Activity-based dialogue analysis as evaluation method

Bilyana Martinovski
Ashish Vaswani
Institute for Creative technology
University of Southern California
martinovski@ict.usc.edu
vaswani@ict.usc.edu

Abstract

This paper uses linguistic activity-based dialogue analysis in order to characterize, evaluate and compare activities. We find that human-human and human-machine interaction via radio are equally efficient but offer different styles of learning and instruction.

1. Introduction

How can we recognize speech activities? How does the language and media determine and reflect the speech activity? Which descriptive features are more basic and what story do they tell us about the activity? In this paper we present activity-based dialogue analysis of radio dialogue during Call for Fire (CFF) training. The quantitative analysis is meant to describe the activity but also to give basis for comparisons with other speech activities. We compare instructive radio talk between humans (HH) and between humans and a machine (HM).

2. Method

Studies of automatic systems of communication typically involve a post hoc evaluation by the users in the form of interviews/questionnaires or scored on performance efficiency. In this study, we suggest to observe dialogue styles in order to characterize, compare and evaluate the relation between activities and between human-human and human-machine interaction.

Most studies on language style have concentrated on written language. Biber [1] analyses speech but studies mostly lexical and grammatical features. Allwood [2] and Martinovski [3], for instance, include features specific to spoken language, such as feedback, turn-taking, pause, etc. The method we adopt here, namely activity-based dialogue analysis [4] involves qualitative and quantitative description of the activity.

2.1. Data and Activity

The data consist of thirteen (13) audio recordings of CFF training sessions, which are part of the ICT Spoken Language Corpus, USC. Seven of the recordings represent radio interaction between humans and six represent interaction between human and a machine, which simulates a human. Dialogue involved in interactive accomplishment of a task, such as CFF dialogues, can be seen as a collaborative activity with set moves, which correspond to tasks. The dialogue involves two participants. The forward observer (FO) is one (or more) soldier who is close to the target. He or they are supposed to observe and identify the target and to communicate what is observed and identified to the fire direction center (FDC). The FDC is located further away from the target and closer to the guns, which are far from the target. The FDC assists and directs the actions of the FO and the guns. It also assists and directs the communication between FO and guns, among FOs and among guns. The goal is to fulfill a mission; most often this means to destroy a target.

2.2. Units and Coding Categories

A turn is identical to a radio transmission. Thus one may have turns followed by the same speaker distinguished by the switching off of the radio for more than one second. A qualitative analysis of the data resulted in three (3) coding categories and twenty five (25) subcategories:

Dialogue Moves: identification (ID), target location (TL) and description (TD), message to observer (MTO), engagement, fire, method of fire and control (MFC), shot, splash, rounds complete (RC), check, adjustment, observer coordinates (OBCO), end of mission (EM), intelligence report (INTEL), situation report (SITREP);

Dialogue Regulators (or Keywords): over, out, roger, standby;

Grounding Acts: confirmation, prompt, preparation, correction, say again.

The moves are specific to the activity (see appendix, also [5]). In short, the process involves a preparatory stage of identifications of participants (ID), an initial stage of suggestions/requests for action by the FO (e.g. moves TL, TD, OBCO, SITREP, see appendix), which are considered by the FDC (e.g. INTEL, SITREP) who then (third stage) informs the FO of the actions of the guns (which may follow the FO’s requests or may not, e.g. moves MTO, MFC, SHOT, SPLASH, see appendix); finally the FO reports the end results of the collaborative mission (e.g. EM).

By confirmations we mean mainly feedback giving repetitions of the previous utterance; prompts are initiation eliciting expression such as ‘go ahead, give oboe’; preparations are acts dedicated to preparing for a move, for ex. ‘be prepared for intel’; corrections initiate repair of own or other speech; ‘say again’ is an other-repair initiation. The dialogue regulators listed above are part of the feedback structure of the military speech genre and indicate end of transmission, end of move, confirmation and readiness for initiation of move, respectively.
3. Quantitative analysis

In the following section we observe what story quantitative language analysis has to tell us about the activity style and functions as a whole, in relation to other activities and in relation between HH and HM interaction within one activity. We use simple measures and ranks of parts of speech, moves, acts, turns, pauses as well as more complex measures of liveliness and caution. The HM data are more than the HH.

