# Rhythm analysis in chats using Natural Language Processing

**Irina-Diana Niculescu**
Politehnica University of Bucharest
313 Splaiul Independentei,
Bucharest, Romania
irinaniculescu93@gmail.com

**Ştefan Trăuşan-Matu**
University Politehnica of Bucharest
313 Splaiul Independentei,
Bucharest, Romania
and
Research Institute for Artificial Intelligence
and
Academy of Romanian Scientists
stefan.trausan@cs.pub.ro

**ABSTRACT**
Rhythm is important both in daily speaking and literary texts, as it is a means of expressing the writer's/speaker's feelings and ideas. A good human-computer interface in natural language, both spoken and written should also pay attention to rhythm. This paper presents an analysis of rhythm in chats, based on the analogy between conversations and literature and the importance of repetition. In the case of chats, rhythm characterizes involvement, which can be an useful indicator of a successful collaborative learning session.

**Author Keywords**
rhythm; natural language processing; chats; repetition

**ACM Classification Keywords**
I.2.7 Natural Language Processing: Text analysis.

## INTRODUCTION

This paper presents a system for rhythm analysis in chats. It continues the research made in this domain in the past by the same authors, that is the analysis and comparison of rhythm in poetry, prose, and political speech, by means of three approaches of dividing in rhythmic units: Solomon Marcus approach [1], Boychuck et al. approach [2], Mihai Dinu approach [3], the last of them only for poems. Solomon Marcus approach [1] consists of dividing a text in verses/sentences, and Boychuck et al. approach [2] delimits a text according to punctuation marks, coordinate conjunctions and subordinate conjunctions. Mihai Dinu approach [3] divides each verse in metrical units that contain a single stressed word and the unstressed words that precede it.

In the implementation of these approaches, we considered that only the content words from a text are stressed. For each rhythmic unit obtained, the rhythmic factors proposed by Solomon Marcus [1] are computed: the rhythmic structure, the rhythmic index, the upper and lower rhythmic limit, the rhythmic diameter. Moreover, the rhythmic factors calculated for the entire texts are of utmost importance, as they represent a means of comparing rhythm in different texts.

## STATE OF THE ART

Regarding rhythm in conversations, Deborah Tannen [4] discusses how "repetition, dialogue, and imagery create involvement in discourse, especially conversational discourse" [4, pg 25]. It is analyzed the similarity between the literary texts and conversations, thus "literary writing elaborates strategies that are spontaneous in conversation" [4, pg 30] and "involvement strategies are the basic force in both conversational and literary discourse by means of sound and sense patterns" [4, pg 31]. At the sound level, the participants "become rhythmically involved", while at the sense level, they are "meaningfully, mythically involved" [4, pg 31].

Deborah Tannen analyzes also the importance of repetition in conversations [4, Ch. 3]. With regard to the link between repetition and rhythm, the repetition of words and phrases from a given text represents the unit of measure of the rhythmicity at a lexical and grammatical level [2]. Indeed, repetition creates a certain automatization of the language [4, pg 61]. In order to prove this fact, Tannen gives the example of shadowing: "repeating what is being heard with a split-second delay" [4, pg 93]. It can sometimes occur that speakers repeat unconsciously what they have heard from another speaker, within seconds of expressing the utterance for the first time [4, pg 93]. In this way, the speaker who use repetition tries to participate in the conversation and understand the sent message.

From another perspective, Trausan-Matu showed that repetition of cue words or phrases may give birth to

artifacts that could be the key to solve problems in computer-supported collaborative learning chats [10].

## IMPLEMENTATION

A software application (Figures 1 and 2 present the graphical interface) was developed, which analyzes the rhythm in conversations. The application was tested on 10 chats in English, annotated in a specific XML format [6]. Each conversation has 4-5 participants, who are discussing in approximately 2 hours the advantages and disadvantages of the following forms of communication: blog, chat, wiki, Google Wave, and forum [6]. By analyzing the chats, it can be observed a different degree of participants' involvement, analyzed in terms of the number of utterances and words. The application was written in the Java programming language and has a graphical interface made using Java Swing and Java Applets.

