Predicting Interaction Quality in Customer Service Dialogs

Svetlana Stoyanchev, Soumi Maiti, and Srinivas Bangalore

Abstract In this paper, we apply a dialog evaluation *Interaction Quality (IQ)* framework to human-computer customer service dialogs. IQ framework can be used to predict user satisfaction at an utterance level in a dialog. Such a rating framework is useful for online adaptation of dialog system behavior and increase user engagement through personalization. We annotated a dataset of 120 human-computer dialogs from two customer service application domains with IQ scores. Our inter-annotator agreement ($\rho = 0.72/0.66$) is similar to the agreement observed on the IQ annotations of publicly available bus information corpus. The IQ prediction performance of an *in-domain* SVM model trained on a small set of call center domain dialogs achieves a correlation of ρ=0.53/0.56 measured against the annotated IQ scores. A *generic* model built exclusively on public LEGO data achieves 94%/65% of the in-domain model's performance. An *adapted* model built by extending a public dataset with a small set of dialogs in a target domain achieves 102%/81% of the in-domain model's performance.

1 Introduction

Automated call/chat centers handle thousands of customer service requests daily and require regular procedures to assess quality of the dialog interaction. While using automation for customer service leads to cost savings, it is important to maintain a high quality of interaction as it affects customers perception of the company and the brand. Analysts in automated customer support call centers measure objective

Svetlana Stoyanchev
Interactions LLC., 25 Broadway, New York, NY, USA e-mail: svetlana.stoyanchev@gmail.com

Soumi Maiti
The Graduate Center, 365 5th Ave, New York, NY, USA e-mail: smaiti@gradcenter.cuny.edu

Srinivas Bangalore
Interactions LLC., 25 Broadway, New York, NY, USA e-mail: sbangalore@interactions.com

task success as well as subjective interaction quality. Task success is determined by objective metrics, such as if a customer request was handled appropriately or whether the required information was elicited from the customers. Subjective metrics estimate customers opinion about the quality of the dialog. While an objective task success measure is an important metric for any interface, a subjective dialog evaluation has important long-term implication. A customer's inclination to recommend the system is measured by the Net Promoter Score questionnaires which is widely used by businesses [9].

In this work, we attempt to quantify the subjective user satisfaction of a customer during conversation with an automated spoken customer service dialog system. We apply Interaction Quality framework, first introduced by Schmitt et al. [12]. IQ score, an integer in the range of 1 to 5, is annotated by a labeler on each dialog turn. It has been experimentally shown to correlate with the overall satisfaction of a dialog system user [13]. IQ framework supports prediction of user satisfaction in an ongoing dialog which allows adaptation of dialog behavior to the perceived user satisfaction [16, 17]. In a post-deployment system analysis, IQ prediction on the last turn of a dialog can be used to identify problematic dialogs and infer conditions that lead to decreased user satisfaction [21, 11].

We apply IQ framework in a new domain of customer service dialogs and evaluate generalization of a Support Vector Machine Model trained on the publicly available *LEGO* corpus to the customer service domain. We annotate a dataset of 120 dialogs for two customer service domains (devices and hospitality) using IQ guidelines. Our inter-annotator agreement is similar to the agreement achieved by the annotators of the LEGO corpus, with Weighed Cohen's κ=.54/.63 and Spearman's Rank Correlation ρ=.72/.66 for each of the domains. We evaluate the automatic IQ prediction using Support Vector Machine Model. The performance of an *in-domain* model trained on a small set of the dialogs in the corresponding call center domain achieves ρ=.53/.56. A *generic* model that was built only on public data achieves 94%/65% of the in-domain model's performance. An *adapted* model built by extending a public dataset with a small set of 30 dialogs in a target domain achieves 102%/81% of the in-domain model's performance.

To our knowledge, this is the first application of the IQ framework to a commercially deployed dialog system. Our results indicate that an IQ model trained on a publicly available corpus can be successfully applied and adapted to predict IQ scores in customer service dialogs.

