ANALYSIS OF USER BEHAVIOR WITH MULTIMODAL VIRTUAL CUSTOMER SERVICE AGENTS

*Ian Beaver and Cynthia Freeman**

NextIT Corporation
Spokane Valley, WA USA
ibeaver@nextit.com, cwu@nextit.com

ABSTRACT

We investigate the occurrence of user restatement when there is no apparent error in Intelligent Virtual Assistant (IVA) understanding in a multimodal customer service environment. We define several classes of response medium combinations and use various statistical measures to determine how the combination of medium and linguistic complexity impacts the user's apparent willingness to accept their query result. Through analysis on $3,000$ sessions with a live customer service IVA deployed on an airline company website and mobile application, we discover that as more media are involved in a response, user restatements increase. We also determine which linguistic complexity measures should be minimized for every response class in order to improve user comprehension.

***Index Terms*—** intelligent virtual agents, automated assistants, customer service automation, multimodal

1. INTRODUCTION

With the continued rise of Intelligent Virtual Assistants (IVA) in the customer service domain [1] and analysts predicting that human contact center agents will be altogether replaced by IVAs in the near future [2], discovering means to optimize these human-computer (HC) interactions is necessary. As a company that builds IVAs primarily for customer service agent replacement, we are interested in cases where there is no apparent misunderstanding on the part of the IVA, but the user continues to restate their query. This may happen because the answer was not specific enough, the user did not fully read or understand the response, or the response was presented in a form that the user did not like. In the latter case, as these IVAs are increasingly multimodal [3, 4], we theorize that not only is the formulation of the response important, which is well explored in literature, but so is the medium it is presented on.

In this paper, we explore interactions with an IVA that communicates with customers over embedded live chat on a large company website as well as the company's mobile application. In both cases, the IVA is exposed on multimodal interfaces that use audio, text, images, User Interface (UI) controls, and web content as media. After tagging numerous features in these interactions, we perform statistical analysis to determine clues in how the IVA response can appear acceptable to a reviewer but still fail to satisfy the user. Our contribution is to provide designers of multimodal IVAs guidance for intelligently selecting the media to present information to the user and the linguistic features to consider in order to minimize confusion.

2. DATA AND FEATURE SELECTION

For our analysis, we selected $3,000$ sessions consisting of $28,000$ turns from a large international airline IVA. The IVA interacts with users on the airline's website and mobile application providing general travel advice such as flight status information, baggage and security rules, and even helps with the booking process. This particular assistant was selected as user interactions are a good middle ground between an IR agent, as it must fetch flight status and travel documents, and a dialogue system, as it contains several tasks such as collecting everything needed to book a flight or transfer award miles between accounts.

The input media supported by this agent are voice, text, UI elements, and web page events. Voice service is provided by the speech application programming interfaces available on the mobile device or browser; therefore, we have no access to Automatic Speech Recognition features or original audio. We are simply given the resulting text translation. Example UI elements may be additional links provided by the agent as suggestions of clarifying questions or drop down selection boxes used for tasks like indicating a country code. Web page events may be clicking on a help icon next to text on a webpage which will launch the agent with a query asking for more information on that topic, or clicking navigation links to pages the agent is designed to help with.

The output media include the agent response in both text and audio form using Text To Speech, related topic links that, when clicked, will submit the topic to the agent for additional

*The authors thank Ben Edwards at IBM Research for statistics advise

information, and pushing web pages related to the current topic to the user's browser window.

2.1. Manual Tagging

Due to the size of the dataset and the expense of manual review, the data was divided equally among three reviewers who were given general definitions of each feature they were tagging. To avoid bias, the chosen reviewers were not familiar with this specific IVA or its knowledge base. They were given only the user turn and agent response text for each conversation. Although the URLs to any associated web content was known to us, this data was recorded three years prior to the analysis and there is no way to verify that the web page content available now is what was actually seen by the users at that time.