For studying and comparing the rhythm in conversations, we used the following metrics:

- the total number of words;
- the total number of repeated words: each word is reduced to its stem. Two different words with the same stem will be considered a repetition (e.g. the words "user", "using" have the same stem, "us"). Function words are ignored unless they are part of a repeated phrase. An important aspect taken into account in the implementation of the repetition metric is the timestamp, i.e. the time difference between the two words repetition [4, page 64]. Thus, we have considered that two words or phrases are repeated if the distance in words between the two occurrences is less than a constant (30 in experiment), in the case of calculating the metrics for the entire chat. At timestamps/sliding windows, where the number of lines/words is obviously smaller, the repetition of two words disregards the distance between them;
- the number of function words;
- the number of content words;
- the total number of unique function words (ignoring repetition) - calculated at the level of their stems;
- the total number of unique content words (ignoring repetition) - calculated at the level of their stems;
- the total number of unique words (ignoring repetition) - calculated at the level of their stems

The functionalities of the software application are:

1. Dividing chat in timestamps of 5 minutes (about 20 utterances). For each timestamp, the metrics defined above are computed.
2. Dividing chat in sliding windows. Dividing chat in fixed timestamps can lead to the partition of collaboration areas in which repetition plays an important role. To avoid this, every line will be the first line of a sliding window with a duration of 5 minutes. For each chat, it will be obtained a number of sliding windows approximately equal with the number of utterances. For each sliding window are computed the metrics defined above.
3. Finding the most frequent words in chat. Only content words are taken into account, function words being ignored. The stem of each word (obtained with the help of Porter Stemmer algorithm) is a key in a dictionary and its corresponding value is the number of occurrences in the

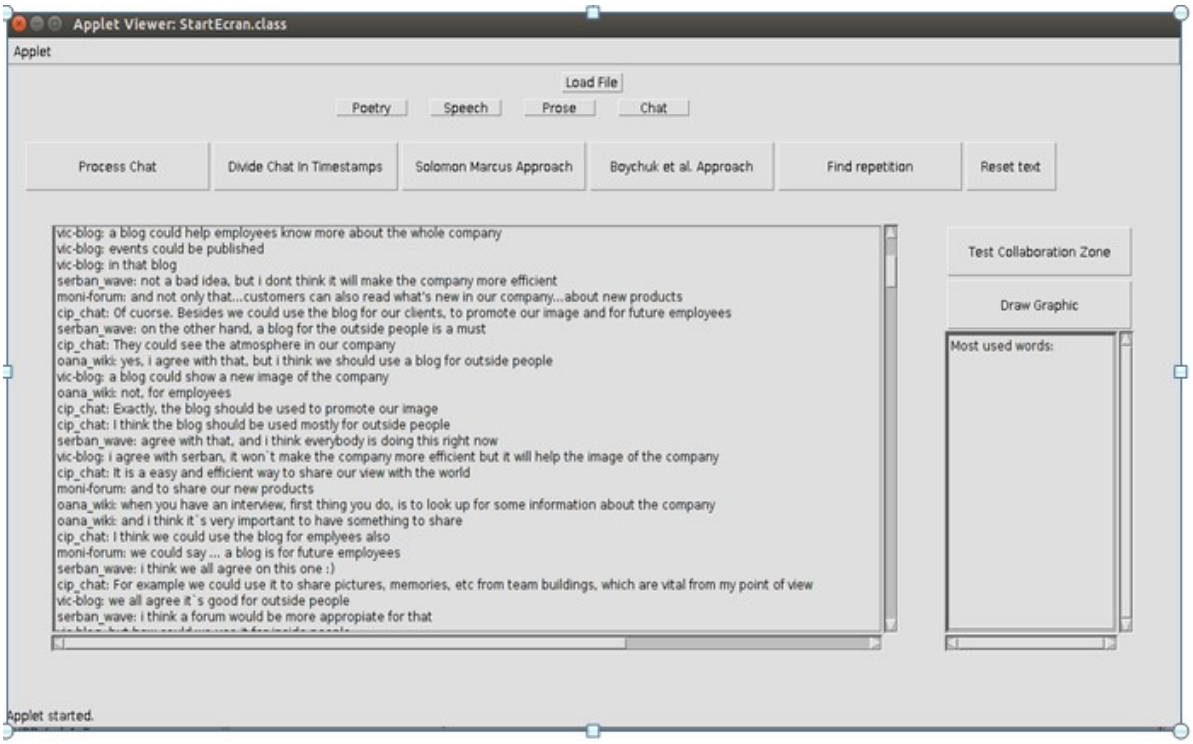


**Figure 1: Graphical user interface of the application**

chat.

