3rd International Workshop on Perceptual Quality of Systems (PQS 2010)
6-8 September 2010, Bautzen, Germany

Finding Patterns in User Quality Judgements

Maria K. Wolters[1], *Florian Gödde*[2], *Sebastian Möller*[2], *Klaus-Peter Engelbrecht*[2]

[1]Centre for Speech Technology Research, University of Edinburgh, United Kingdom
[2]Quality and Usability Laboratories, Technical University of Berlin, Germany
maria.wolters@ed.ac.uk, (florian.goedde|sebastian.moeller|klaus-peter.engelbrecht)@telekom.de

Abstract

User quality judgements can show a bewildering amount of variation that is difficult to capture using traditional quality prediction approaches. Using clustering, an exploratory statistical analysis technique, we reanalysed the data set of a Wizard-of-Oz experiment where 25 users were asked to rate the dialogue after each turn. The sparse data problem was addressed by careful a priori parameter choices and comparison of the results of different cluster algorithms. We found two distinct classes of users, positive and critical. Positive users were generally happy with the dialogue system, and did not mind errors. Critical users downgraded their opinion of the system after errors, used a wider range of ratings, and were less likely to rate the system positively overall. These user groups could not be predicted by experience with spoken dialogue systems, attitude to spoken dialogue systems, affinity with technology, demographics, or short-term memory capacity. We suggest that evaluation research should focus on critical users and discuss how these might be identified.

Index Terms: clustering, perceived quality, spoken dialogue systems, evaluation, user modeling

1. Introduction

The holy grail of user modeling for evaluation is an algorithm that predicts how users will rate the quality of a system given a system description and a set of user characteristics. One of the main challenges for such an algorithm is the sheer amount of variation in quality judgements. In this study, we propose a bottom-up approach to this problem. Given a set of evaluations of very similar dialogues, is it possible to find groups of users that follow similar strategies for assigning quality ratings? If these groups can be identified, then how much of the variation in judgements do they explain? Can users be assigned to these groups based on characteristics such as affinity with technology or cognitive abilities?

The rest of this paper is structured as follows. In Section 2, we review the research that motivated the present study. The corpus used for this study is described in Section 3. Next, in Section 4, we explain the main exploratory statistical analysis techniques that we used to uncover and characterise user groups, cluster analysis. We found two groups, one of critical users and one of positive users. Both groups are described in detail in Section 5. None of the user characteristics we measured were able to predict group membership. In Section 6, we discuss the implications for formative evaluation and user simulations. Finally, in Section 7, we stress the need for our findings to be replicated with different data sets, in particular real-world interactions with dialogue systems.

2. Background

Although methods exist which predict average user satisfaction ratings for spoken dialogue systems, a substantial amount of variation in these ratings remains unexplained [1, 2, 3]. We conjecture that at least some of this unexplained variation can be explained by relevant user characteristics, and that users who tend to rate systems in a certain way form distinct groups. This is supported by the results of [4], who found that quality prediction models performed better when users' short-term memory and affinity with technology were taken into account.

As their next step, Engelbrecht *et al.* analysed a corpus of interactions between users and a simulated spoken dialogue system which was designed to make pre-defined errors at certain points in the dialogue. Three types of errors were included, partial understanding (PA), failure to understand an utterance (FA), and incorrect extraction of a concept (IC). The system was a modified version of the BoRIS system [5]. 25 users were asked to perform 5 restaurant searches. After each turn, users indicated their satisfaction with the dialogue so far using a scale from 1 (= poor) to 5 (= excellent). For each dialogue, task success was measured. Users also gave summary quality judgements after each dialogue. After completing all five tasks, they filled in a detailed 37-item questionnaire. On the questionnaire, all statements were rated on a the five-point scale where 1 = "strongly disagree" and 5 = "strongly agree", except for overall quality, which was rated using the same scale as the dialogue turns. Memory was assessed using digit span. Affinity with technology and attitude to spoken dialogue systems were measured using short questionnaires. Users were also asked to indicate whether they had any experience of using spoken dialogue systems.

Although there were clear, significant effects of errors on ratings, these were comparatively small [6]. Users with higher affinity with technology and a more positive attitude to SDS were less likely to penalise errors than users with low scores on both measures. Based on qualitative inspection of their data and interviews with users, Engelbrecht *et al.* found that the users differed substantially in the priorities they assigned to different usability issues and in their leniency, but they did not not estimate the size of each user group or identify relevant measures of rating behaviour. They concluded that user satisfaction predictions might be improved by identifying groups of users with similar rating behaviour.