3.1. Characteristics of the dialogue style

First we will observe the activity as a whole and in comparison to other activities.

3.1.1. Parts of speech

By checking the parts of speech we can notice that this activity is greatly nominalized and it is oriented towards precision, because the most frequent parts of speech are nouns and cardinals. We can conclude that many utterances do not consist of grammatical sentences but of listing of numbers, verbs as nouns and nouns as names. Most of the nouns appear without determiners.

Table 1. Parts of speech (POS) ranks in numbers.

<table>
<thead>
<tr>
<th>POS</th>
<th>Num</th>
<th>POS</th>
<th>Num</th>
</tr>
</thead>
<tbody>
<tr>
<td>noun</td>
<td>2626</td>
<td>pronoun</td>
<td>154</td>
</tr>
<tr>
<td>cardinals</td>
<td>2568</td>
<td>adverb</td>
<td>121</td>
</tr>
<tr>
<td>preposition</td>
<td>1227</td>
<td>‘to’</td>
<td>102</td>
</tr>
<tr>
<td>verb</td>
<td>774</td>
<td>coordin-</td>
<td>54</td>
</tr>
<tr>
<td>determiner</td>
<td>570</td>
<td>ing con-</td>
<td></td>
</tr>
<tr>
<td>adjective</td>
<td>529</td>
<td>junction</td>
<td></td>
</tr>
<tr>
<td>hesitation</td>
<td>194</td>
<td>modality</td>
<td>18</td>
</tr>
<tr>
<td>sounds</td>
<td></td>
<td>existential</td>
<td>5</td>
</tr>
<tr>
<td>Total</td>
<td></td>
<td></td>
<td>8951</td>
</tr>
</tbody>
</table>

Many moves, (e.g. orders, information, and confirmations), are given in an elliptic form and counted as nouns, f.ex. ‘fire’, ‘splash’, ‘shot’ etc. The pronouns are more rear than the hesitation sounds, which indicates the need of precise identification of speakers.

3.1.2. Move and acts

The most frequent moves and acts in the activity are identification, confirmation and dialogue regulators. Identification is used to identify the speaker and the addressee. Confirmations are repetitions of already given information or order which purpose is establishment of common ground. The keywords’ function is to regulate the dialogue i.e. they are part of the grounding efforts. Thus the most dominant function of the activity is exercise of security and reliability of information including credibility of participants of which the establishment of common ground is an essential part. The acts contributing to grounding are in bold in Table 2.

Table 2. Moves and acts ranks for all CCF data.

<table>
<thead>
<tr>
<th>rank</th>
<th>move and num of words</th>
<th>rank</th>
<th>move and num of words</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>id 1984</td>
<td>12</td>
<td>splash 164</td>
</tr>
<tr>
<td>2</td>
<td>confirm 1462</td>
<td>13</td>
<td>shot 144</td>
</tr>
<tr>
<td>3</td>
<td>keyword 1306</td>
<td>14</td>
<td>rc 93</td>
</tr>
<tr>
<td>4</td>
<td>sitrep 617</td>
<td>15</td>
<td>obco 91</td>
</tr>
<tr>
<td>5</td>
<td>tl 590</td>
<td>16</td>
<td>engage 88</td>
</tr>
<tr>
<td>6</td>
<td>intel 560</td>
<td>17</td>
<td>adjustment 52</td>
</tr>
<tr>
<td>7</td>
<td>mto 408</td>
<td>18</td>
<td>say again 52</td>
</tr>
<tr>
<td>8</td>
<td>prompt 383</td>
<td>19</td>
<td>mfc 46</td>
</tr>
<tr>
<td>9</td>
<td>em 376</td>
<td>20</td>
<td>correction 39</td>
</tr>
<tr>
<td>10</td>
<td>td 230</td>
<td>21</td>
<td>prep 33</td>
</tr>
<tr>
<td>11</td>
<td>fire 216</td>
<td>22</td>
<td>check 10</td>
</tr>
<tr>
<td></td>
<td></td>
<td>23</td>
<td>standby 6</td>
</tr>
</tbody>
</table>

