Analysis of an Extended Interaction Quality Corpus

Stefan Ultes, María Jesús Platero Sánchez, Alexander Schmitt, and Wolfgang Minker

Abstract The Interaction Quality paradigm has been suggested as evaluation method for Spoken Dialogue Systems and several experiments based on the LEGO corpus have shown its suitability. However, the corpus size was rather limited resulting in insufficient data for some mathematical models. Hence, we present an extension to the LEGO corpus. We validate the annotation process and further show that applying SVM estimation results in similar performance on the original, the new and the combined data. Finally, we test previous statements about applying a Conditioned Hidden Markov Model or Rule Induction classification using the new data set.

1 Introduction

Assessing the performance of Spoken Dialogue Systems (SDSs) is still an open issue, although research has been conducted in this field for over a decade. The task may be solved using objective and subjective criteria. Here, objective criteria contain measures like dialogue length or success rate which are easily measurable and offer a direct connection to commercial interests. Subjective criteria usually contain the user experience or the user satisfaction. While the latter two are unarguably in the focus of the system users, both are much harder to measure automatically.

Stefan Ultes
Ulm University, Ulm, Germany, e-mail: stefan.ultes@uni-ulm.de

María Jesús Platero Sánchez
University of Granada, Granada, Spain, e-mail: plasez@correo.ugr.es

Alexander Schmitt
Ulm University, Ulm, Germany, e-mail: alexander.schmitt@uni-ulm.de

Wolfgang Minker
Ulm University, Ulm, Germany, e-mail: wolfgang.minker@uni-ulm.de

Interaction Quality (IQ) as defined by Schmitt et al. [6] is another subjective criterion and may be regarded as a more objective version of user satisfaction. The main difference is that instead of asking the actual users, experts rate the dialogues. In previous work, we have shown that Interaction Quality may well be used instead of user satisfaction [18]. A number of automatic estimation approaches have been investigated by us [6, 11, 14, 16] and others [4]. Our focus, however, was on applying IQ for online-adaption of the dialogue [13, 17, 9, 10].

However, the size of the available data in the *LEGO* corpus [7] for the experiments posed a critical limitation especially for experiments casting the problem as a sequential classification task [11]. Hence, in this contribution, we present *LEGOext*, an extension of the *LEGO* corpus[1]. We compare the corpus characteristics of both the original and the new data in order to validate the labeling process. We analyze the performance of previously applied classification approaches on the new extended feature set. Furthermore, we compare the classification performance on the old and new data including cross-corpus analysis.

The outline of this work is as follows: the general idea of the Interaction Quality paradigm is presented in Section 2 including a brief description of the original *LEGO* corpus. The extension of this corpus along with an extended analysis and validation of the annotation process is presented in Section 3. Several different classification methods are applied and evaluated in Section 4 followed by a short discussion of the findings in Section 5.

2 The Interaction Quality Paradigm

The general idea of the Interaction Quality (IQ) paradigm—IQ being defined as user satisfaction annotated by expert raters—is to derive a number of interaction parameters from the dialogue system and use those as input variables to train a statistical classifier targeting IQ. Interaction quality is modeled on a scale from 5 to 1 representing the ratings "satisfied" (5), "slightly unsatisfied" (4), "unsatisfied" (3), "strongly unsatisfied" (2), and "extremely unsatisfied" (1).

The IQ paradigm originally presented by Schmitt et al. [6] is based on automatically deriving interaction parameters from the SDS and feed these parameters into a statistical classification module which predicts the IQ level of the ongoing interaction at the current system-user-exchange. The interaction parameters are rendered on three levels (see Figure 1): the exchange level, the window level, and the dialogue level. The exchange level comprises parameters derived from SDS modules Automatic Speech Recognition (ASR), Spoken Language Understanding (SLU), and Dialogue Management (DM) directly. Parameters on the window and the dialogue level are sums, means, frequencies or counts of exchange level parameters. While dialogue level parameters are computed out of all exchanges of the dialogue up to

[1] *LEGOext* and *LEGO* are publicly available under http://nt.uni-ulm.de/ds-lego.