2.1.1. User Turn Features

The reviewers were asked to identify the sentiment of the user turn as one of *positive*, *neutral*, or *negative*. Although no kappa was available for us, humans have been shown to provide 78% agreement in this task for call center data [5]. The reviewers were also asked to identify if the human appeared to be using sarcasm. We are interested in the usage and effect of sarcasm on HC interactions as automatic sarcasm detection has been shown to be a difficult problem [6, 7, 8]. Although it can be difficult for humans to detect sarcasm without various clues [6, 8, 9], inter-annotator agreement is good when constrained to a single utterance [9]. It has been shown to be used more frequently in computer mediated conversation [10] but was not considered as a specific feature in any of [3, 5, 11, 12]. So, we investigate it here.

The third feature tagged was the presence of *escalation* language. We define an escalation request as any attempt by the user to complete their task with a different party than the IVA. In the DARPA Communicator data [13] used in [11], this feature is not available. The authors also note the lack of this ability may lead to errors persisting longer than they would in a real system such as ours. In [3], the IVAs evaluated were in the domain of personal assistants and not customer service so this feature was also not present. However, in [14], the importance of this feature in discovering communication issues in the customer service domain is covered in detail.

In [14], three types of user-initiated escalations and how to differentiate them by conversational context are given. The first type occurs when users immediately request to speak to a different party; they do not want to use the automated system. For example:

```
Agent: Hello, how can I help you today?
User: can you transfer me to reservations?
```

The second type of escalations occur after the IVA directs the user to contact a different department or service, and, in response, either the user asks to be transferred there or requests contact information. An example of this would be:

```
Agent: To change the name on your ticket, please
       call reservations.
User: can you transfer me to reservations?
```

The final type, which we refer to as class 3, are those conversations where the IVA attempted to resolve an issue for the user, failed to do so, and the user requested an alternative party. The presence of a class 3 escalation in a conversation signals that the user recognizes the IVA is unable to resolve their issue, and they are seeking help from a different party. We instructed the reviewers to tag both the presence of escalation language, and, if present, if it met the definition of a class 3 request.

The final user feature that was manually tagged was if the turn appeared to be a restatement of a previous turn. In [11], this was broken into four categories: repeating command, repeating info, rephrasing info, and reword query. As they only tagged 40 conversations in this manner, and we have $3,000$, we collapsed these four categories into one to indicate if the previous response did not resolve the user's task. As noted in [11], this may occur multiple times within the same conversation, and every query may result in multiple restatements until the user is satisfied or gives up.

2.1.2. Agent Response Features

The reviewers also tagged two features of each response by the IVA. First they considered every input-response pair and determined if the response was acceptable given the user utterance. If not, they tagged the response as a misunderstanding on the part of the IVA. This corresponded to roughly 8 of the 13 *evidence of misunderstanding* features defined by [11] and the five *response-level errors* defined in [12], but once again, due to the volume of data we used a single tag for misunderstanding on the part of the IVA for any reason.

The second response feature was what we call the *response medium class*. We define five classes of response based on which output media are referred to within it. The first class (**C1**) is one that uses only a single medium for the response. This corresponds to responses seen in a typical text or speech only IVA. In the second class (**C2**), the agent provides an answer or partial answer, and, for further reference, displays web content in the browser or mobile application. For example:

"Children age five and older are allowed to travel without an adult companion. Please review this Web page for more information about our Unaccompanied Minor policies."

Notice that although the IVA provides a response, it directs the user to the web content that the IVA pushed to their screen.

The third class (**C3**) is when the agent does not attempt to answer the query but responds by redirecting the user to a different medium, typically web page content, for the user to find the answer. An example would be *"Please see this page for information on the forms of payment that we accept."*. Class four responses (**C4**) consist of response text with links for suggested clarifications or related topics appended.

These links give the user an indication of similar or more specific knowledge the IVA has on the current topic and allows the user to simply click on them instead of formulating clarifications or new queries. Once clicked, the IVA will respond to the query associated with the link. The response text may look similar to this:

"To upgrade to a premium cabin, you can pay a fee or use airline miles by selecting the Upgrade Reservation link when viewing your reservation. More upgrade options may be available when checking in online or at a kiosk. For more information about upgrading on the day of departure, select the link below."

In this instance, the IVA provides a response to a query asking about a seat upgrade, and, in addition, supplies the indication that it has further information they may wish to know if this upgrade is happening on the day of departure. By clicking on the displayed link, which may be titled *Day of Departure Upgrades*, the IVA will then provide them information specific to that scenario.