36.7% of all words are dedicated to grounding. If we include in the grounding the identification ‘ritual’ of the participants it will dominate the activity. 13.2% of the words are involved in intelligence and situation reports. These are given usually by the FDC and situation reports, which could be given by both participants are also of high frequency higher then moves which realize the desired missions such as shot and splash. This means that there is more effort spent in coordination and preparation of action than on action. We observe though that there are much more prompts than preparations which compensates for the reports. This is explained by the fact that the FDC uses the reports as a medium for explicit instruction of trainees and after such instructions often gives prompts for action.

3.1.3. Cross-activity and cross-linguistics comparison

Can statistical measures indicate something about the functions of the utterances or the predominant type of communicative acts in an activity? Although the correspondence between the function of utterances/activity and the overall impression of the language style has yet to be studied we offer here a short evaluation and characterization of the studied activity by comparing the data across-activities and languages using two complex measures, namely measures of liveliness [6] and caution [7]. The measure of interactive liveliness is created by comparison of simple measures such as stress (StressP-Tok – the percent of stressed words as tokens in relation to the total amount of words as tokens), overlap (ToOVP-Tok – percent of overlapped words as tokens in relation to the total amount of words as tokens), mean length utterance (MLU - the mean of the number of tokens in an utterance) and pauses (PausP-Tok - the percent of pauses as tokens in relation to the total amount of words as tokens). In the caution measure, the more pauses, stressed words, own communication management (OCM e.g. hesitations, self cut-offs) [8] and numerals (cardinals) there is the more cautious the activity is. The applied formulas (see also Appendix for explanation of measures) are:
Interactive liveliness

(2. ToOVPTok) + OvePUT + (2.StressPTok) -
(2.PauPTok) – MLU

Total number of tokens

Interactive caution

(2.Pauses) + Stress + (2.OCM) + (2.FB) + Numerals –
(2.Pronouns) – (2.Overlap)

Total number of tokens

Table 3. Cross-activity comparison of liveliness and caution measures applied to American English and Swedish (http://www.qualitative-research.net/eqs-texte/3-00/3-00allwoodetal-e.htm).

<table>
<thead>
<tr>
<th>Complex measure</th>
<th>Average number of speakers</th>
<th>Liveliness</th>
<th>Caution</th>
</tr>
</thead>
<tbody>
<tr>
<td>All</td>
<td>9</td>
<td>30.1</td>
<td></td>
</tr>
<tr>
<td>Sermon</td>
<td>5</td>
<td>-82.6</td>
<td></td>
</tr>
<tr>
<td>Auction</td>
<td>7</td>
<td>-70.3</td>
<td></td>
</tr>
<tr>
<td>CFF radio training</td>
<td>2</td>
<td>-0.002</td>
<td>1.4</td>
</tr>
<tr>
<td>Courts</td>
<td>6</td>
<td>9.4</td>
<td>-0.1</td>
</tr>
<tr>
<td>Tasking</td>
<td>3</td>
<td>11.3</td>
<td></td>
</tr>
<tr>
<td>Meeting</td>
<td>11</td>
<td>33.9</td>
<td></td>
</tr>
<tr>
<td>Dinner</td>
<td>9</td>
<td>91.2</td>
<td>-0.7</td>
</tr>
</tbody>
</table>

The US CCF army training is far more cautious than Swedish courtroom interrogation and formal dinner and by implication may be interpreted as a more precision-oriented type of verbal activity. Court examinations are more cautious and precise than dinner talk. Caution is related also to the required commitment to the acts and the activity. Thus caution and precision in references (the large number of numerals, names and nouns) is a reflection of this activity’s purpose, namely a secure transfer of precise sensitive information in stressful conditions. This purpose calls for even more caution and precision than the delivery of justice.