2 Related work

Roy et al. [10] describe a comprehensive analytics tool for evaluating agent behavior in human call centers. In addition to objective measures of system functions, subjective measures of user satisfaction are also used in evaluation of dialog systems. Subjective measures are evaluated using a questionnaire with a set of subjective

Corpus	Num dialogs	Num turns	Avg length	Avg IQ	Avg weighed kappa	Rho
Public *LEGO* dataset						
LEGO1	237	6.3K	26.9	3.5	.54	.72
LEGO2	437	11.1K	25.4	3.9	.58	.72
Proprietary *INTER* dataset						
INTER-D	60	419	6.98	4.3	.54	.72
INTER-H	60	393	6.55	4.4	.62	.66

Table 1: Statistics on the LEGO and INTER data sets.

questions [5, 4]. Net Promoter Score proposes to simplify the measure to a single question of a hypothetical recommendation [9].

A body of research in dialog focuses on automatic estimation of user perception of the dialog quality using objective dialog measures, such as percentages of `timeout`, `rejection`, `help`, `cancel`, and `barge-in` [20, 1]. Predicting subjective ratings assigned by the actual user requires a set up where users rate the system. However, such an evaluation is not always feasible with real users of commercial systems. Evanini et al. [3] collect labels from external listeners, not the callers themselves, and show high degree of correlation among several human annotators and automatic predictors.

Interaction Quality framework has been validated in user studies showing that IQ scores assigned by expert annotators correlate with user's ratings [13, 12]. Support Vector Machines (SVM) classification of IQ score was shown to be the most effective method for predicting IQ. Sequence methods have also been evaluated but did not outperform SVM [18]. Using a Recurrent Neural Network for prediction of IQ shows promising results [7].

3 Data

In our study we use two data sets: a publicly available *LEGO* dataset and a proprietary *INTER* dataset (see Table 1). *LEGO* dataset consists of the logs and extracted features from the *Let's Go!* bus information dialog system annotated with Interaction Quality [8, 14]. *LEGO1* contains 237 dialogs and is a subset of *LEGO2* which contains 437 dialogs. *INTER* dataset consists of logs from deployed *Interactions LLC* customer support dialog systems implemented with a knowledge-based dialog manager, statistical speech recognition (ASR) and natural language understanding (NLU) components. The Interactions dialog systems use a human-in-the-loop approach: when the NLU's confidence is low, human agent performs NLU by listening to an utterance. The dialog starts with a system's generic *'How may i help you?'* question. Once the reason for the call is established, the system proceeds to collect domain-specific details, including dates, names, or account and phone numbers. Some of the calls are fulfilled within the automatic system while others are

forwarded, together with the collected information, to a human agent. In our experiments, we use only the human-computer portion of the customer service dialogs.

From the Interactions data set, we choose two customer support domains: devices (*INTER-D*) and hospitality (*INTER-H*). For each domain, we randomly select 60 dialogs with the lengths ranging between 5 and 10 turns. Two labelers annotated IQ on each dialog turn according to the same guidelines as those used for annotating *LEGO* corpus [14]. Annotations are performed with an in-house implemented iPython notebook interface [6]. We measure agreement using Weighted Cohen's Kappa and Spearman's Rank Correlation [2, 15]. Both of these measures take into account ordinal nature of the scores reducing the discount of disagreements the smaller the difference is between two ratings. To evaluate inter-annotator-agreement, twenty of the *INTER* dialogs in each domain are annotated by both annotators. We observe similar agreement on the *INTER* and *LEGO* datasets with κ between .54 and .62. and ρ between .66 and .72.

Cohen's kappa is a statistical measure for inter-annotator agreement. It measures agreement between two annotators for categorical items.

$$\kappa = \frac{\rho_o - \rho_e}{1 - \rho_e} \tag{1}$$

Where ρ_o is the relative observed agreement between raters and ρ_e is the probability of chance agreement.