The final class (**C5**) contains all three media of information. The IVA provides high level information while directing the user to web content containing specifics, and suggests related or alternative queries all at once. An example of this would be:

"You may change or cancel your reservation within 24 hours of purchase without incurring a change fee. If the original fare is not available when you make your change, you will be asked to pay the difference. See this Web page for more details on our 24-hour flexible booking policy. If you purchased your ticket more than 24 hours ago and would like to change or cancel it, select the appropriate link below."

2.2. Automatic Tagging

In order to measure the complexity of the IVA response, we pass the text through the L2 Syntactic Complexity Analyzer (L2SCA). This generates 14 different measures covering length of production units, amounts of coordination, amounts of subordination, degree of phrasal sophistication and overall sentence complexity [15]. Although the L1 Lexical Complexity Analyzer generates more complexity indicators [16], it is also limited to text containing at least 50 words. As we are dealing with microtexts that are typically shorter than 50 words and the L2SCA does not have this limitation, we use it for measuring complexity.

3. RESULTS AND ANALYSIS

We first consider conditional probabilities and the co-occurrence of certain tagged features with misunderstanding and restatement. This proceeds with a statistical test for significance on the difference between syntactical complexities of agent responses depending on whether the next user turn is a restatement. Finally, we look at the response medium classes and their relationship to user restatements.

3.1. Co-occurrence Matrices and Conditional Probabilities

We begin with an analysis of the co-occurrence of features such as the presence of sarcasm or negative sentiment with misunderstanding. A co-occurrence matrix is created on a conversational (Table 1) and turn (Table 2) basis.

For example, in Table 1, there are 639 conversations that consist of a user turn that was misunderstood (**M**) by the agent and also a user turn that was a restatement (**RE**) of some previous user turn in the same conversation. **E3** signifies that the conversation contains a class 3 escalation. **SN** and **SA** represent the presence of negative sentiment and sarcasm, respectively. Note that while 60% of conversations contain at least one restatement, only 35% of those conversations also contained misunderstanding on the part of the IVA.

	M	**E3**	**SN**	**SA**	**RE**
M	1019	215	136	18	639
E3	215	453	82	11	286
SN	136	82	197	21	114
SA	18	11	21	31	18
RE	639	286	115	18	1812

Table 1. Co-occurrence matrix by conversation.

It is also worth investigating the co-occurrence of features on an turn basis. Since we believe that a user is more likely to express frustration or request an escalation *after* an agent misunderstanding has occurred, we consider whether a turn was misunderstood directly before such events.

For example, in Table 2, there are 566 user turns where the previous turn was misunderstood (**PM**) by the agent but the current turn was a restatement of some previous user turn.

	PM	**E3**	**SN**	**SA**	**RE**
PM	1408	135	95	10	566
E3	135	1065	20	4	416
SN	95	20	337	15	86
SA	10	4	15	46	4
RE	566	416	86	4	4111

Table 2. Co-occurrence matrix by turn.

Using Table 2, it is easy to determine the probability of a previous misunderstood turn given that the current turn contains sarcasm, restatement, etc. We determine the conditional probability for every combination of such features. For example, the probability of a previous turn being misunderstood given that the current turn contains sarcasm and a restatement ($p(\mathbf{PM}|\mathbf{SA} \wedge \mathbf{RE})$) is .25. Only probabilities greater than or equal to .2 are listed in Table 3 below.

X	Y	p(X|Y)
PM	**SN**	.28
PM	**SA**	.22
PM	**E3∧SN**	.2
PM	**SN∧SA**	.2
PM	**SA∧RN**	.25
¬ M ∧ NR	**SN**	.21
¬ M ∧ NR	**SN∧SA**	.2

Table 3. Conditional probabilities for misunderstanding and restatement.

We also considered if a user was more likely to restate the current turn in the next turn (**NR**) if his or her utterance was misunderstood by the agent, but this only yielded a probability of .14. However, it is also interesting to note that there are many instances where the turn was *not* misunderstood (**¬ M**) by the agent, but the user repeated his or her request anyway!

This led us to investigating why such restatements would occur. We theorized that the complexity of the agent response and the media in which it is presented may lead to the user restating a request in the next turn.