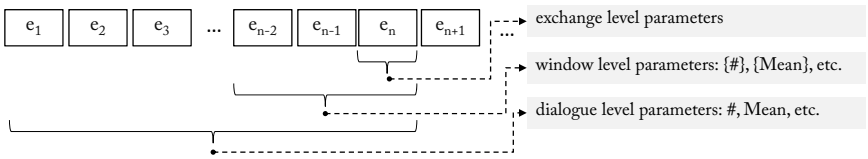


Fig. 1: This figure originally published by Schmitt et al. [6] shows the three parameter levels constituting the interaction parameters: the exchange level containing information about the current exchange, the window level, containing information about the last three exchanges, and the dialogue level containing information about the complete dialogue up to the current exchange.

Table 1: Statistics of the two corpora *LEGO* and *LEGOext* and of the combined corpus *LEGOv2*. Shown are the recording year, the number of calls, the number of exchanges, the average dialogue length in number of exchanges, and the inter-rater agreement.

Corpus	Year	#calls	#exchanges	avg. Length	κ
LEGO	2006	200	4,885	25.4	.54
LEGOext	2007	201	4,753	22.6	.50
LEGOv2		401	9,638	24.0	.52

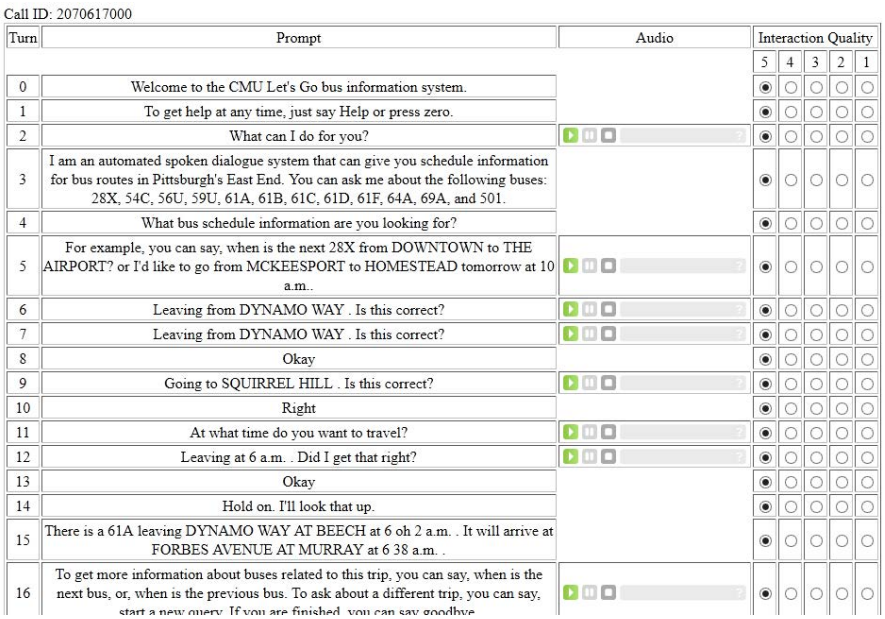
the current exchange, window level parameters are only computed out of the last three exchanges.

These interaction parameters are used as input variables to a statistical classification module. The statistical model is trained based on annotated dialogues of the Lets Go Bus Information System in Pittsburgh, USA [5]. For the original *LEGO* corpus [7], 200 calls from 2006 consisting of 4,885 exchanges have been annotated by three different raters resulting in a rating agreement of $\kappa = 0.54$[2]. Furthermore, the raters had to follow labeling guidelines to enable a consistent labeling process [7].

3 Corpus Statistics

In order to extend the *LEGO* corpus, an additional 201 calls to the Let's Go Bus Information System from 2007 consisting of 4,753 exchanges have been annotated to constitute the *LEGOext* corpus. Three different raters being advanced students of computer science were asked to annotate each system-user-exchange with one out of five satisfaction labels. They used an online form (Figure 2) showing the complete call providing system output and user input as well as audio recordings of each user utterance and of the complete calls. Following the same rating guidelines as in the original *LEGO* corpus [7], the three raters achieved an overall inter-rater agreement of $\kappa = 0.5$[2]. General statistics for both corpora as well as the combined corpus *LEGOv2* are depicted in Table 1.