10.21437/PQS.2010-8

3. Data

In this study, we reanalysed the data described in [6, 7]. For each user, we computed the number of times they used each of the ratings from 1—5. These five variables cover overall rating tendencies. We then subdivided the ratings into five main groups: no error in the previous utterance (OK), partial understanding error in the previous utterance (PA), failure to understand the previous utterance (FA), incorrect extraction of a concept in the previous utterance (IC), and errors in both previous utterances (Two). For each of these groups, we computed the average rating, the range of ratings, and the average change in ratings from the previous turn. All variables represent averages across all five dialogues. This set of fifteen variables quantifies users' reactions to errors.

The first dialogue lasted an average of 10 turns, the second and fourth dialogue required 8 turns, dialogue 3 lasted for 7 turns, and the final dialogue was around 11 turns long. On average, across all five dialogues, each user experienced 36 turns without errors, 7 turns with failure to understand (FA), 3 turns with partial comprehension (PA), and 3 turns where not all concepts were extracted correctly (IC). Dialogues 1–3 contained an average of 3 errors each, while the final two dialogues contained two errors each.

4. Method

Since we are interested in finding groups of users, the statistical method of choice is cluster analysis. Clustering algorithms look for coherent groups of similar items in a set of n data points. The number and composition of groups that are found can vary greatly. Results are affected by many factors, including the exact definitions of similarity and coherence used, the direction of search (top-down, starting with one large cluster, or bottom-up, starting with n small clusters), and, for non-deterministic algorithms, relevant initialisation choices.

When working with data about user ratings of spoken dialogue systems, these problems are exacerbated by the relative lack of data. If clustering algorithms are to detect groups reliably, they need a sufficient number of exemplars per group. Since interactions with an SDS take time, collecting data on user behaviour is expensive, unless quality judgements are collected by survey for a system that has been deployed in the field. Data sets with 50+ users that contain both dialogue transcriptions and quality judgements are rare.

We addressed the sparse data problem by using a five-step strategy which reflects good analytical practice:

1. Choose a good set of features
2. Restrict the number of clusters to a small number
3. Use multiple clustering algorithms and distance measures that make different assumptions about the size and shape of clusters
4. Test whether algorithms yield high-quality, stable clusters
5. For high-quality solutions, check whether the clusters make sense

The first step, choosing the right feature vector, can be automated based on statistical properties of the candidate features or based on experiments with target clustering algorithms. Alternatively, features can be selected that characterise the behaviour patterns of interest.

Table 1: *Agglomerative Clustering Algorithms Tested.*

Algorithm	Merge Criterion
average	average distance between points
median	distance between medians
centroid	distance between centroids
complete	distance of farthest elements
single	distance of closest elements
ward	sum-of-squares error
mcquitty	distance; C = A + B new clus., D old clus. $\Delta(C,D) = (\Delta(A,D) + \Delta(B,D))/2$

The second step, restricting the number of clusters, helps prevent apparently optimal partitions that contain many clusters with very few elements. Such partitions are typically difficult to interpret and generalise. Although algorithms such as MClust [8] can automatically determine an optimal number of clusters based on a model quality measure such as the Bayes Information Criterion, in practice, these approaches benefit from prior information about the expected maximum number of clusters.

The third step is particularly important if there are no strong reasons to assume that the clusters which correspond to the final groups have a particular shape. Otherwise, it is easy to assume that there is no underlying structure in the data because the output of a particular clustering algorithm does not yield any high quality output. A useful package is the R package clusterSim [9], which exhaustively explores many standard clustering algorithms, combined with several different normalisation strategies and distance measures, and ranks the results according to a range of different quality measures.

Agglomerative clustering algorithms start out with each data point in a separate cluster and then merge the two clusters with the shortest distance until the desired amount of clusters has been reached. They mostly differ with respect to the way in which the distance between clusters is defined. Table 1 gives an overview of the agglomerative clustering algorithms used.

In partitioning-based approaches, clusters are defined by their centres. Data points are assigned to the cluster with the closest centre. These approaches start with a random set of centres (build stage), which are then adapted (swap stage) unless the partitioning cannot be optimised further. In pam (partitioning around medoids, [10]), initial cluster centres are data points, and cluster centres are swapped with other data points until the classification can no longer be improved. Partitioning-based algorithms are useful if one wishes to characterise groups through "typical" users.