3.2. Complexity and the Mann-Whitney U Test

We first only considered turns where the previous turn was not a misunderstanding. We then partitioned this dataset to user turns where their next turn is a restatement (**NR**) and user turns where their next turn is not a restatement (**¬ NR**). As mentioned in 2.2, the fourteen measures of syntactic complexity are discussed in detail in [15] and are defined in the Appendix. A Mann-Whitney U test was conducted to determine if **NR** tends to have stochastically greater values than **¬ NR** for each of the 14 features of complexity. We also repeat this test on different media of agent response using the five classes described in 2.1.2. The final row, the average number of words for an agent response (**AW**), is not part of the 14 features but is included to give some intuition on the difference between classes.

For example, consider an agent response belonging to class **C5**. The complexity value for **VP/T** is likely to be higher if the next turn is a restatement than if the next turn was not a restatement (last column, last row of Table 4). But if we instead consider all classes, **VP/T** has the reverse effect (first column, last row of Table 4). See section 4 for a discussion of the implications of these results.

3.3. Complexity and Logistic Regression

The previous analysis indicates that there are significant differences in agent response complexity between turns which lead to a restatement and turns that don't on an individual class basis. Now we wish to consider the effects of all response medium classes on the next turn being a restatement while *controlling* for syntactical complexity. We fit a logistic regression model where the independent variables are the value of a complexity feature and x_i, where $x_i = 1$ if the agent response belongs to medium class i (see Table 5). We only choose to use complexity features that are not entirely composed of a row of F's (regardless of significance) in Table 4. The dependent variable is whether or not the next turn is a restatement. To ensure that our predictors maintain independence, one class (**C1**) is not included in the regression, and it serves as the base class [17]. We use a likelihood-ratio test [18] to test for statistical significance. Significance was present for all features; response class has a significant effect on restatement.

	All	**C1**	**C2**	**C3**	**C4**	**C5**
MLC	F	F^{*}	F	F	T^{**}	F^{***}
MLS	T^{***}	F^{***}	F	F^{*}	T^{**}	F^{**}
MLT	F	F^{***}	F^{**}	F	F	F^{**}
C/S	F^{***}	T^{***}	F^{*}	F	F^{***}	F^{*}
C/T	F	F	F^{***}	F	F^{*}	F^{*}
CT/T	F^{***}	T^{***}	F^{*}	F	T^{***}	T^{***}
DC/C	F^{***}	F	F^{**}	F	T^{***}	T^{***}
DC/T	T^{***}	F	F^{**}	F	T^{***}	T^{***}
CP/C	F^{***}	F^{***}	F^{***}	F^{***}	F^{***}	F^{***}
CP/T	F^{***}	F^{***}	F^{***}	F^{***}	F^{***}	F^{***}
T/S	F^{***}	F^{***}	F^{***}	F^{**}	F^{***}	F
CN/C	F^{***}	F^{***}	T^{***}	F	F^{**}	F
CN/T	T^{***}	T^{**}	F^{**}	F	F	F^{***}
VP/T	F^{***}	F^{**}	F^{***}	T^{***}	T^{***}	T^{***}
AW	38.66	35.89	42.48	20.49	43.28	60.19

Table 4. Mann-Whitney U results comparing **NR** and **¬ NR** using the 14 complexity features in [15] for all data (**All**) and separate classes (**C1-5**). A value of T indicates that the values in **NR** tend to be greater than the values in **¬ NR** for the complexity feature. Otherwise, the value is F. $^{*} : p \leq 0.1$, $^{**} : p \leq 0.05$, and $^{***} : p \leq 0.01$. The final row represents the average number of words for an agent response in that class (**AW**).