Spearman's rho is used to measure rank correlation between two variables.

$$\rho = \frac{\sum_i (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{(\sum_i (x_i - \bar{x}))^2 (\sum_i (y_i - \bar{y}))^2}} \tag{2}$$

Where x_i and y_i are corresponding ranks and $\bar{x}$, $\bar{y}$ are the mean ranks.

4 Experiments

4.1 Features

The features used in IQ prediction include the features from the automatic speech recognizer (ASR), dialog manager state (DM), user utterance modality, duration, and text/NLU (see Table 2). In our experiments, we use a set of features that may be automatically extracted from system log (AUTO), a subset of the features distributed with the *LEGO* corpus [14]. The AUTO features excluding text, NLU, and dialog state are GENERIC and also transferable across domains.[1] We automatically extract the subset of generic features from the *INTER* corpus. We scale all numeric features to have mean 0 and variance 1 using *sklearn.preprocessing.StandardScaler* with default settings.

[1] *barge-in* feature is GENERIC but not recorded in the *INTER* dataset.

Feature set	Features
ASR (utt/total/mean/window)	**ASR success?, ASR failure?, timeout?, reject?, ASR score**, barge-in?
DM (utt/total/mean/window)	**reprompt?, confirm?, acknowledge?**
DM (this utt)	**prompt type (request/ack/confirm), role, loop**, DM-state
Modality	**voice?, dtmf?, unexpected?**
Duration	**words per utt, utt duration, turn number, dialog-duration**
Text/NLU	system-prompt, user-utt, semantic-parse

Table 2: Feature set from the *LEGO* corpus. Binary features are marked with *?*. Generic features also available in the *INTER* corpus are shown in **bold**.

4.2 Method

In our experiments, we use SVM classification which has been shown to outperform sequence models in the IQ prediction task [19].[2] For the evaluation metrics, we follow [19] and report Unweighed Average Recall (UAR), linearly weighted Cohen's Kappa (w-κ), and Spearman's Rank coefficient ρ.

Unweighted Average Recall is defined as the sum of class wise recalls r_c divided by the number of classes $|C|$.

$$UAR = \frac{1}{|C|} \sum_{c \in C} r_c \tag{3}$$

Recall r_c for each class c is defined as,

$$r_c = \frac{1}{|R_c|} \sum_{i=1}^{R_c} \delta_{h_i r_i} \tag{4}$$

Where δ is the Kronecker-delta, h_i and r_i are the i^{th} pair of hypothesis and reference. $|R_c|$ is the total number of ratings per class c.

For the evaluation on *LEGO* data, we conduct a 10-fold cross validation by splitting the data set into training and test sets on dialog level. For the evaluation on *INTER* data, we consider three types of conditions: CROSS, ADAPT, and DOMAIN. For the CROSS condition, we train the model on *LEGO*1 data and evaluate on all 60 dialogs from the test corpus. For the ADAPT condition, we run a 10-fold validation experiment by selecting part of the *INTER* data for training and interpolating it with *LEGO*1. Each fold is tested on the remaining *INTER* dialogs. For the ADAPT20 condition we use 20% (12) *INTER* dialogs and for the ADAPT50 condition, we use 50% (30) *INTER* dialogs. For the DOMAIN condition, we perform a 10-fold cross-validation on the full 60 dialogs of the *INTER* data.[3]

[2] We use *linearSVC* from the the *sklearn* package with the default parameters.

[3] We report the results on *INTER* corpus using LEGO1 for training as it achieved higher scores than the models trained on LEGO2.

