Cluster-based Prediction of User Ratings for Stylistic Surface Realisation

Nina Dethlefs, Heriberto Cuayáhuitl, Helen Hastie, Verena Rieser and Oliver Lemon
Heriot-Watt University, Mathematical and Computer Sciences, Edinburgh
n.s.dethlefs@hw.ac.uk

Abstract

Surface realisations typically depend on their target style and audience. A challenge in estimating a stylistic realiser from data is that humans vary significantly in their subjective perceptions of linguistic forms and styles, leading to almost no correlation between ratings of the same utterance. We address this problem in two steps. First, we estimate a mapping function between the linguistic features of a corpus of utterances and their human style ratings. Users are partitioned into clusters based on the similarity of their ratings, so that ratings for new utterances can be estimated, even for new, *unknown* users. In a second step, the estimated model is used to re-rank the outputs of a number of surface realisers to produce stylistically adaptive output. Results confirm that the generated styles are recognisable to human judges and that predictive models based on clusters of users lead to better rating predictions than models based on an average population of users.

1 Introduction

Stylistic surface realisation aims not only to find the best realisation candidate for a semantic input based on some underlying trained model, but also aims to adapt its output to properties of the user, such as their age, social group, or location, among others. One of the first systems to address stylistic variation in generation was Hovy (1988)'s PAULINE, which generated texts that reflect different speaker attitudes towards events based on multiple, adjustable features. Stylistic variation in such contexts can often be modelled systematically as a *multidimensional variation space* with several continuous dimensions, so that varying stylistic scores indicate the strength of each dimension in a realisation candidate. Here, we focus on the dimensions of *colloquialism, politeness* and *naturalness*. Assuming a target score on one or more dimensions, candidate outputs of a data-driven realiser can then be ranked according to their predicted affinity with the target scores.

In this paper, we aim for an approach to stylistic surface realisation which is on the one hand based on natural human data so as to reflect stylistic variation that is as natural as possible. On the other hand, we aim to minimise the amount of annotation and human engineering that informs the design of the system. To this end, we estimate a mapping function between automatically identifiable shallow linguistic features characteristic of an utterance and its human-assigned style ratings. In addition, we aim to address the high degree of variability that is often encountered in subjective rating studies, such as assessments of recommender systems (O'Mahony et al., 2006; Amatriain et al., 2009), sentiment analysis (Pang and Lee, 2005), or surface realisations, where user ratings have been shown to differ significantly ($p<0.001$) for the same utterance (Walker et al., 2007). Such high variability can affect the performance of systems which are trained from an average population of user ratings. However, we are not aware of any work that has addressed this problem principally by estimating ratings for both *known* users, for whom ratings exists, and *unknown* users, for whom no prior ratings exist. To achieve this, we propose to partition users into clusters of individuals who assign similar ratings to linguistically similar utterances, so that their ratings can be estimated more accurately than

based on an average population of users. This is similar to Janarthanam and Lemon (2014), who show that clustering users and adapting to their level of domain expertise can significantly improve task success and user ratings. Our resulting model is evaluated with realisers not originally built to deal with stylistic variation, and produces natural variation recognisable by humans.

2 Architecture and Domain

We aim to with generating restaurant recommendations as part of an interactive system. To do this, we assume that a generator input is provided by a preceding module, e.g. the interaction manager, and that the task of the surface realiser is to find a suitable stylistically appropriate realisation. An example input is *inform(food=Italian, name=Roma)*, which could be expressed as *The restaurant Roma serves Italian food*. A further aspect is that users are initially *unknown* to the system, but that it should adapt to them over time by discovering their stylistic preferences. Future work involves integrating the surface realiser into the PARLANCE[1] (Hastie et al., 2013) spoken dialogue system with a method for triggering the different styles. Here, we leave the question of when different styles are appropriate as future work and focus on being able to generate them.

The architecture of our model is shown in Figure 1. Training of the regression model from stylistically-rated human corpora is shown in the top-left box (grey). Utterance ratings from human judges are used to extract shallow linguistic features as well as to estimate user clusters. Both types of information inform the resulting stylistic regression model. For surface realisation (top-right box, blue), a semantic input from a preceding model is given as input to a surface realiser. Any realiser is suitable that returns a ranked list of output candidates. The resulting list is re-ranked according to stylistic scores estimated by the regressor, so that the utterance which most closely reflects the target score is ranked highest. The re-ranking process is shown in the lower box (red).