Conceptually, the fourth step, quantitative evaluation of clusters, should be simple—good clusters are compact units that are well-separated within the feature space. However, in practice, it is difficult to describe what exactly well-separated means. For comparing many different cluster partitions automatically, we used the Calinski criterion, a classic measure of clustering quality. To obtain the value of this criterion, the sum of squared within-cluster distances is divided by the sum of squared between-cluster distances, and the re-

sult is adjusted for the number of clusters by multiplying with $(K-1)/(n-K)$, where K = number of clusters, n=number of data points.

Once the space of potential partitions is reduced, cluster solutions are compared graphically. First, the feature set is reduced to two dimensions for ease of visualisation. In this study, we used the first two principal components for each data set. For each candidate partition of the data set, clusters are plotted along two dimensions obtained by data reduction and the degree of separation and shape of the resulting clusters is observed.

Another important aspect of cluster quality is their stability [11]. If clusterings are determined for different, overlapping subsets of the original data set, the resulting clusters should be highly similar. A useful similarity measure for two sets of clusterings C_1, C_2 is the Jaccard coefficient. Let a be the number of data points that are in the same cluster in both C_1 and C_2 and b the number of data points that are in the same cluster in one set, but in different clusters in the other set. Then, the Jaccard coefficient is defined as $J = a/(a+b)$. It is 1 if both clusterings agree, and below 1 if they disagree. In this study, we estimated stability using the clusterboot algorithm [11, 12] and 100 bootstrap samples per clustering algorithm tested.

Finally, the fifth step is checking whether the clusters can be interpreted in a meaningful way. This step is important because we are using clustering to generate hypotheses about the typology of people who interact with spoken dialogue systems. Although strange clusters that are difficult to interpret may indicate an underlying regularity in the data that is not adequately captured by the feature space, more often than not, such clusters are just artifacts that result from overfitting or bad initialisation.

5. Results

5.1. Identifying Clusters

We ran an exhaustive search of the eight clustering algorithms used in clusterSim for 2, 3, and 4 clusters on several subsets of the original feature set. The 20 best solutions consistently featured only 2 and 3 clusters. Therefore, we discarded the 4 cluster option.

In order to establish basic parameters, we compared five distance metrics (Euclidean, Squared Euclidean, Manhattan, Chebyshev, and GDM1 [9]) and five data normalisation techniques (standardisation using mean and standard deviation, standardisation using median, unitisation, unitisation with minimum 0, and normalisation to the range [-1,1]). The standard Euclidean distance and normalisation to [-1,1] gave consistently good results.

The boxplot in Figure 1 shows the rankings of all eight clustering methods according to the Calinski criterion. Single, centroid, and median linking methods give consistently bad results, with the Ward, average, mcquitty, and complete linking methods scoring highest. This indicates that the underlying groups in this data set are relatively compact. Pam, the only partitioning algorithm in clusterSim, performs slightly worse than the best agglomerative approaches. For the following analysis, we retained both the best agglomerative method (Ward) and the partitioning method pam, because we are interested in characterising our groups by typical users. We used

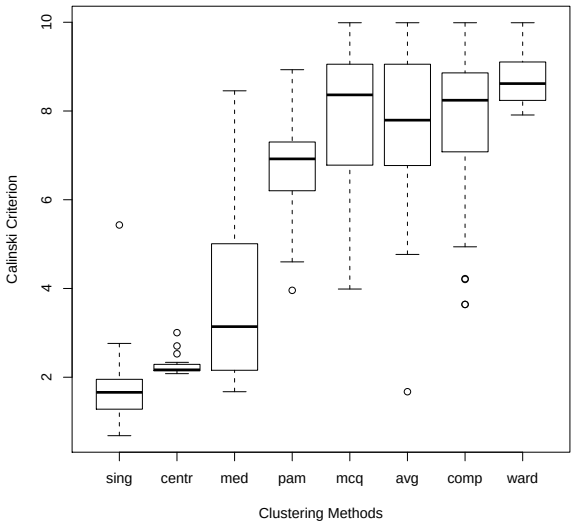

Figure 1: *Calinski Score for Eight Clustering Methods.*

the full original feature set as defined in Section 3.