[2] UAR, κ and ρ are defined in Section 4.1


Call ID: 2070617000

Turn	Prompt	Audio	Interaction Quality				
			5	4	3	2	1
0	Welcome to the CMU Let's Go bus information system.		◉	○	○	○	○
1	To get help at any time, just say Help or press zero.		◉	○	○	○	○
2	What can I do for you?		◉	○	○	○	○
3	I am an automated spoken dialogue system that can give you schedule information for bus routes in Pittsburgh's East End. You can ask me about the following buses: 28X, 54C, 56U, 59U, 61A, 61B, 61C, 61D, 61F, 64A, 69A, and 501.		◉	○	○	○	○
4	What bus schedule information are you looking for?		◉	○	○	○	○
5	For example, you can say, when is the next 28X from DOWNTOWN to THE AIRPORT? or I'd like to go from MCKEESPORT to HOMESTEAD tomorrow at 10 a.m..		◉	○	○	○	○
6	Leaving from DYNAMO WAY . Is this correct?		◉	○	○	○	○
7	Leaving from DYNAMO WAY . Is this correct?		◉	○	○	○	○
8	Okay		◉	○	○	○	○
9	Going to SQUIRREL HILL . Is this correct?		◉	○	○	○	○
10	Right		◉	○	○	○	○
11	At what time do you want to travel?		◉	○	○	○	○
12	Leaving at 6 a.m. . Did I get that right?		◉	○	○	○	○
13	Okay		◉	○	○	○	○
14	Hold on. I'll look that up.		◉	○	○	○	○
15	There is a 61A leaving DYNAMO WAY AT BEECH at 6 oh 2 a.m. . It will arrive at FORBES AVENUE AT MURRAY at 6 38 a.m. .		◉	○	○	○	○
16	To get more information about buses related to this trip, you can say, when is the next bus, or, when is the previous bus. To ask about a different trip, you can say, start a new query. If you are finished, you can say goodbye		◉	○	○	○	○

Fig. 2: The online form used by the expert raters for annotating the *LEGOext* corpus.

Comparing the agreement κ and correlation ρ[3] of the individual IQ ratings between the two corpora depicted in Table 2 shows that the annotation process using the guidelines results in similar agreement.

Table 2: Agreement (κ) and correlation (ρ) in IQ ratings of the 3 raters in *LEGO* and *LEGOext*. Expert ratings show similar correlations among each other.

LEGOext	R1/R2	R1/R3	R2/R3	Mean
κ	.40	.51	.59	**.50**
ρ	.67	.66	.73	**.69**

LEGO	R1/R2	R1/R3	R2/R3	Mean
κ	.64	.48	.51	**.54**
ρ	.79	.68	.70	**.72**

Since the aim is to model a general opinion on Interaction Quality, i.e., mirroring the IQ score other raters (and eventually users) agree with, the final label is determined empirically. Majority voting for deriving the final IQ label is not applicable since many exchanges are labeled with three different ratings, i.e., each of the three raters opted for a different score, thus forming no majority for either score. Therefore, the mean of all rater opinions is considered as possible candidate for the final class label:

[3] UAR, κ and ρ are defined in Section 4.1

$$rating_{mean} = \lfloor \left(\frac{1}{R} \sum_{r=1}^{R} IQ_r \right) + 0.5 \rfloor \, . \tag{1}$$

Here, IQ_r is the Interaction Quality score provided by rater r. $\lfloor y \rfloor$ denotes the highest integer value smaller than y. Every value IQ_r contributes equally to the result that is finally rounded to the closest integer value.

Furthermore, the median is considered, which is defined as

$$rating_{median} = select(sort(IQ_r), \frac{R+1}{2}) \, , \tag{2}$$

where *sort* is a function that orders the ratings IQ_r of all R raters ascendingly and $select(list, i)$ chooses the item with index i from the list $list$. In other words, the IQ score separating the higher half of all ratings to the lower half is selected as final IQ score.