Features (condition)	Train	Test	UAR	W-κ	ρ	%
Evaluation on *LEGO* data						
AUTO	LEGO1	LEGO1	.50	.61	.78	-
AUTO w/o text	LEGO1	LEGO1	.49	.61	.77	-
GENERIC	LEGO1	LEGO1	.49	.61	.77	-
AUTO	LEGO2	LEGO2	.47	.55	.70	-
AUTO w/o text	LEGO2	LEGO2	.45	.52	.66	-
GENERIC	LEGO2	LEGO2	.44	.48	.61	-
Evaluation on *INTER* data.						
GENERIC (DOMAIN)	INTER-D	INTER-D	.42	.38	.53	100%
GENERIC (CROSS)	LEGO1	INTER-D	.42	.30	.50	94%
GENERIC (ADAPT20)	LEGO1+20%INTER-D	INTER-D	.38	.31	.49	92%
GENERIC (ADAPT50)	LEGO1+50%INTER-D	INTER-D	.36	.35	.54	102%
GENERIC (DOMAIN)	INTER-H	INTER-H	.45	.38	.57	100%
GENERIC (CROSS)	LEGO1	INTER-H	.40	.26	.37	65%
GENERIC (ADAPT20)	LEGO1+20%INTER-H	INTER-H	.38	.31	.42	74%
GENERIC (ADAPT50)	LEGO1+50%INTER-H	INTER-H	.41	.37	.46	81%

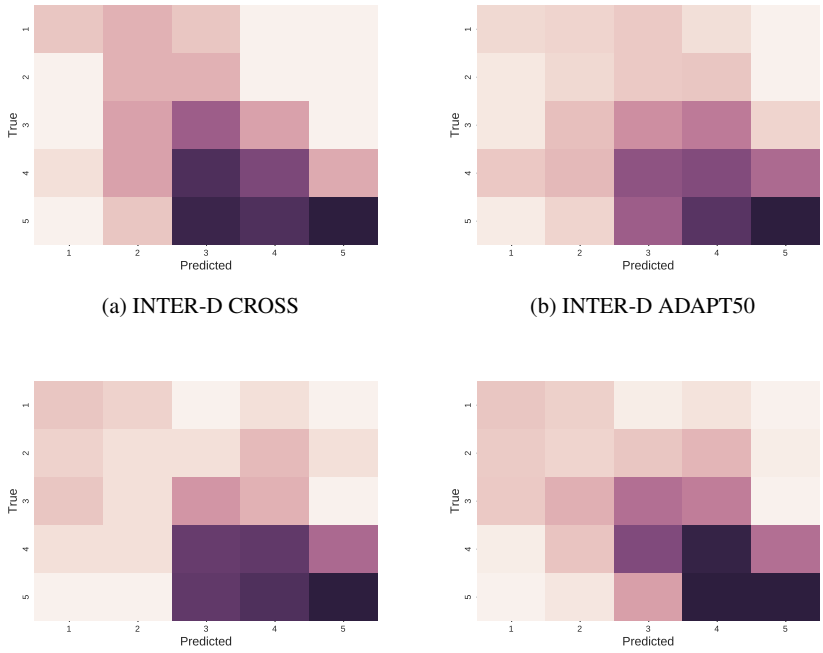
Table 3: Results of SVM classification on *LEGO* and *INTER* dataset: Unweighed Average Recall (*UAR*), Weighed Cohen's Kappa (w-κ), and Spearman's Rank Correlation (ρ). The last column shows the % of the ρ in the CROSS and the ADAPT conditions in relation to the DOMAIN condition.

4.3 Results

Table 3 shows the results of an SVM classifier predicting IQ score on *LEGO* and *INTER* datasets described in Section 3. With the AUTO features on *LEGO*1, SVM classifier achieves $UAR = .50$ w-$\kappa = .61$ and $\rho = .78$.[4] Performance of an SVM classifier with the AUTO features on *LEGO*2 is lower than on *LEGO*1: $UAR = .47$ w-$\kappa = .55$ and $\rho = .70$. We observe that removing non-generic features results in a small drop in performance on *LEGO*1 (ρ drops from .78 to .77). However, the drop is larger on *LEGO*2 when non-generic features are removed (ρ drops from .70 to .61). *LEGO*2 is a superset of *LEGO*1. Although all *LEGO* data originates from the same domain of bus information, part of *LEGO*2 was collected at a later time with a potentially different system components affecting homogeneity of features, such as ASR confidence scores or dialog manager's logic. These results are consistent with the previous work that showed that cross-train-testing on *LEGO*1 and (*LEGO*2 − *LEGO*1) yields a drop in IQ prediction performance in comparison to using training and testing on *LEGO*1.