Next, we examined the stability of the clusters obtained by both methods. Stability was assessed by computing the Jaccard similiarity between the clusters derived from 100 bootstrapped samples and the original clusters. Following [11], a Jaccard coefficient of 0.75 and above is taken to indicate stable clusters, while values between 0.6 and 0.75 indicate a potential pattern in the data set. For the two-cluster Ward solution, the Jaccard coefficients are 0.77 and 0.76, just over the criterion, while for the two-cluster pam solution, the similarities are 0.74 and 0.76, with one stable cluster and one potentially stable one. For the three-cluster solutions, two of the three clusters have Jaccard values below 0.75 for both clustering algorithms (Ward: 0.66, 0.71; Pam: 0.68, 0.64). We conclude that the three-cluster solutions are slightly unstable, and that two-cluster solutions provide a more reliable basis for hypothesis generation.

The results of both pam and the Ward algorithm agree for 21 of the 25 participants. Since Ward yields more stable clusters, we will continue with the clustering derived using this algorithm. The resulting clusters are shown in Figure 2. The clusters are reasonably well-separated apart from a brief area of overlap at the mid point of both dimensions. Experiments with other clustering approaches such as Bayesian maximum-entropy clustering as implemented in MClust [8] or traditional k-means clustering all yielded results that were inferior to the Ward clustering described above.

5.2. Interpreting Clusters

As a first step of exploring what the clusters mean, let us consider the "prototypical" users selected by pam for each of the two groups. For these users, we examined those attributes whose values were below -0.5 or above 0.5 in the normalised feature set, where feature values range from -1 to $+1$. For the first cluster, the prototyp-

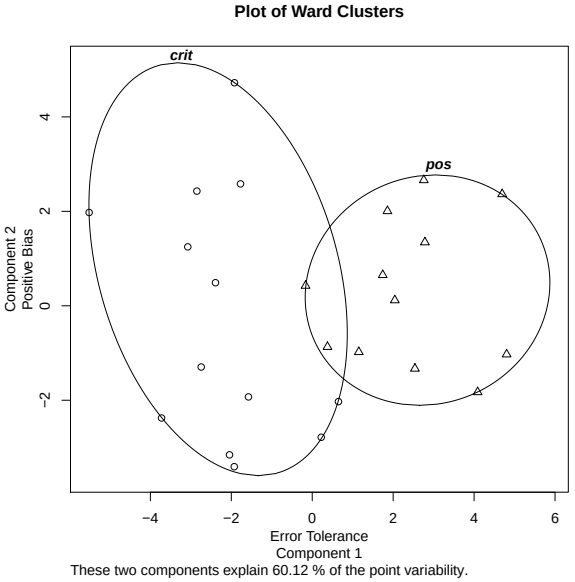

Figure 2: *Clusters Projected Onto First Two Principal Components.*

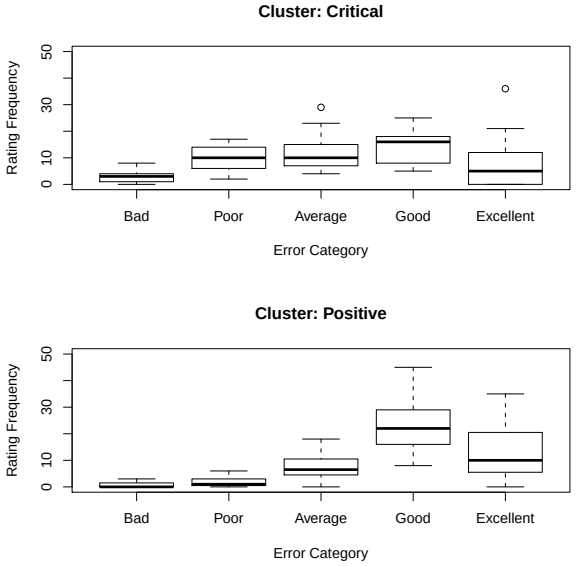

Figure 3: *Overall Frequency of Ratings for Critical and Positive Users.*

ical user is number 8. This person assigns exceptionally many "poor" ratings and tends to downgrade their rating after a completely or partially misunderstood utterance as well as after two system errors in a row. The second cluster is centred on participant number 20. This person is disproportionately enthusiastic, often assigns the highest score, 5 ("excellent"), and reacts positively to misunderstandings and sequences of two errors in a row.