Table 3 shows the agreement between the mean and median labels with the single user ratings. Clearly, the median represents the better choice of final label given the higher values in κ, ρ, and UAR[4]. This validates the findings for the original experiments in the *LEGO* corpus.

Table 3: Agreement of single rater opinions to the merged label when determined by mean and median, measured in UAR, κ, and ρ. On the left side is *LEGOext*, on the right side *LEGO*.

LEGOext		
	Mean Label	**Median Label**
UAR		
Rater1	.550	.648
Rater2	.410	.512
Rater3	.600	.844
Mean	.520	**.668**
Cohen's weighted κ		
Rater1	.612	.806
Rater2	.507	.577
Rater3	.493	.601
Mean	.539	**.661**
Spearman's ρ		
Rater1	.843	.891
Rater2	.905	.846
Rater3	.782	.799
Mean	.843	**.845**

LEGO		
	Mean Label	**Median Label**
UAR		
Rater1	.623	.737
Rater2	.612	.720
Rater3	.545	.605
Mean	.593	**.687**
Cohen's weighted κ		
Rater1	.763	.815
Rater2	.767	.814
Rater3	.657	.658
Mean	.729	**.762**
Spearman's ρ		
Rater1	.901	.900
Rater2	.911	.907
Rater3	.841	.814
Mean	**.884**	.874

The distribution of the final IQ label is shown in Figure 3. For the *LEGOext* corpus, label "5" has been assigned much more frequently while all others have been assigned less often compared to the *LEGO* corpus. This increase in overall

[4] UAR, κ and ρ are defined in Section 4.1

system performance may be a result of an improved system as the 2007 version of Let's Go represents an updated system.

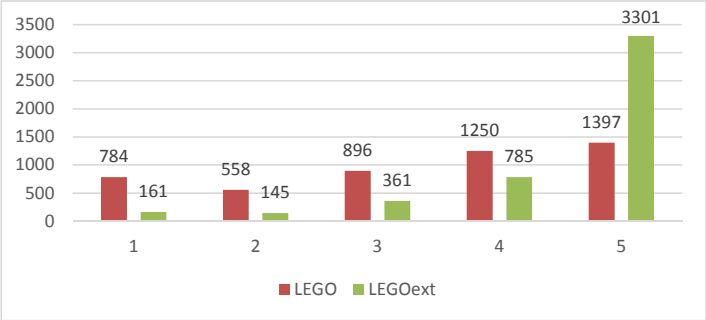


Fig. 3: The distribution of the final label scores along with there absolute number of occurrences for the *LEGO* and the *LEGOext* corpus.

Naturally, this also results in a higher average IQ score for the *LEGOext* corpus: it achieves an average IQ of 4.46 while the *LEGO* corpus achieves 3.39 averaged over all labelled system-user exchanges.

4 IQ Modelling

For evaluating the performance of IQ with the new data set, three classification algorithms have been applied. The main evaluation has been conducted using a Support Vector Machine (SVM) [19] with linear Kernel in accordance to Schmitt et al. [6]. Furthermore, IQ recognition has been cast as a sequence recognition problem with a Conditioned Hidden Markov Model (CHMM) [13] using the JaCHMM library [12]. A difference between a CHMM and an HMM is that a CHMM directly predicts a class probability $p(\omega|\mathbf{x},\lambda)$ for sequence $\mathbf{x}$ while a conventional HMM only provides a probability $p(\mathbf{x}|\lambda)$ that the given model λ represents the observation sequence $\mathbf{x}$. The CHMM was included as initial tests have resulted in bad performance which was attributed to having not enough data [11]. Finally, experiments using Rule Induction (RI) [3] are conducted.

The SVM experiments were conducted using 10-fold cross-validation on the exchange level, i.e., the exchanges were assigned to one of ten subsets without regarding the call they belong to. In each fold, one subset is selected for evaluation while the remaining nine are used for training. By that, each sample is used for evaluation without having it within the training set at the same time. As the CHMM is based on the IQ value evolving over the course of the dialogue, 6-fold cross-validation on the call-level has been applied. Here, each complete call has been assigned to one out of six subsets.