| No chat | Start ID | End ID | No. of words | No. of repeated words | Percent of repeated words |
|---|---|---|---|---|---|
| 1 | 1 | 344 | 2832 | 456 | 16.1 |
| 2 | 1 | 296 | 2409 | 576 | 23.91 |
| 3 | 1 | 419 | 4865 | 795 | 16.34 |
| 4 | 1 | 261 | 3314 | 549 | 16.56 |
| 5 | 1 | 430 | 5533 | 673 | 12.16 |
| 6 | 1 | 392 | 2949 | 457 | 15.49 |
| 7 | 1 | 284 | 2246 | 431 | 19.18 |
| 8 | 1 | 397 | 3181 | 542 | 17.03 |
| 9 | 1 | 203 | 1428 | 261 | 18.27 |
| 10 | 1 | 311 | 3201 | 439 | 13.71 |

**Table 1: Metrics calculated for the 10 chats**

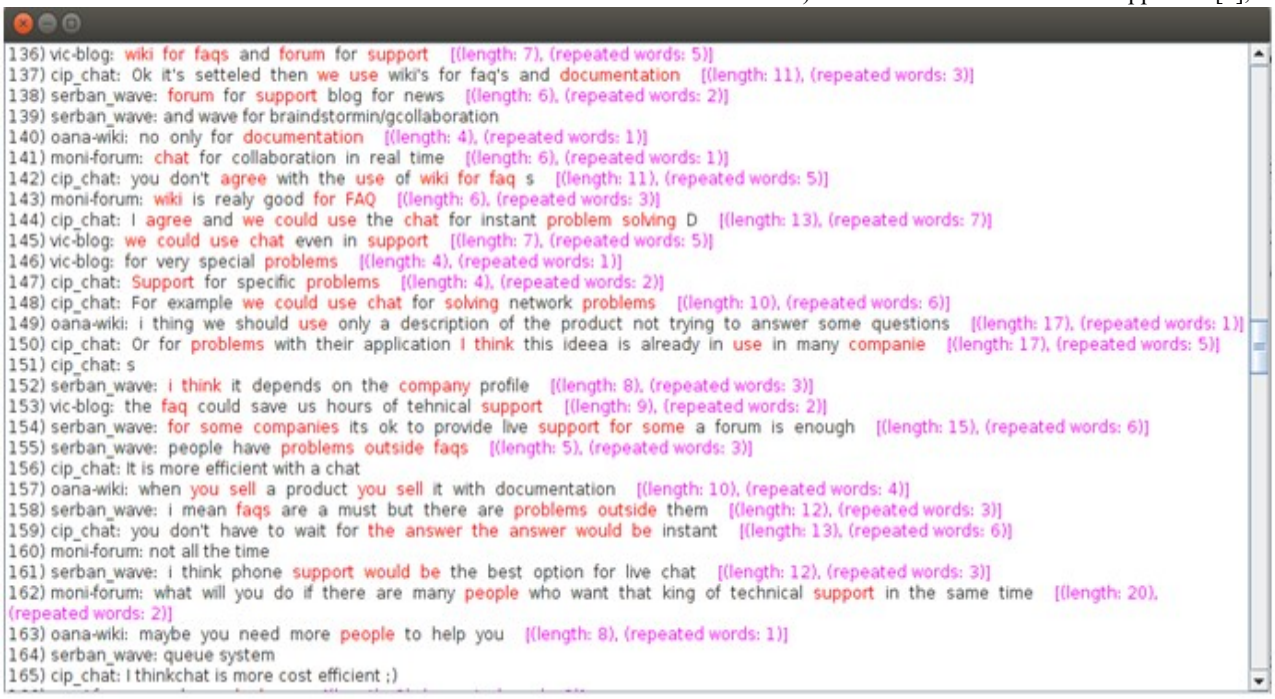


**Figure 2: Highlighting repeated words/phrases in a chat**

4. Applying Solomon Marcus approach [1] and Boychuck et al. approach [2]. Considering the analogy that Deborah Tannen does between literature and conversation [4], the idea to use these two approaches in a chat appeared. In a conversation, more people exchange the role of listener and speaker, so rhythm can be analyzed at the level of the entire conversation (i.e., the utterances of all speakers, taken as a whole) and at the level of each speaker. The rhythm of a conversation can therefore be seen as a sum of the rhythm of all "voices" (in a generalized way [7, 8]) that are present in the conversation. In order to analyze the rhythm of the entire conversation, in mathematical terms, the next steps were followed:

- concatenating all the utterances from the chat to get a text ready to be processed
- dividing the obtained text in rhythmic units according to the approach chosen; in Solomon Marcus approach, the text is divided into sentences and in Boychuk et al. approach the text is delimited according to punctuation marks, coordinate conjunctions, and subordinate conjunctions.
- for each rhythmic unit, the rhythmic factors such as the rhythmic structure, the rhythmic index, the upper rhythmic limit, the lower rhythmic limit, the rhythmic diameter can be calculated using the formulas proposed by Marcus [1]

5. Applying a modified Solomon Marcus approach (at the level of words). In Solomon Marcus standard approach [1], the rhythmic unit is divided into syllables and stressed syllables are found. The rhythmic factors are computed depending on the distance between the accents found in the

unit. To analyze the rhythm of conversation, we made an analogy between a stressed syllable and a stem frequently used in a chat. Thus, the rhythmic factors will be calculated with the same formula, for the most widely used 15 words / stems of the chat. For example, the rhythmic structure is an array, consisting of elements that represent the distance between two repeated instances of the same stem. The length of the rhythmic structure will thus be equal to the number of repetitions of the stem in the chat. A chat is seen as a single rhythmic unit (thus the length of the rhythmic unit is equal to the number of words in chat), and the stem on which this approach is applied is seen as a rhythmic nucleus.

## RESULTS

The results obtained for the ten chats are presented in Table 1. Obviously, the ratio between the number of repeated words and the total number of words is related to the value of the constant (30), which represents the maximum distance between two occurrences of a word in order to be considered as a repetition. Increasing this constant will also proportionally increase the ratio. The number of words, 30, is found in about 2-3 lines and of course, repetition can occur after a larger number of replies. For example, by setting the distance 100, the ratio becomes around 30%, and by setting the distance 200, the ratio varies around 40%. We noticed that for a distance greater than 300, the ratio does not increase significantly, remains around 40%.
It should also be noticed that the function words are ignored when counting repetitions, if they are not part of a repeated phrase such as: function word 1 + function word 2, function word + content word, content word + function word.

By dividing chats into timestamps, it was observed that the first and last utterances are introducing or ending ideas, as it is normal in a conversation, so the number of words is quite small. Both the total number of words (the maximum value obtained exceeds 400 words, but for most of the intervals, it varies between 120-250 words) and the number of repeated words increase in the following timestamps. The ratio between the number of repeated words and the number of words reaches a maximum value of 40%, but the values for most of the intervals vary between 20% - 35%. Table 2 presents the results obtained for this approach for one of the chats used for testing.

| Start ID | End ID | No of words | No of repeated words |
|---|---|---|---|
| 1 | 8 | 6 | 0 |
| 9 | 34 | 86 | 18 |
| 35 | 50 | 131 | 37 |
| 51 | 67 | 127 | 43 |
| 68 | 87 | 201 | 65 |
| 88 | 106 | 196 | 45 |
| 107 | 132 | 292 | 75 |
| 133 | 149 | 203 | 73 |
| 150 | 171 | 241 | 66 |
| 172 | 192 | 225 | 56 |
| 193 | 215 | 183 | 56 |
| 216 | 240 | 171 | 39 |
| 241 | 263 | 216 | 50 |
| 264 | 286 | 176 | 53 |
| 287 | 306 | 210 | 72 |
| 307 | 323 | 139 | 16 |
| 324 | 344 | 29 | 8 |

**Table 2: The results from dividing one of the chats into timestamps**

Dividing a chat in sliding windows (an example of the results obtained for this approach is presented in Figure 3) leads to the observation that the intervals or windows for which higher percentage of repetition are obtained, correspond with the collaborative areas manually annotated [5]. In this paper, the same chats are analyzed than those in a previous one, with different methods [5]. Thus, we can conclude that repetition is most common in the parts of the chat where participants discuss and support their view in a dynamic and enthusiastic way. Moreover, in conversations, participants reuse the same words, instead of synonyms, probably because they were recently heard.