This characterisation is reinforced by the first two principal components of the principal component analysis which was used to plot the clusters in Figure 2. In order to interpret these components, we looked at the features that have a loading of above 0.6 or below -0.6 on each component. For the first principal component, features that load negatively are the range of ratings for non-errors and failures to understand and the average change in ratings for non-errors. Features that load positively are the average change in rating for incompletely extracted concepts, failures to understand, and two errors in a row. In summary, this dimension corresponds to "error tolerance"—people who score highly on this component tend to use a narrower range of values, and are more likely to upgrade their assessment of the system after an error. For the second principal component, features that load negatively are the frequency of "poor" and "average" ratings, while features that load positively are the frequency of "excellent" ratings and mean ratings of all errors except for partial understanding and non-error turns. This dimension can be summarised as "positive bias". People who score highly on this dimension tend to use positive ratings, even if the system makes an error.

Going back to Figure 2, we see that the cluster on the right consists of people who score highly on error tolerance, and tend to have a positive bias. People associated with the cluster on the left vary considerably in overall bias, but is less likely to react positively to errors. These patterns are reflected in distribution of the actual feature values across clusters. Table 2 shows features for which significant differences could be detected using the Kruskal-Wallis test. Users in the left cluster in Figure 2 are more likely to use the ratings 1 ("bad") and 2 ("poor"), while users in the right cluster are more likely to use the rating 4 ("good"). The size of this effect is illustrated in Figure 3. People in the left cluster also assign a lower rating after errors, and they are more likely to choose a lower rating for the error turn than for the non-error turn before. Failure to understand leads to a rating that is on average one point lower than the previous rating, and for utterances where the system failed to extract a concept, ratings drop by almost two points.

From these observations, we can deduce that the first cluster represents *critical* users, while the second cluster consist of more *positive* users. Critical users differentiate between error and non-error turns (p =0.000); positive users do not (p =0.156. This difference is illustrated in Figure 4. Critical users also use different ranges of values for error versus non-error turns (p =0.000), while positive users use the same range of rating for each class of turns (p =0.102).

5.3. Differences in Quality Ratings

The overall quality ratings for each dialogue are somewhat higher in the positive group, but this difference is only significant for the fourth dialogue (cf. Table 3). There were clear differences between both groups on seven of the final evaluation criteria, including overall impression (cf. Table 4). Users in the critical group tended to rate the system as slightly below average, while users in the positive group rated it as above average. Users in the critical group were also more sensitive to system errors. They were more likely to find the system unreliable, and to complain about the frequency of errors. Positive users were more likely to feel comfortable with the sys-

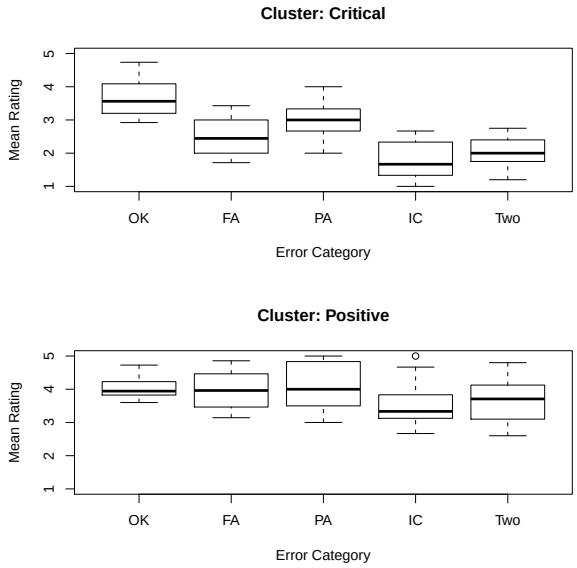

Figure 4: *Mean Ratings for Each Class of Turn (No Error, Failure to Understand, Partial Understanding, Incorrect Concept, Two Errors in a Row) for Critical and Positive Users.*

Table 2: *Significant Differences in Features by Cluster.*

Feature	Critical		Positive		P
	M	SD	M	SD	
Rating "Bad"	2.9	2.6	0.8	1.1	0.019
Rating "Poor"	9.9	4.9	1.9	2.0	0.000
Rating "Good"	13.9	6.2	23.0	10.3	0.022
Mean rating, PA	3.0	0.5	4.1	0.8	0.001
Mean rating, FA	2.5	0.6	4.0	0.6	0.000
Mean rating, IC	1.8	0.6	3.5	0.7	0.000
Rating range, PA	2.4	0.8	3.2	0.7	0.018
Mean change, FA	-0.9	0.5	-0.1	0.4	0.001
Mean change, IC	-1.9	0.8	-1.0	0.6	0.006

tem, to feel that they obtained the desired information, and to feel that the system understood them.