4.1 Evaluation Metrics

Three commonly applied evaluation metrics will be used in this contribution: unweighted average recall(UAR), Spearman's Rho and Cohen's Kappa. The latter two also represent a measure for similarity of paired data. All measures will be briefly described in the following:

Unweighted Average Recall The Unweighted Average Recall (UAR) is defined as the sum of all 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 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 represent the corresponding hypothesis-reference-pair of rating i, and $|R_c|$ the total number of all ratings of class c. In other words, UAR for multi-class classification problems is the accuracy corrected by the effects of unbalanced data.

Cohen's Kappa To measure the relative agreement between two corresponding sets of ratings, the number of label agreements corrected by the chance level of agreement divided by the maximum proportion of times the labelers could agree is computed. κ is defined as

$$\kappa = \frac{p_0 - p_c}{1 - p_c} \ , \tag{5}$$

where p_0 is the rate of agreement and p_c is the chance agreement [1]. As US and IQ are on an ordinal scale, a weighting factor w is introduced reducing the discount of disagreements the smaller the difference is between two ratings [2]:

$$w = \frac{|r_1 - r_2|}{|r_{max} - r_{min}|} \ . \tag{6}$$

Here, r_1 and r_2 denote the rating pair and r_{max} and r_{min} the maximal and minimal rating. This results in $w = 0$ for agreement and $w = 1$ if the ratings have maximal difference.

Spearman's Rho The correlation of two variables describes the degree by that one variable can be expressed by the other. *Spearman's Rank Correlation Coefficient* is a non-parametric method assuming a monotonic function between the two variables [8]. It is defined by

$$\rho = \frac{\sum_i (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_i (x_i - \bar{x})^2 \sum_i (y_i - \bar{y})^2}} \ , \tag{7}$$

Table 4: Results of SVM classification for all feature groups for each corpus separately.

		LEGOext			*LEGO*		
	# feat.	UAR	κ	ρ	UAR	κ	ρ
ASR	29	.378	.287	.494	.458	.535	.689
SLU	5	.221	.093	.239	.260	.219	.311
DM	17	.424	.382	.521	.477	.563	.726
AUTO	51	.463	.482	.604	.512	.614	.764

where x_i and y_i are corresponding ranked ratings and $\bar{x}$ and $\bar{y}$ the mean ranks. Thus, two sets of ratings can have total correlation even if they never agree. This would happen if all ratings are shifted by the same value, for example.

4.2 Support Vector Machine

Three different experiments using a Support Vector Machine have been conducted with the new data. First, the *LEGOext* corpus has been analysed using different feature groups to identify their contribution to the overall performance. The `AUTO` group contains all (automatically derivable) features and subsumes the `ASR`, `SLU`, and `DM` feature groups which contain features belonging to the corresponding dialogue system module (cf. Section 2). The features used correspond to the list of features and their categorization of the *LEGO* corpus [7] and will not be restated here.

The results of SVM experiments on the *LEGOext* corpus are presented in Table 4 and show an UAR of 0.46 for the `AUTO` feature group. Furthermore, the results are compared with the performance of the *LEGO* corpus. It can be seen that, although *LEGOext* achieved lower performance, both corpora result in similar performances. Moreover, the `DM` feature group contributes most to the over all performance having `ASR` second and `SLU` third. This is notable as it shows that besides the `ASR` parameters, the `DM` parameters also have a major impact on the system performance.

A second experiment has been conducted using the combined *LEGOv2* corpus. The results are depicted in Table 5. With an overall performance of UAR 0.51 for the `AUTO` feature group, evaluating on the combined data achieves similar performance compared to each corpus separately. Evaluating the different feature groups furthermore also shows similar results compared to the performance on each corpus separately. However, for the combined data set, the *ASR* feature group contributes most to the overall performance.

Finally, the cross-corpus performance, i.e., training with one corpus and evaluating with the other corpus, has been investigated for all feature groups. Hence, no cross-validation has been applied. The results are depicted in Table 6. While perfor-

Table 5: Results of SVM classification on the combined data set *LEGOv2*.