Next, we evaluate the IQ prediction performance on the customer service dialogs in two domains: devices (INTER-D) and hospitality (INTER-H). In the DOMAIN condition, where the classifier is trained and tested on the data from the same domain, the classification achieves $UAR = .42$, w-$\kappa = .38$ and $\rho = .53$ on the domain INTER-D and $UAR = .45$, w-$\kappa = .38$ and $\rho = .57$ on the domain H. In

[4] This result is a comparable to the result in [19] on *LEGO* corpus.



(a) INTER-D CROSS

(b) INTER-D ADAPT50

(c) INTER-H CROSS

(d) INTER-H ADAPT50

Fig. 1: Error analysis with confusion matrix

the CROSS condition, the classifier achieves $UAR = .42$, w-$\kappa = .30$ and $\rho = .50$ on the domain D and $UAR = .40$, w-$\kappa = .26$ and $\rho = .37$ on the domain INTER-H. The performance for both domains in the CROSS condition is lower than in the DOMAIN condition. Next, we evaluate the ADAPT conditions. Interpolating a model with domain-specific data consistently yields an improvement in w-κ: $ADAPT50 > ADAPT20 > CROSS$ but not in the UAR measure. w-κ and ρ metrics account for the ordinal nature of the IQ class by penalizing less smaller error. Hence the results suggest that with the addition of the in-domain training data, the classification results is *closer* to the human ranking but does not always yield more exact matches of scores.

4.4 Error Analysis

We analyze the errors made by a classifier on the INTER dataset. Figure 1 shows a heat map of true and predicted IQ scores for the CROSS and ADAPT-50 condi-

INTER-D							
True(row)/*Pred(col)*	*1*	*2*	*3*	*4*	*5*	**total**	**Rec**
1	3	5	3	0	0	11	.27
2	0	5	5	0	0	10	0.5
3	0	7	25	7	0	39	0.64
4	1	7	65	38	6	117	0.32
5	0	3	80	64	95	241	0.39
prec	0.75	0.19	0.14	0.35	0.94	408	0.43
INTER-H							
True(row)/*Pred(col)*	*1*	*2*	*3*	*4*	*5*	**total**	**Rec**
1	3	2	0	1	0	6	0.5
2	2	1	1	4	1	9	0.11
3	3	1	9	5	0	18	0.5
4	1	1	52	56	21	131	0.43
5	0	0	56	69	104	229	0.45
prec	0.33	0.2	0.08	0.41	0.82	387	0.40

Table 4: Error analysis confusing matrix for the cross-domain experimental condition CROSS.

tions.[5] We observe that the most weight is concentrated on the higher true scores $(4,5)$ as the dataset is skewed towards the higher scores. In all four heat maps, the weight is concentrated around the diagonal indicating that the prediction error tends to be within a range of +/-2 points. However, majority of the weight is not *on* the diagonal reflecting a low *UAR* which does not take error size into account.

In the CROSS condition the classifier tends to assign lower scores for the turns labeled as 4 and 5, mislabeling them as 3 and 4. This appears to be corrected in the ADAPT-50 condition where the weight shifts closer to the diagonal. For the INTER-H dataset, in the ADAPT50 condition we observe a weight shift towards the diagonal across all scores. For the INTER-D dataset, however, the lower scores (1,2) are misclassified more frequently as (3,4).

We note that the *INTER* dataset is small and highly skewed towards higher scores with the average IQ of 4.2/4.3. Both INTER-D and INTER-H contain very few examples with lower IQ scores. Table 4 shows a confusion matrix with *precision* and *recall* scores for each label for the CROSS condition. More annotated data with lower IQ labels is needed to validate the performance on turns labeled with lower IQ scores.

4.5 Feature Analysis

In this section we explore the relationship between different features and Interaction Quality. We fit a linear regression model using the numeric features of the INTER-

[5] The heat map is drawn on a logarithmic scale.