3.4. Response Medium and Restatements

In Table 6, we show the distribution and occurrence of response medium class. Notice that the response classes which provide additional links (**C4**,**C5**) appear to have slightly higher probability of being followed by a restatement (see last row) . In addition, the class which uses all three media has the highest probability of users restating their query. Somewhat surprising to us was the almost identical performance between classes on a single medium and those that

	Intercept	Complexity	C2	C3	C4	C5
MLC	-1.0007	$-.0023$	.0451	.0581	.2658	.5330
MLS	-1.3013	.0161	.0744	.0355	.2962	.4584
C/S	-1.2544	.1444	.1015	.0527	.3300	.5452
CT/T	-1.1614	.2888	.1046	.0683	.3164	.4570
DC/C	-1.1233	.3035	.0714	.0912	.2842	.4126
DC/T	-1.0812	.1078	.0634	.0801	.2766	.4670
CN/C	-1.0891	.0564	.0159	.0725	.2296	.5076
CN/T	-1.0953	.0521	.0339	.0728	.2841	.4956
VP/T	-1.1998	.0791	.0802	.0912	.2891	.4802

Table 5. Coefficients for a logistic regression model where the dependent variable is whether or not the next turn is a restatement. The independent variables are the value of a complexity feature and x_i which represents class membership for class i. **C1** is the base class; thus, it is not included in the table. Statistical signifiance (99% CI) was present for all features.

	C1	C2	C3	C4	C5
Unique Responses	241	112	110	127	54
Occurrence overall	7056	1428	1414	2896	1206
Occurrence with **NR**	1961	395	385	918	445
p(**NR**|**Ci**)	0.278	0.277	0.272	0.317	0.369

Table 6. Distribution of IVA responses by medium

redirect to web content in part or whole. Adding web content (**C2**) to responses on a single medium (**C1**) did not appear to change user behavior. However, adding web content (**C5**) to responses already using two media (**C4**) *did* appear to change user behavior.

4. DISCUSSION

In section 3.1, we investigated the manually tagged features and how they relate to both agent misunderstanding and user restatement. Unfortunately, not only was the incidence of sarcasm much lower than we had hoped, but it was also a poor indication of agent misunderstanding. With only 46 occurrences of sarcasm out of $14,000$ user turns, the effort of implementing automatic sarcasm detection in this domain would not be justified. The highest indicator of misunderstanding, negative sentiment, (see row 1 of Table 3) is also not reliable. In fact, no combinations of these features seem to provide any reliable means to detect prior misunderstanding of any type.

The fact that sarcasm, negative sentiment, restatement, and even class 3 escalations are all far more likely when the previous turn was *not* misunderstood seems to indicate that the wording and medium selection of the IVA response are as important as agent understanding. Obviously, without correct understanding, it is impossible to provide correct response. However, correct understanding and even apparently acceptable response does not seem to ensure resolution for the user; users appeared to restate their issue over six times more often when there was *no* apparent misunderstanding on the part of the IVA than when there was (see **RE** column in Table 2). Remember that this determination of misunderstanding is based off of the *response text* which accounts for correct natural language understanding as well as natural language generation. This greatly justifies our work.

In the customer service domain, neither the system nor the user want to spend more effort than necessary on completing their tasks [19]. These restatements on the part of the user are probably not exploratory, where the user is probing the knowledge boundaries of the IVA out of curiosity as is expected with chatterbots. This, instead, indicates that the user's query is not being answered efficiently. This difference could be due to poor response wording which we cannot completely control for. Asking for reviewers to tag for responses that do not appear to answer the user's question has problems of its own; what is clear to a reviewer may not be clear to the original user. However, the difference between restatement with and without prior misunderstanding is so great that this cannot be the only reason. To help explain this difference, we turn our attention to response medium and complexity.

Considering Table 4, we can safely eliminate the five features with a row consisting only of *F* as they appear to have no effect on restatements *regardless of the medium used*. For the remaining features, a statistically significant *T* indicates that the incidence of a high value for the complexity feature may affect the user's comprehension and cause a restatement. Of interest are instances where *T* and *F* appear with high significance in the same row. Agent responses belonging to classes 3,4, and 5 that have a higher value for the complexity feature **VP/T** are more likely to have a restatement on the next user turn. However, this is not the case if we consider all classes. Without detailed analysis on the agent responses, it is hard to determine why this is the case.

Instead, we take the effect of the complexity features on individual response classes by looking at the columns of Table 4. Looking at each class individually, it does appear that

some classes are more sensitive to response complexity than others. Consider **C1**, for instance, that has three significant positive features and seven significant negative features, using $p \leq 0.05$. This implies that the designers of text-only responses need to pay attention to **C/S**, **CT/T**, and **CN/T** as complexity features. Compare this to **C4** where designers of responses using text *and* additional links need to minimize the incidence of six different complexity features.