	# feat.	UAR	κ	ρ
ASR	29	.453	.483	.622
SLU	5	.257	.141	.342
DM	17	.446	.443	.538
AUTO	51	.508	.583	.694

Table 6: Results of SVM classification trained on one corpus and evaluated on the other for all feature groups.

	Train	Eval	UAR	κ	ρ
ASR			.319	.357	.504
SLU	*LEGO*	*LEGOext*	.275	.239	.372
DM			.311	.330	.480
AUTO			.331	.379	.554
ASR			.302	.129	.441
SLU	*LEGOext*	*LEGO*	.245	.019	.134
DM			.441	.257	.474
AUTO			.390	.322	.558

mance decreases, the results are clearly above the majority baseline[5] for all feature groups. The finding that the DM parameters contribute most to the overall system performance is further emphasized as using only those yield the best cross-corpus performance. This means that these feature groups contribute most to the generalization ability of the IQ paradigm.

4.3 Conditioned Hidden Markov Model

As previous studies investigating the applicability of the Conditioned Hidden Markov Model for IQ recognition resulted in low performance presumably due to lack of data, the *LEGOv2* corpus has been used to repeat the original experiments of Ultes et al. [11]. The results are shown in Table 7 along with the results of the original experiment. Unfortunately, the performance has not increased. Two possible reasons have been identified: either the amount of data is still not sufficient or the CHMM is not a suitable model for IQ estimation. The latter might be attributed to the choice of Gaussian mixture models to model the observation probability.

[5] Majority baseline means that the majority class is always predicted. This would result in an UAR of 0.2 for a five class problem.

Table 7: Results of CHMM classification using the *LEGOv2* corpus compared with previous results of the *LEGO* corpus only [11].

	LEGOv2			*LEGO*		
# HS	UAR	κ	ρ	UAR	κ	ρ
5	.39	.399	.542	.38	.4	.56
6	.379	.405	.562	.38	.39	.57
7	.376	.402	.561	.35	.4	.59
8	.336	.27	.385	.37	.41	.59
9	.394	.406	.562	.39	.43	.6
10	.38	.412	.567	.37	.39	.55
11	.389	.417	.566	.36	.41	.58

Table 8: Performance of Rule Induction for cross-corpus evaluation.

Train	Eval	UAR	κ	ρ
LEGOext	*LEGO*	.374	.235	.513
LEGO	*LEGOext*	.293	.264	.436

4.4 Rule Induction

As Rule Induction has shown to perform better than SVMs in previous work [16], RI has also been applied for IQ recognition. However, the claim was that RI produces a lot of specialized rules which result in worse generalizability of the model [15]. To investigate this, the cross-corpus experiment has been repeated using RI as classification method. Again, no cross-validation has been applied due to the experiment characteristics. The results in Table 8 clearly show that RI achieves lower performance on the cross-corpora task for the `AUTO` feature set compared to the SVM. This confirms that using RI results in specialized models not as capable of generalizing than the SVM.

5 Discussion and Conclusion

In this work, we have presented an extension to the *LEGO* corpus adding 201 calls taken from the Let's Go Bus Information System in Pittsburgh, PA, USA. The new calls have been annotated with IQ labels from three different expert raters. The annotation statistics were similar to the statistics of the original corpus thus validating the annotation procedure. This has been underpinned by the performance of SVM classification of IQ on different feature groups achieving an UAR of 0.5 on the combined feature set. Furthermore, cross-corpus classification experiments have been conducted showing the transferability of IQ recognition for different system versions. The DM feature group has been identified as having a major contribution to IQ recognition performance both for evaluation within the corpus as well as for cross-corpus evaluation. Finally, a Conditioned Hidden Markov Model has shown

to not increase performance having more data and Rule Induction has shown to be not as generalizable as Support Vector Machines thus validating claims in previous work.