5.4. Can Individual Differences Predict Clusters?

Despite clear and consistent differences in quality ratings both during and after the dialogues, none of the user characteristics measured during the original study is associated with cluster membership. The distribution of genders and experience with SDS is almost completely equal between clusters. 6 of all 12 male users and 6 of all 13 female users were in the positive group, the rest were in the critical group. Likewise, of the 4 users with no previous SDS experience, 2 were in the positive group, and of the 21 experienced users, 10 were in the positive group—again, almost exactly half. Users also do not differ with respect to task success ($p = 0.847$), affinity with technology ($p = 0.584$), attitude to spoken dialogue systems ($p = 0.723$), or digit span performance (short-term memory capacity, $p = 0.869$).

Table 3: *Differences in Overall Dialogue Quality Judgements by Cluster, Kruskal-Wallis Test for Significance.*

Dialogue	Critical		Positive		P
	M	SD	M	SD	
1	3.1	0.7	3.5	0.9	0.258
2	2.2	1.2	2.6	0.8	0.208
3	2.4	0.7	2.9	1.0	0.117
4	2.6	1.1	3.6	0.6	0.017
5	2.7	1.0	3.3	1.0	0.111

Table 4: *Significant Differences in Questionnaire Ratings by Cluster, Kruskal-Wallis Test for Significance.*

Criterion	Critical		Positive		P
	M	SD	M	SD	
Overall impression	2.8	0.3	3.5	0.5	0.009
System unreliable	3.2	0.8	2.2	0.6	0.006
Obtained relevant info.	3.3	0.9	3.9	0.3	0.021
System understood me	2.4	0.8	3.2	0.7	0.018
Many system errors	3.5	0.8	2.8	0.6	0.019
Using system was fun	2.5	1.0	3.3	0.8	0.043
Comfortable w/ system	2.9	0.9	4.0	0.6	0.002

6. Discussion

Even though the number of data points is relatively small, we successfully identified a meaningful underlying structure in users' quality judgements. Half the users belonged to a cluster that can be characterised as "positive". These users tend to give mainly positive ratings and appreciate attempts at error recovery. The other half is more "critical". These users assign a greater variety of ratings and penalise system errors. Both clusters are substantiated by a range of converging evidence, prototypical users found through clustering, principal component analysis, and analysis of the distribution of feature values across clusters. The clusters also agree with the qualitative observations by Engelbrecht *et al.* [6] summarised in Section 2. This high level of consistency and convergence suggests that what we have found indeed reflects an aspect of the real structure of the data set.

The result has implications for both formative and summative evaluation of spoken dialogue systems. For formative evaluation, it may be best to work with critical users only, since they will have a lower error tolerance and may therefore be more likely to highlight usability problems. For the summative evaluation of a finished system, it may be best to analyse data from both user groups separately, since as Figure 4 illustrates, the same analysis may yield different results for both groups.

Whenever simulated users are used to learn dialogue management policies using a reward function that is based on user satisfaction or user quality ratings [13, 14], a case can be made for deriving the reward function and training policies mainly from interactions between the

critical users and the system, not from positive users. Since positive users do not tend to change their overall ratings based on dialogue behaviour, a reward function based on their judgements may severely underestimate the effect of system errors or other usability problems. Moreover, models that are trained with simulated positive users may not provide learning algorithms with sufficient information for exploring the space of potential policies, because they will rate almost all interactions positively. A model trained on the positive user group detected in this study might even encourage dialogue policies that make frequent errors, because positive users tended to revise their opinion of the system upwards after an error, not downwards. If positive users differ from critical users in their dialogue behaviour, policies may need to be learned with simulated users from both groups, but the parameters of the learning algorithm may need to be adjusted depending on the user group.

Another important finding is that none of the user characteristics that were assessed in the corpus allow us to predict what user group a person will belong to. At first glance, this appears to contradict Engelbrecht *et al.*'s finding that affinity with technology and attitude to SDS correlate with higher ratings after errors. However, those analyses concerned individual features, whereas the clusters represent more generic patterns of behaviour.