INTER-D		INTER-H	
Weights	Features	Weights	Features
-5.8245	**#reprompt**	+9.6005	**%ASR success**
+5.5439	**%ASR success**	-5.2867	**Mean ASR score**
+2.9553	**%ASR Timeout| reject**	+ 5.0647	**%ASR Timeout| reject**
+2.9553	**%ASR reject**	+5.0647	**%ASR reject**
+2.1197	Mean ASR score (w)	-4.8677	**#reprompt**
-2.0003	**Mean ASR score**	+2.1534	**%reprompt**
+1.7735	**%reprompt**	-1.6206	user turn number
-1.2341	**% unexpected modality**	-1.5383	unexpected modality(w)
-0.8767	ASR reject (w)	+1.0256	dialog duration
-0.8767	ASR timeout/reject (w)	+0.8635	**% unexpected modality**

Table 5: Feature analysis with linear regression

D and INTER-H datasets. The weights of the linear model allows us to measure the impact of each feature on IQ prediction. To explore numeric features within INTER dataset, we use all 60 domain specific dialogs to train a linear model for each domain. In Table 5 we list the top 10 numeric features and their weights identified by the linear models from each domain. The common top features are marked as **bold**. The feature names with (w) implies the feature was calculated on a window of previous utterances(window size = 3).

We noted that seven features(#reprompt, %ASR success, %ASR Timeout| reject, %ASR reject, Mean ASR score, %reprompt and %unexpected modality) are assigned a top-10 score by a linear model in both datasets. We found #reprompt to be most important feature in INTER-D. The negative sign to weight -5.8245 further denotes that more re-prompt in dialogs results in a lower IQ. %ASR success is the second most important features in INTER-D. This feature is also the most important feature in INTER-H. The feature has positive weight in both dataset indicating that the dialogs with higher %ASR success have higher IQ. We also find the window features are very important for INTER-D. Mean ASR score in last 3 dialogs increase the predicted IQ score. More ASR reject or timeout in previous 3 turns decrease the predicted IQ. For the INTER-H dataset we see that the higher user turn number affects IQ negatively indicating that lower score is more likely to appear later in dialog.

5 Conclusions and Future Work

In this work, we apply *Interaction Quality* dialog evaluation framework to predict user satisfaction in human-computer customer service dialogs. We annotate 120 dialogs from two deployed customer service applications and use them to evaluate IQ prediction performance. The inter-annotator agreement Weighed Cohen's κ of .54/.63 and Spearman's Rank Correlation ρ of .72/.66 for each of the datasets indicate that IQ model can be applied in customer service domain. The performance

of an in-domain model trained only on the call center domain achieves ρ=.53/.56. A generic model built only on public data achieves 94%/65% of the in-domain performance. A generic model built by extending a public dataset with a small set of 30 dialogs in a target domain achieves 102%/81% of the in-domain performance. The results of the cross-domain evaluation show that a model built on a publicly available LEGO corpus can be directly applied to customer service dialogs. Further adaptation to the domain yields an improved performance.

In the future work, we will further analyze the features used in predicting IQ and their effect on domain adaptation. We will apply Recurrent Neural Network model to predict *change* in IQ score (UP/SAME/DOWN) using a distributed representation that captures subjectivity of the task and diverging views of the annotators.