In Table 6, we see that restatements are more probable with the addition of links (**C1** $\rightarrow$ **C4**) than with the addition of web content (**C1** $\rightarrow$ **C2**). There are a few plausible explanations for this. Because the classes containing link references appear more sensitive to complexity in Table 4 than those without them, these responses may be more confusing to the user. Or, more likely, the suggestions on the part of the IVA alert the user that it has more specific or related information to their query that the user did not originally specify.

This would be the case in a scenario where the user asks "How do I change my seat?", and the IVA responds with instructions on how to modify their seat, but, in addition, includes a link named *Same Day Seat Changes*. This link indicates to the the user that changing their seat on the same day as their flight is a *different* task than what he or she originally asked for. The user then clicks on the link to get information on what was their real task all along; they just didn't know there was a distinction when they stated their query. In this scenario, the restatement is not actually a bad thing since the IVA has informed the user on the difference in task.

The difficulty is separating these scenarios from those where the user truly did not understand or was not satisfied with the response. This may indicate that the higher restatement probabilities in Table 6 for classes with links is artificially inflated, but to what extent is unknown without more granular tagging. However, as mentioned in section 3.4, adding web content to the text-only medium (**C1** $\rightarrow$ **C2**) did not increase probability of restatement but adding the same content to the one containing links (**C4** $\rightarrow$ **C5**) does. Therefore, the increased probability of restatement in **C5** cannot be entirely explained to informing the user of task specialization through additional links.

To further investigate this increase, we compare the classes directly, controlling for complexity features. In Table 5, we see that every class has a significant positive relationship with restatement. However, as links are added (**C4**) the coefficient markedly increases which agrees with the probabilities seen in Table 6 row 4. When all three media are combined (**C5**), the coefficient doubles from **C4** even though the only difference between them is web content, which, as stated before, did not have any effect on **C2**! This clearly shows that controlling for any measured complexity feature **C5** response types will be most likely to produce a restatement.

Length of response does not appear to be a factor in this increase as well since **C3** is over 40% shorter than **C1** on average, which in turn is 15% shorter than **C2**. However, all three have nearly identical probability of being followed by a restatement in Table 6.

5. CONCLUSION AND FUTURE WORKS

As we have shown in our discussion, when designing multimodal IVAs for customer service, it is not as simple as avoiding long responses and choosing words carefully to ensure user comprehension. By using more than two media for a response, the probability of user dissatisfaction in what appears to be an acceptable response rises. While the inclusion of similar topic links can help the user discover specific answers they did not explicitly request, these responses appear more sensitive to complexity. Therefore, once the media of response are chosen, it is necessary to perform complexity analysis on the response text and minimize the use of features that positively correlate to restatements in Table 4.

In the future, we plan to apply these methods to data collected from IVAs in different domains and compare the results to this paper. As pointed out, we do not have visibility into *how* the user restated, so separating the restatements into classes such as in [11] would be informative. At the same time, defining a method to determine when people are clicking links with additional information versus a true restatement would help us better understand the probabilities in Table 6. Deeper linguistic study can be done on the response text to understand what the complexity features are indicating about the response classes.

This paper has revealed several considerations that IVA designers need to take into account when working in a multimodal environment. By limiting the use of concurrent media and doing additional complexity analysis, user experience with multimodal IVAs can be improved.

6. APPENDIX

The linguistic features of complexity as defined in L2SCA [15] are listed below. A T-unit is the smallest word group that can be considered a grammatical sentence regardless of punctuation.

- **MLS**: mean length of sentence
- **MLT**: mean length of T-unit
- **MLC**: mean length of clause
- **C/S**: clauses per sentence
- **VP/T**: verb phrases per T-unit
- **C/T**: clauses per T-unit
- **DC/C**: dependent clauses per clause
- **DC/T**: dependent clauses per T-unit
- **T/S**: T-units per sentence
- **CT/T**: complex T-unit ratio
- **CP/T**: coordinate phrases per T-unit
- **CP/C**: coordinate phrases per clause
- **CN/T**: complex nominals per T-unit
- **CN/C**: complex nominals per clause

7. REFERENCES

[1] Erica Marois, "Using intelligent virtual agents to improve the customer experience: Brains before beauty," ICMI Blog, 2013.