It was also noted that in all the 10 chats used for testing, the number of unique words (calculated at the level of stems) is about half the total number of words in the approach of sliding windows. Figure 4 highlights this observation.

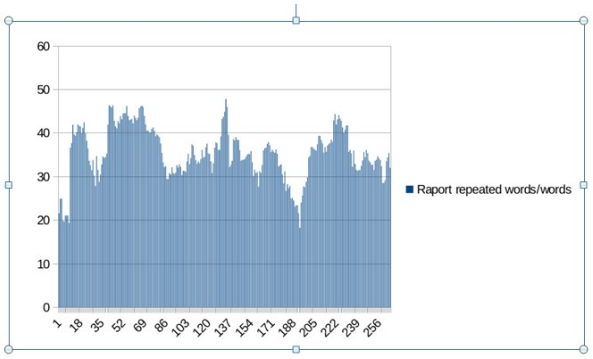


**Figure 3: The ratio between the number of repeated words and the number of words computed for one of the chats, using the approach of sliding windows**

The most repeated words in chats are: "blog", "chat", "us" (the stem for "use", "using", "user", etc.), "think", "agre" (the stem for "agree" and the words derived from it), "wiki", "ye" (the stem for "yes"), "inform", "good", "wave", "forum" etc. The presence of the words "blog", "chat", "wave", "forum", "wiki" list is explained by the fact that the chats are a discussion of these forms of information transmission and communication, so it is normal that these words are used frequently. In this case, we did not take into account the distance between repetitions. Furthermore, the repetition is given by the occurrences of the stems of the words, not by the occurrences of the words themselves. As an example, Figure 5 is a graphic that outlines the frequently used terms in one of the chats used for testing.

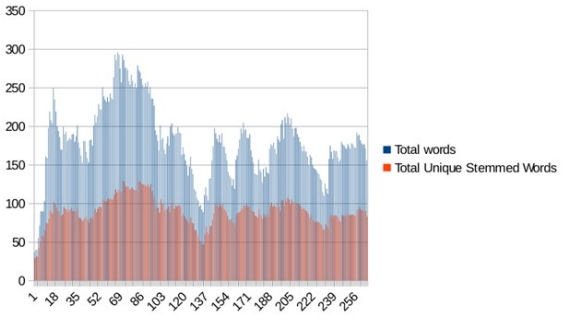


**Figure 4: Diagram illustrating the number of words and the number of unique words (calculated at the level of the stems) in one of the chats, using the approach of sliding windows**

After applying Solomon Marcus and Boychuck et al. approaches, we obtained the following results (Table 3 and Table 4), computed at the level of the entire chat.

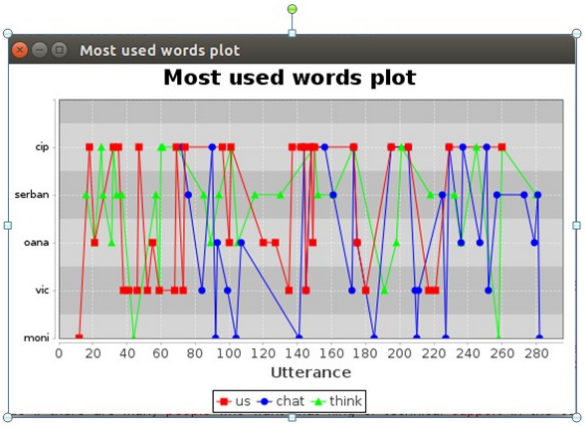


**Figure 5: Graphical representation of frequently used words in one of the chats according to their occurrences in the participants' utterances**

| Chat | The rhythmic index | The rhythmic upper limit | The rhythmic lower limit | The rhythmic diameter | The number of rhythmic units |
|---|---|---|---|---|---|
| 1 | 8 | 14 | 1 | 13 | 443 |
| 2 | 6 | 14 | 1 | 13 | 314 |
| 3 | 8 | 13 | 1 | 12 | 551 |
| 4 | 5 | 13 | 1 | 12 | 360 |
| 5 | 6 | 12 | 1 | 11 | 513 |
| 6 | 7 | 15 | 1 | 13 | 396 |
| 7 | 5 | 11 | 1 | 9 | 294 |
| 8 | 9 | 16 | 1 | 15 | 421 |
| 9 | 5 | 16 | 1 | 14 | 195 |
| 10 | 9 | 16 | 1 | 15 | 356 |