References

1. Beringer, N., Kartal, U., Louka, K., Schiel, F., Türk, U., et al.: Promise – a procedure for multimodal interactive system evaluation. In: Multimodal Resources and Multimodal Systems Evaluation Workshop Program Saturday, June 1, 2002, p. 14 (2002)
2. Cohen, J.: Weighted kappa: Nominal scale agreement with provision for scaled disagreement or partial credit. Psychological Bulletin **70**, 213–220 (1968)
3. Evanini, K., Hunter, P., Liscombe, J., Suendermann, D., Dayanidhi, K., Pieraccini, R.: Caller experience: a method for evaluating dialog systems and its automatic prediction. In: Spoken Language Technology Workshop, 2008. SLT 2008. IEEE, pp. 129–132. IEEE (2008)
4. Hartikainen, M., Salonen, E.P., Turunen, M.: Subjective evaluation of spoken dialogue systems using SERVQUAL method. In: INTERSPEECH (2004)
5. Hone, K.S., Graham, R.: Towards a tool for the subjective assessment of speech system interfaces (SASSI). Nat. Lang. Eng. **6**(3-4), 287–303 (2000)
6. Pérez, F., Granger, B.E.: IPython: a system for interactive scientific computing. Computing in Science and Engineering **9**(3), 21–29 (2007)
7. Pragst, L., Ultes, S., Minker, W.: Recurrent Neural Network Interaction Quality Estimation, pp. 381–393. Springer Singapore, Singapore (2017)
8. Raux, A., Langner, B., Black, A., Eskenazi, M.: Let's Go Public! Taking a Spoken Dialog System to the Real World. In: Proceedings of Eurospeech (2005)
9. Reichheld, F.F.: The one number you need to grow. Harvard business review **81(12)**, 46–54 (2004)
10. Roy, S., Mariappan, R., Dandapat, S., Srivastava, S., Galhotra, S., Peddamuthu, B.: Qa rt: A system for real-time holistic quality assurance for contact center dialogues. In: Thirtieth AAAI Conference on Artificial Intelligence (2016)
11. Schmitt, A., Hank, C., Liscombe, J.: Detecting problematic dialogs with automated agents. In: Proceedings of the 4th IEEE Tutorial and Research Workshop on Perception and Interactive Technologies for Speech-Based Systems: Perception in Multimodal Dialogue Systems, pp. 72–80. Springer-Verlag, Berlin, Heidelberg (2008)
12. Schmitt, A., Schatz, B., Minker, W.: Modeling and predicting quality in spoken human-computer interaction. In: Proceedings of the SIGDIAL 2011 Conference, pp. 173–184. Association for Computational Linguistics (2011)
13. Schmitt, A., Ultes, S.: Interaction quality: Assessing the quality of ongoing spoken dialog interaction by experts - and how it relates to user satisfaction. Speech Communication **74**, 12–36 (2015)
14. Schmitt, A., Ultes, S., Minker, W.: A parameterized and annotated spoken dialog corpus of the cmu let's go bus information system. In: Proceedings of the Eight International Conference on

Language Resources and Evaluation (LREC'12). European Language Resources Association (ELRA), Istanbul, Turkey (2012)
15. Spearman, C.: The proof and measurement of association between two things. American Journal of Psychology **15**, 88–103 (1904)
16. Suendermann, D., Liscombe, J., Pieraccini, R.: Minimally invasive surgery for spoken dialog systems. In: INTERSPEECH, pp. 98–101 (2010)
17. Ultes, S., Kraus, M., Schmitt, A., Minker, W.: Quality-adaptive spoken dialogue initiative selection and implications on reward modelling. In: Proceedings of the 16th Annual Meeting of the Special Interest Group on Discourse and Dialogue, pp. 374–383. Association for Computational Linguistics, Prague, Czech Republic (2015)
18. Ultes, S., Minker, W.: Interaction quality estimation in spoken dialogue systems using hybrid-HMMs. In: Proceedings of the SIGDIAL 2014 Conference, The 15th Annual Meeting of the Special Interest Group on Discourse and Dialogue, 18-20 June 2014, Philadelphia, PA, USA, pp. 208–217 (2014)
19. Ultes, S., Sánchez, M.J.P., Schmitt, A., Minker, W.: Analysis of an extended interaction quality corpus. In: Natural Language Dialog Systems and Intelligent Assistants, pp. 41–52. Springer (2015)
20. Walker, M., Kamm, C., Litman, D.: Towards developing general models of usability with paradise. Natural Language Engineering **6**(3&4), 363–377 (2000)
21. Walker, M.A., Langkilde-Geary, I., Hastie, H.W., Wright, J.H., Gorin, A.: Automatically training a problematic dialogue predictor for a spoken dialogue system. J. Artif. Int. Res. **16**(1), 293–319 (2002)