With regard to the definition of liveliness the sermon is one of the least lively, it is after all a monologue. The auction, which consists mainly of offers and confirmations, the task-oriented, the court examination and the CCF training activity are much less lively than the dinner conversations. The talk-over-dinner is the most lively activity. Thus we may say that if we imagine the activities as a continuum on a scale of liveliness the auction represents one pole and the dinner the other pole. The CCF training talk style appears to be in the middle of this scale with similarity to the court talk.

We may rather speculatively observe that question-answer fixed activities are less lively but more cautious. As expressed by the complex measures above, the CFF training has much less overlap than court and dinner talk, mainly due to the radio transmission style of interaction and transcription. However, the radio training has a tendency for more stressed tokens and more pauses than the dinner talk, as shown in Table 4.

Table 4. Number of tokens, utterances, overlaps, pauses, and stressed words in the Swedish dinner and court data and American English CFF radio training.

<table>
<thead>
<tr>
<th>Activity</th>
<th>Word</th>
<th>Type</th>
<th>Turn</th>
</tr>
</thead>
<tbody>
<tr>
<td>Dinner</td>
<td>30738</td>
<td>3971</td>
<td>2365</td>
</tr>
<tr>
<td>Court</td>
<td>33409</td>
<td>3667</td>
<td>2045</td>
</tr>
<tr>
<td>CCF</td>
<td>8951</td>
<td>630</td>
<td>896</td>
</tr>
</tbody>
</table>

Note: OvePUT - percent of overlapped words as tokens in relation to the total number of utterances in a group.

StrPTok – percent of emphatic, stressed words as tokens in relation to the total amount of words as tokens in a group.

PauPTok - percent of pauses in relation to the total amount of words as tokens in a group.

In fact, the pauses in the CCF data have much longer duration than the pauses in all the other activities we reflected on. The average duration of a pause in the CFF radio talk is 1 sec, which is much more than the average conversation pause. Sometimes the pause takes more than 2 minutes. If we involve the pause duration in complex measures we might get different results: the CFF training would appear even more cautious and much less lively. Such a result would reflect the fact the CFF activity is mediated by radio whereas the rest are face-to-face. The emphatics may indicate the instruction pathos but may also reflect the higher level of stress (also, impatience, irritation etc.).

3.2. Comparing human-human and human-machine interaction style

The CFF radio training data consist of human-human talk and human-machine talk, the distinction between which we disregarded in the general characterization of the activity as a whole. We observe how the HH and HM interaction styles within the same activity relate.

3.2.1. Moves and acts

The HH CFF radio talk has 615 unique words which is 8.6% of all HH, whereas the HM has 111 i.e. 6.3% of all HM words. Thus despite the greater amount of reports in HH talk (see Table 5), the vocabulary of the HM is not much different but f. ex. the hesitation sounds are more frequent in and thus characteristic of HH talk.
3.2.2. Parts of Speech

With regards to parts of speech realization, pronouns, hesitation sounds and modals are more frequent in HH talk and thus characterize the HH talk.

Table 6. Parts of speech (POS) ranks in number and in percent of own group for HH and HM talk.