[2] Auxbreak, "The call center industry may come to an end soon according to some experts," Call Center Talk Blog, 2015.

[3] Jiepu Jiang, Ahmed Hassan Awadallah, Rosie Jones, Umut Ozertem, Imed Zitouni, Ranjitha Gurunath Kulkarni, and Omar Zia Khan, "Automatic online evaluation of intelligent assistants," in *Proceedings of the 24th International Conference on World Wide Web*. ACM, 2015, pp. 506–516.

[4] Michael Johnston, John Chen, Patrick Ehlen, Hyuckchul Jung, Jay Lieske, Aarthi Reddy, Ethan Selfridge, Svetlana Stoyanchev, Brant Vasilieff, and Jay Wilpon, "Mva: The multimodal virtual assistant," in *15th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, 2014, p. 257.

[5] Shourya Roy, Ragunathan Mariappan, Sandipan Dandapat, Saurabh Srivastava, Sainyam Galhotra, and Balaji Peddamuthu, "Qa rt: A system for real-time holistic quality assurance for contact center dialogues," in *Thirtieth AAAI Conference on Artificial Intelligence*, 2016.

[6] Konstantin Buschmeier, Philipp Cimiano, and Roman Klinger, "An impact analysis of features in a classification approach to irony detection in product reviews," in *Proceedings of the 5th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*, 2014, pp. 42–49.

[7] Antonio Reyes, Paolo Rosso, and Tony Veale, "A multidimensional approach for detecting irony in twitter," *Language resources and evaluation*, vol. 47, no. 1, pp. 239–268, 2013.

[8] Roberto González-Ibánez, Smaranda Muresan, and Nina Wacholder, "Identifying sarcasm in twitter: a closer look," in *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies: short papers-Volume 2*. Association for Computational Linguistics, 2011, pp. 581–586.

[9] Ellen Riloff, Ashequl Qadir, Prafulla Surve, Lalindra De Silva, Nathan Gilbert, and Ruihong Huang, "Sarcasm as contrast between a positive sentiment and negative situation.," in *EMNLP*, 2013, vol. 13, pp. 704–714.

[10] Jeffrey T Hancock, "Verbal irony use in face-to-face and computer-mediated conversations," *Journal of Language and Social Psychology*, vol. 23, no. 4, pp. 447–463, 2004.

[11] John Aberdeen and Lisa Ferro, "Dialogue patterns and misunderstandings," in *ISCA Tutorial and Research Workshop on Error Handling in Spoken Dialogue Systems*, 2003.

[12] Ryuichiro Higashinaka, Kotaro Funakoshi, Masahiro Araki, Hiroshi Tsukahara, Yuka Kobayashi, and Masahiro Mizukami, "Towards taxonomy of errors in chat-oriented dialogue systems," in *16th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, 2015, p. 87.

[13] Marilyn A Walker, Alexander I Rudnicky, Rashmi Prasad, John S Aberdeen, Elizabeth Owen Bratt, John S Garofolo, Helen Wright Hastie, Audrey N Le, Bryan L Pellom, Alexandros Potamianos, et al., "Darpa communicator: cross-system results for the 2001 evaluation.," in *INTERSPEECH*, 2002.

[14] Ian Beaver and Cynthia Freeman, "Detection of user escalation in human-computer interactions," in *INTERSPEECH*, 2016.

[15] Xiaofei Lu, "Automatic analysis of syntactic complexity in second language writing," *International Journal of Corpus Linguistics*, vol. 15, no. 4, pp. 474–496, 2010.

[16] Xiaofei Lu, "The relationship of lexical richness to the quality of esl learners oral narratives," *The Modern Language Journal*, vol. 96, no. 2, pp. 190–208, 2012.

[17] Christopher J Nachtsheim, John Neter, Michael H Kutner, and William Wasserman, "Applied linear regression models," *McGraw-Hill Irwin*, 2004.

[18] George Casella and Roger L Berger, *Statistical inference*, vol. 2, Duxbury Pacific Grove, CA, 2002.

[19] Emiel Krahmer, Marc Swerts, Mariët Theune, and Mieke Weegels, "Problem spotting in human-machine interaction," 1999.