**Table 3: Results for Solomon Marcus's approach, considering the whole chat as a text**

| Chat | The rhythmic index | The rhythmic upper limit | The rhythmic lower limit | The rhythmic diameter | The number of rhythmic units |
|---|---|---|---|---|---|
| 1 | 8 | 14 | 1 | 12 | 626 |
| 2 | 6 | 10 | 1 | 9 | 600 |
| 3 | 8 | 13 | 1 | 12 | 1148 |
| 4 | 5 | 12 | 1 | 11 | 730 |
| 5 | 6 | 12 | 1 | 10 | 1198 |
| 6 | 7 | 12 | 1 | 11 | 731 |
| 7 | 7 | 11 | 1 | 9 | 552 |
| 8 | 8 | 11 | 1 | 9 | 745 |
| 9 | 4 | 10 | 1 | 9 | 314 |
| 10 | 6 | 14 | 1 | 13 | 677 |

**Table 4: Results for Boychuck et al. approach, considering the whole chat as a text**

After applying the modified Solomon Marcus approach, it was observed that the rhythmic index obtained for the most used word (stem) has a value between 60-80. Obviously, as the length of the rhythmic unit is the number of words in the chat, the index increases inversely proportional with the rhythmic length. The lower rhythmic limits for the 15 most used words are less than 10, and in their corresponding rhythmic structures, elements less than 20 can be observed (a proof that repetition occurs at a low timestamp).

## CONCLUSIONS

In order to analyze rhythm in conversations, the research was based on the previously used approaches for prose, poetry and political discourse [9], as well as on the repetition of words [10]. The results obtained for the ten chats show that the areas of significant repetition correspond with collaborative areas. Moreover, it was noted that the results for Solomon Marcus's approach are similar with the results obtained for Boychuck et al approach. The percentage of words repetition vary between 12% - 24% at the level of the entire chats, and the same percentage reaches up to 40% for the sliding windows approach.

In conclusion, natural language processing is helpful in order to analyze texts rhythm, as by automating the process of finding rhythm, comparisons between different texts can be made.

**Acknowledgments.** This research was partially supported by the FP7 2008-212578 LTfLL project.

## REFERENCES

1. Marcus, S., Poetica Matematică, Academy of the Socialist Republic of Romania Publishing House, Bucharest (1970).
2. Boychuk, E., Paramonov, I., Kozhemyakin, N. and Kasatkina, N., Automated Approach for Rhythm Analysis in French, *Proceeding of the 15th Conference of FRUCT Association* (Finnish-Russian University Cooperation in Telecommunications. (2014)
3. Dinu, M., Ritm şi rimă în poezia românească, Cartea Românească Publishing House, Bucharest, (1986).
4. Tannen, D., Talking Voices - Repetition, Dialogue and Imagery in Conversational Discourse, Cambridge University Press, Second Edition (2007).
5. Dascălu, M., Trăuşan-Matu, S., Dessus, P., and McNamara, D.S., Dialogism: A Framework for CSCL and a Signature of Collaboration (2015).
6. Trăuşan-Matu, S., Dascălu, M., Rebedea, T.E., Gartner, A., Corpus de conversaţii multi-participant şi editor pentru adnotarea lui, Revista Română de Interacţiune Om-Calculator 3(1) (2010), 53-64
7. Trausan-Matu, S., Stahl, G., Sarmiento, J., Supporting Polyphonic Collaborative Learning, E-service Journal, vol. 6, nr. 1, Indiana University Press, (2007), pp. 58-74.
8. Trausan-Matu, S., Dascalu, M., Rebedea, T., PolyCAFe—automatic support for the polyphonic analysis of CSCL chats, *International Journal of Computer-Supported Collaborative Learning*, 06/2014; Volume 9(2), Springer, (2014), pp. 127-156
9. Niculescu, I.D., Trausan-Matu, S., Rhythm analysis of texts using Natural Language Processing. In A. Iftene & J. Vanderdonckt (Eds.), *Romanian Conference on Human-Computer Interaction (RoCHI 2016)*, (2016), pp.197-112
10. Trausan-Matu, S., Repetition as Artifact Generation in Polyphonic CSCL Chats, *Third International Conference on Emerging Intelligent Data and Web Technologies*, IEEE Conference Publications, pp. 194-198 (2012)