<table>
<thead>
<tr>
<th>POS ranks</th>
<th>Human-human</th>
<th>Human-machine</th>
</tr>
</thead>
<tbody>
<tr>
<td>noun</td>
<td>2106 (29.3%)</td>
<td>520 (29.6%)</td>
</tr>
<tr>
<td>cardinal</td>
<td>1905 (26.5%)</td>
<td>663 (37.7%)</td>
</tr>
<tr>
<td>preposition</td>
<td>1007 (14.0%)</td>
<td>220 (12.5%)</td>
</tr>
<tr>
<td>verb</td>
<td>659 (9.2%)</td>
<td>115 (6.5%)</td>
</tr>
<tr>
<td>determiner</td>
<td>490 (6.8%)</td>
<td>80 (4.6%)</td>
</tr>
<tr>
<td>adjective</td>
<td>411 (5.7%)</td>
<td>118 (6.7%)</td>
</tr>
<tr>
<td>hesitation</td>
<td>190 (2.6%)</td>
<td>4 (0.2%)</td>
</tr>
<tr>
<td>sounds</td>
<td>149 (2.1%)</td>
<td>5 (0.3%)</td>
</tr>
<tr>
<td>pronoun</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

The grounding and regulatory moves in HH are 55.2 % of all moves vs. 27.3 % in HM. However, this does not mean that HM is more efficient, both HH and HM have almost the comparable amount of EM reports, 4.1 and 4.8 respectively and similar amount of Shot moves, 1.6 and 1.5 respectively. MTO is more frequent in HM, which means that the radiobot FDC is more active than the human instructor FDC. The human FDC instructs with explanations in INTEL and STREP while the machine FDC trains by action, with much less explanations or discussions. This is most obvious if we compare the frequency of Intel reports in the HH and the HM data: 7.8 % vs. 0 %, respectively. Also, the HH shows more diversity of moves, e.g. it has moves such as check and adjustment, which were absent or less frequent in HM.

The HH uses meta-grounding acts such as prompts to stimulate the trainee to perform an action whereas the HM simply performs the move.

3.2.2. Parts of Speech

With regards to parts of speech realization, pronouns, hesitation sounds and modals are more frequent in HH talk and thus characterize the HH talk.

Table 6. Parts of speech (POS) ranks in number and in percent of own group for HH and HM talk.

<table>
<thead>
<tr>
<th>POS ranks</th>
<th>Human-human</th>
<th>Human-machine</th>
</tr>
</thead>
<tbody>
<tr>
<td>noun</td>
<td>2106 (29.3%)</td>
<td>520 (29.6%)</td>
</tr>
<tr>
<td>cardinal</td>
<td>1905 (26.5%)</td>
<td>663 (37.7%)</td>
</tr>
<tr>
<td>preposition</td>
<td>1007 (14.0%)</td>
<td>220 (12.5%)</td>
</tr>
<tr>
<td>verb</td>
<td>659 (9.2%)</td>
<td>115 (6.5%)</td>
</tr>
<tr>
<td>determiner</td>
<td>490 (6.8%)</td>
<td>80 (4.6%)</td>
</tr>
<tr>
<td>adjective</td>
<td>411 (5.7%)</td>
<td>118 (6.7%)</td>
</tr>
<tr>
<td>hesitation</td>
<td>190 (2.6%)</td>
<td>4 (0.2%)</td>
</tr>
<tr>
<td>sounds</td>
<td>149 (2.1%)</td>
<td>5 (0.3%)</td>
</tr>
<tr>
<td>pronoun</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

The cardinals dominate the HM talk, which indicates that there is greater precision in the HM performance due also to more frequent TL and OBCO moves (see Table 5 above). There is more variation in verb tense forms in the HH talk although and the HM talk is more ‘now’ oriented, which is partly due to infrequent if any reflection and discussion of events.

In the HH radio talk there are preferences for particular nominalized formulations (in bold in Example 1 below) with indefinite determiners in front of ID names, confirmative regulators, STREPs, prompts, or target descriptions (TD):

Example 1.

\[
g91: \text{a roger over , KEYWORDS}
g19: \text{ok . CONFIRM.}
g91: \text{a steel one niner , ID}
g91: \text{this is gator niner one , ID}
g91: \text{ah end of mission , EM}
g91: \text{break , KEYWORDS}
g91: \text{a we'll engage that mission again , SITREP}
g91: \text{over , KEYWORDS}
g91: \text{a go ahead and just send your correction . -}
g91: \text{PROMPT over , KEYWORDS}
g91: \text{a roger , KEYWORDS a distance of a seven}
g91: \text{one zero , a direction of five six four zero , TL}
g91: \text{a one zsu in open TD over . KEYWORDS}
g19: \text{roger , KEYWORDS}
\]

These are specific for the HH radio talk and express the ritualistic repetitive fixed format of the sub-activities, which sound like known moves in a game rather than spontaneous speech.