References

[1] Cohen, J.: A coefficient of agreement for nominal scales. In: Educational and Psychological Measurement, vol. 20, pp. 37–46 (1960)
[2] Cohen, J.: Weighted kappa: Nominal scale agreement provision for scaled disagreement or partial credit. Psychological bulletin **70**(4), 213 (1968)
[3] Cohen, W.W.: Fast effective rule induction. In: Proceedings of the 12th International Conference on Machine Learning, pp. 115–123. Morgan Kaufmann (1995)
[4] El Asri, L., Khouzaimi, H., Laroche, R., Pietquin, O.: Ordinal regression for interaction quality prediction. In: IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 3245–3249. IEEE (2014)
[5] Raux, A., Bohus, D., Langner, B., Black, A.W., Eskenazi, M.: Doing research on a deployed spoken dialogue system: One year of let's go! experience. In: Proc. of the International Conference on Speech and Language Processing (ICSLP) (2006)
[6] 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, Portland, Oregon, USA (2011)
[7] Schmitt, A., Ultes, S., Minker, W.: A parameterized and annotated spoken dialog corpus of the cmu let's go bus information system. In: International Conference on Language Resources and Evaluation (LREC), pp. 3369–337 (2012)
[8] Spearman, C.E.: The proof and measurement of association between two things. American Journal of Psychology **15**, 88–103 (1904)
[9] Ultes, S., Dikme, H., Minker, W.: Dialogue management for user-centered adaptive dialogue. In: Proceedings of the 5th International Workshop On Spoken Dialogue Systems (IWSDS) (2014)
[10] Ultes, S., Dikme, H., Minker, W.: First insight into quality-adaptive dialogue. In: International Conference on Language Resources and Evaluation (LREC), pp. 246–251 (2014)
[11] Ultes, S., ElChabb, R., Minker, W.: Application and evaluation of a conditioned hidden markov model for estimating interaction quality of spoken dialogue systems. In: J. Mariani, L. Devillers, M. Garnier-Rizet, S. Rosset (eds.) Proceedings of the 4th International Workshop on Spoken Language Dialog System (IWSDS), pp. 141–150. Springer (2012)
[12] Ultes, S., ElChabb, R., Schmitt, A., Minker, W.: Jachmm: A java-based conditioned hidden markov model library. In: IEEE International Conference on

Acoustics, Speech and Signal Processing (ICASSP), 2013, pp. 3213–3217. IEEE (2013)

[13] Ultes, S., Heinroth, T., Schmitt, A., Minker, W.: A theoretical framework for a user-centered spoken dialog manager. In: Proceedings of the Paralinguistic Information and its Integration in Spoken Dialogue Systems Workshop, pp. 241 – 246. Springer (2011)

[14] Ultes, S., Minker, W.: Improving interaction quality recognition using error correction. In: Proceedings of the 14th Annual Meeting of the Special Interest Group on Discourse and Dialogue, pp. 122–126. Association for Computational Linguistics (2013). URL http://www.aclweb.org/anthology/W/W13/W13-4018

[15] Ultes, S., Minker, W.: Interaction quality: A review. Bulletin of Siberian State Aerospace University named after academician M.F. Reshetnev (4), 153–156 (2013). URL http://www.vestnik.sibsau.ru/images/vestnik/ves450.pdf

[16] Ultes, S., Minker, W.: Interaction quality estimation in spoken dialogue systems using hybrid-hmms. In: Proceedings of the 15th Annual Meeting of the Special Interest Group on Discourse and Dialogue (SIGDIAL), pp. 208–217. Association for Computational Linguistics (2014). URL http://www.aclweb.org/anthology/W14-4328

[17] Ultes, S., Schmitt, A., Minker, W.: Towards quality-adaptive spoken dialogue management. In: NAACL-HLT Workshop on Future directions and needs in the Spoken Dialog Community: Tools and Data (SDCTD 2012), pp. 49–52. Association for Computational Linguistics, Montréal, Canada (2012). URL http://www.aclweb.org/anthology/W12-1819

[18] Ultes, S., Schmitt, A., Minker, W.: On quality ratings for spoken dialogue systems – experts vs. users. In: Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, pp. 569–578. Association for Computational Linguistics (2013)

[19] Vapnik, V.N.: The nature of statistical learning theory. Springer-Verlag New York, Inc., New York, NY, USA (1995)