained for chats 7-10 modified to have 4 participants

Test No.	Mod 7	Mod 8	Mod 9	Mod 10	Min mod	Max mod	Min 6-8	Max 6-8	Min 4-5	Max 4-5
Test 1	0.42	0.35	0.28	0.3	0.28	0.42	0.19	0.31	0.3	0.4
Test 2	0.13	0.13	0.22	0.19	0.13	0.22	0.02	0.07	0.08	0.18
Test 3	0.17	0.1	0.05	0.02	0.02	0.17	0.003	0.3	0.07	0.31
Test 4	1.02	1.01	1	1.01	1	1.02	1.01	1.04	0.19	0.79
Test 5	0.055	0.018	0.009	0.02	0.009	0.055	0.134	0.028	0.011	0.047
Test 6	4	30	106	60	4	106	1	16	6	52
Test 7	41	232	188	115	41	232	21	137	48	345
Test 8	0.2	0.2	0.29	0.34	0.2	0.34	0.19	0.34	0.19	0.28
Test 9	0.13	0.1	0.11	0.17	0.5	1.53	0.03	0.07	0.08	0.15
Test 10	0.18	0.136	0.12	0.19	0.12	0.19	0.09	0.13	0.12	0.18
Test 11	1.88	1.98	2.01	1.91	0.1	0.17	2.77	4.09	1.73	2.69
Test 12	2.14	2.36	2.38	2.14	1.88	2.01	3.39	5.1	2.12	2.95
Test 13	1.53	1.49	0.5	0.68	2.14	2.38	0.51	1.46	0.76	1.69

Table 5. The gold standard values provided for the 3 chats along with the scores computed by our application and with the revised values

Chat	Reviewer	Application	Modified application score
Chat 11	9.627	5.24	9.6
Chat 12	7.574	4.76	9.12
Chat 13	8.777	5.39	9.75

6 Conclusion and Further Work

In this paper we have presented an application that evaluates the quality of a chat according to a number of predefined conversation topics and to the personal involvement of the participants. During the verification, we have shown that the models that should be used to evaluate the chats are dependent on the number of participants: they are different for small (4-5 participants) and medium (6-8 participants) teams, and we expect that these models are also different for 2-3 participants and for more than 8 participants.

The good correlation between the application and the domain experts obtained at both the validation and verification stages recommends it as a reliable application. Also, the large number of tests, gives a lot of flexibility to the user, allowing him/her to give more or less importance to some of the tests and therefore to evaluate exactly the aspects considered to be important.

In the meantime, an evaluator can make a complex analysis of the chats by correlating the results of the different tests, this way identifying the causes that lead to the obtained results and thus being able to take the right decision in the evaluation.

Acknowledgments. This research presented in this paper was supported by project No.264207, ERRIC-Empowering Romanian Research on Intelligent Information Technologies/FP7-REGPOT-2010-1.

References

1. Adams, P.H., Martell, C.H.: Topic detection and extraction in chat. In: Proceedings of the 2008 IEEE International Conference on Semantic Computing, pp. 581–588 (2008)
2. Bakhtin, M.M.: Problems of Dostoevsky's Poetics. Ardis (1993)
3. Bakhtin, M.M.: The Dialogic Imagination: Four Essays. University of Texas Press (1981)
4. Halliday, M.A.K., Hasan, R.: Cohesion in English. Longman, London (1976)
5. Holmer, T., Kienle, A., Wessner, M.: Explicit referencing in learning chats: Needs and acceptance. In: Nejdl, W., Tochtermann, K. (eds.) EC-TEL 2006. LNCS, vol. 4227, pp. 170–184. Springer, Heidelberg (2006)
6. Jurafsky, D., Martin, J.H.: Speech and Language Processing. In: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition, 2nd edn. Pearson Prentice Hall, London (2009)
7. Kontostathis, A., Edwards, L., Bayzick, J., McGhee, I., Leatherman, A., Moore, K.: Comparison of Rule-based to Human Analysis of Chat Logs. In: Conferencia de la Asociación Española para la Inteligencia Artificial (2009)
8. Rebedea, T., Trausan-Matu, S., Chiru, C.-G.: Extraction of Socio-semantic Data from Chat Conversations in Collaborative Learning Communities. In: Dillenbourg, P., Specht, M. (eds.) EC-TEL 2008. LNCS, vol. 5192, pp. 366–377. Springer, Heidelberg (2008)
9. Schmidt, A.P., Stone, T.K.M.: Detection of topic change in IRC chat logs, http://www.trevorstone.org/school/ircsegmentation.pdf
10. Sfard, A.: On reform movement and the limits of mathematical discourse. Mathematical Thinking and Learning 2(3), 157–189 (2000)
11. Tannen, D.: Talking Voices: Repetition, Dialogue, and Imagery in Conversational Discourse. Cambridge University Press, Cambridge (1989)
12. Trausan-Matu, S., Rebedea, T.: A polyphonic model and system for inter-animation analysis in chat conversations with multiple participants. In: Gelbukh, A. (ed.) CICLing 2010. LNCS, vol. 6008, pp. 354–363. Springer, Heidelberg (2010)
13. Voloshinov, V.N.: Marxism and the Philosophy of Language. New York Seminar Press (1973)