4. Conclusion

In this short paper, we use an activity-based method for analysis and evaluation of dialogue. We start with qualitative analysis in order to produce coding categories, which we then use in a quantitative linguistic study. We find recognizable and specific characteristics of activities, which could be even richer in an extended version of the study. In comparison to other activities across languages, the CFF radio training is livelier only in comparison to sermons and auctions but more cautious compare to courtroom examinations and dinners. Question-answer fixed activities appear less lively but more cautious. The complex measures of liveliness and caution would be improved if information about the duration of pauses is included. Furthermore, the data show that the HH interactive training relies on explicit explanations, coordination and meta-grounding acts whereas the HM relies on repetition of pre-recorded actions with not much explicit explanation, less speech management and coordination, but also less modal expression and hesitation sounds. Thus
both training modes are equally efficient but differ in style of instruction and learning.

Future work on the topic will involve other speech-feature analysis, such as intonation in relation to function of task moves and grounding.

5. Acknowledgements

"The project or effort described here has been sponsored by the U.S. Army Research, Development, and Engineering Command (RDECOM). Statements and opinions expressed do not necessarily reflect the position or the policy of the United States Government, and no official endorsement should be inferred."

6. References


Appendix

Appendix A Example of a CFF dialogue

Note: FO will be G91 (Gator niner one) and FDC will be S19 (Steel one niner)

Phase I:
G91: steel one niner, this is gator niner one, radio check, over
S19: gator niner one, this is steel one niner, radio check good, over
G91: steel one niner, gator niner one, prepare to copy obco, over
S19: go ahead gator niner one
G91: obco follows, grid one two three five, five six six four, over
S19: roger, I read obco is grid one two three five, five six six four, over
G91: that’s a good copy steel one niner, over
S19: roger that gator niner one
Phase II:
G91: steel one niner, this is gator niner one, fire for effect, over
S19: fire for effect out
G91: grid, three five four four, three six four six, attitude one six hundred over
S19: grid, three five four four, three six four six, attitude one six hundred over
G91: five bmps stationary over
S19: gator niner one, steel one niner, check that grid over
G91: correction, grid, four five four four, three six four six over
S19: correction, grid, four five four four, three six four six out
Phase III:
S19: uh message to observer, mike four rounds, here, over
G91: mike four round here out
S19: target number, alpha bravo, zero zero zero one, over
G91: target number, alpha bravo, zero zero zero one, break, direction six one hundred over
S19: direction six one hundred out
S19: shot, rounds complete over
G91: shot, rounds complete out
S19: splash over
G91: splash out
Phase IV:
G91: end of mission, target neutralized break, one BMP destroyed, four neutralized
S19: roger one bpm destroyed, four neutralized

Appendix B Description of measures

MLU average number of words per utterance
PauPTok percent of pauses in relation to the total amount of words as tokens in a group
StrPTok percent of stressed words as tokens in relation to the total amount of word-tokens in a group
OvePuT percent of overlapped words as tokens in relation to the total number of utterances in a group

Measure of interactive caution:

\[
\frac{(2 \text{Pauses}) + \text{Stress} + (2 \text{OCM}) + (2 \text{FB}) + \text{Numerals} - (2 \text{Pronouns}) - (2 \text{Overlap})}{\text{total number of tokens}}
\]

Measure of interactive caution:

\[
\frac{(2 \text{ToOVPtok}) + \text{OvePuT} + (2 \text{StressPTok}) - (2 \text{PausPTok})}{\text{MLU} \times \text{total number of tokens}}
\]