Although the two user groups are relatively well-defined, the clusters have not yet been incorporated into a formal approach to quality prediction; it is not clear to what extent they would improve different models. An additional challenge here is the sparseness of the data set. Each cluster consists of half the original data set, which makes it more difficult to derive good estimations for the parameters of a PARADISE-style model.

Finally, the original set up of Engelbrecht *et al.*'s experiment is relatively artificial. Not only are users asked to interact with simulated spoken dialogue systems that have been primed to make errors, they are also asked to rate the current interaction after each turn. It is not clear how this setup affects users' overall quality judgements and, by extension, the user groups we found in the data.

7. Future Work

Our analysis needs to be replicated on different data sets to ascertain whether similar user groups emerge. If our categories hold across data sets, then strategies need to be developed for identifying critical users. There are two potential solutions. If there are differences between user groups in terms of dialogue act patterns, linguistic structure, content of utterances, these could be used to automatically predict the overall rating behaviour of a user based on a few brief, well-designed interactions with a SDS. One might also want to investigate the predictive power of other user characteristics that can be measured easily before data collection, such as personality.

Most importantly, we need to examine data sets of interactions between real users and deployed spoken dialogue systems where task success matters to the user. It is probably fair to assume that many of the users who were in the "positive" group in this study would be less lenient if system errors delayed or thwarted an actual restaurant booking.

8. Acknowledgements

We would like to thank Felix Hartard for collecting the original data set. This research was funded by the Quality and Usability Labs.

9. References

[1] M. Walker, D. J. Litman, C. A. Kamm, and A. Abella, "PARADISE: Framework for evaluating spoken dialogue agents," in *Proceedings of the 35th Annual Meeting of the Association of Computational Linguistics*, 1997.

[2] E. Frøkjær, M. Hertzum, and K. Hornbæk, "Measuring usability: are effectiveness, efficiency, and satisfaction really correlated?" in *Proceedings of CHI 2000, The Hague, Amsterdam*, 2000, pp. 345–352.

[3] S. Möller, K.-P. Engelbrecht, and R. Schleicher, "Predicting the quality and usability of spoken dialogue services," *Speech Communication*, vol. 50, no. 8-9, pp. 730 – 744, 2008.

[4] K.-P. Engelbrecht, S. Möller, R. Schleicher, and I. Wechsung, "Analysis of paradise models for individual users of a spoken dialog system," *Proc. ESSV*, 2008.

[5] S. Möller, *Quality of Telephone-Based Spoken Dialogue Systems*. New York, NY: Springer, 2005.

[6] K. P. Engelbrecht, F. Hartard, F. Gödde, and S. Möller, "A closer look at quality judgments of spoken dialog systems," in *Proc. Interspeech*, 2009.

[7] K.-P. Engelbrecht, F. Gödde, F. Hartard, H. Ketabdar, and S. Möller, "Modeling user satisfaction with hidden markov model," in *SIGDIAL '09: Proceedings of the SIGDIAL 2009 Conference*. Morristown, NJ, USA: Association for Computational Linguistics, 2009, pp. 170–177.

[8] C. Fraley and A. E. Raftery, "Model-based clustering, discriminant analysis and density estimation," *Journal of the American Statistical Association*, vol. 97, pp. 611–631, 2002.

[9] M. Walesiak, "Cluster analysis with ClusterSim computer program and R environment," *Acta Universitatis Lodziniensis, Folia Oeconomica*, vol. 216, pp. 303–311, 2008.

[10] L. Kaufman and P. J. Rousseeuw, *Finding groups in data*. New York NY: Wiley, 1990.

[11] C. Hennig, "Dissolution point and isolation robustness: robustness criteria for general cluster analysis methods," *Journal of Multivariate Analysis*, vol. 99, no. 6, pp. 1154–1176, 2008.

[12] ——, *fpc: Fixed point clusters, clusterwise regression and discriminant plots*, 2009, r package version 1.2-7. [Online]. Available: http://CRAN.R-project.org/package=fpc

[13] J. Schatzmann, K. Weilhammer, M. Stuttle, and S. Young, "A survey of statistical user simulation techniques for reinforcement-learning of dialogue management strategies," *Knowledge Engineering Review*, vol. 21, pp. 97–126, 2006.

[14] V. Rieser and O. Lemon, "Automatic learning and evaluation of user-centered objective functions for dialogue system optimisation," in *Proc. of LREC*, vol. 8, 2008.