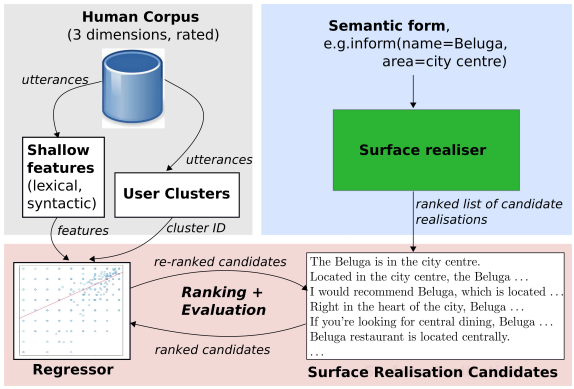

Figure 1: Architecture of stylistic realisation model. Top left: user clusters are estimated from corpus utterances described by linguistic features and ratings. Top right: surface realisation ranks a list of output candidates based on a semantic input. These are ranked stylistically given a trained regressor.

3 Related Work

3.1 Stylistic Variation in Surface Realisation

Our approach is most closely related to work by Paiva and Evans (2005) and Mairesse and Walker (2011), discussed in turn here. Paiva and Evans (2005) present an approach that uses multivariate linear regression to map individual linguistic features to distinguishable styles of text. The approach works in three steps. First, a factor analysis is used to determine the relevant stylistic dimensions from a corpus of human text using shallow linguistic features. Second, a hand-crafted generator is used to produce a large set of utterances, keeping traces of each generator decision, and obtaining style scores for each output based on the estimated factor model. The result is a dataset of *<generator decision, style score>* pairs which can be used in a correlation analysis to identify the predictors of particular output styles. During generation, the correlation equations inform the generator at each choice point so as to best express the desired style. Unfortunately, no human evaluation of the model is presented so that it remains unclear to what extent the generated styles are perceivable by humans.

Closely related is work by Mairesse and Walker (2011) who present the PERSONAGE system, which aims to generate language reflecting particular personalities. Instead of choosing generator decisions by considering their predicted style scores, however, Mairesse and Walker (2011) directly predict generator decisions based on target personality scores. To obtain the generator, the authors first generate a corpus of utterances which differ randomly in their linguistic choices. All utterances are rated by humans indicating the

[1] http://parlance-project.eu

extent to which they reflect different personality traits. The best predictive model is then chosen in a comparison of several classifiers and regressors. Mairesse and Walker (2011) are the first to evaluate their generator with humans and show that the generated personalities are indeed recognisable.

Approaches on replicating personalities in realisations include Gill and Oberlander (2002) and Isard et al. (2006). Porayska-Pomsta and Mellish (2004) and Gupta et al. (2007) are approaches to politeness in generation, based on the notion of face and politeness theory, respectively.

3.2 User Preferences in Surface Realisation

Taking users' individual content preferences into account for training generation systems can positively affect their performance (Jordan and Walker, 2005; Dale and Viethen, 2009). We are interested in individual user perceptions concerning the *surface realisation* of system output and the way they relate to different stylistic dimensions. Walker et al. (2007) were the first to show that individual preferences exist for the perceived *quality* of realisations and that these can be modelled in trainable generation. They train two versions of a rank-and-boost generator, a first version of which is trained on the average population of user ratings, whereas a second one is trained on the ratings of individual users. The authors show statistically that ratings from different users are drawn from different distributions ($p<0.001$) and that significantly better performance is achieved when training and testing on data of individual users. In fact, training a model on one user's ratings and testing it on another's performs as badly as a random baseline. However, no previous work has modelled the individual preferences of *unseen* users–for whom no training data exists.

4 Estimation of Style Prediction Models

4.1 Corpora and Style Dimensions

Our domain of interest is the automatic generation of restaurant recommendations that differ with respect to their *colloquialism* and *politeness* and are as *natural* as possible. All three stylistic dimension were identified from a qualitative analysis of human domain data. To estimate the strength of each of them in a single utterance, we collect user ratings for three data sets that were collected under different conditions and are freely available.

Corpus	Colloquial	Natural	Polite
LIST	3.38 ± 1.5	4.06 ± 1.2	4.35 ± 0.8
MAI	3.95 ± 1.2	4.32 ± 1.0	4.27 ± 0.8
CLASSIC	4.29 ± 1.1	4.20 ± 1.2	3.64 ± 1.3

Table 1: Average ratings with standard deviations. Ratings between datasets (except one) differ significantly at $p<0.01$, using the Wilcoxon signed-rank test.

- LIST is a corpus of restaurant recommendations from the website The List.[2] It consists of professionally written reviews. An example is *"Located in the heart of Barnwell, Beluga is an excellent restaurant with a smart menu of modern Italian cuisine."*
- MAI is a dataset collected by Mairesse et al. (2010),[3] using Amazon Mechanical Turk. Turkers typed in recommendations for various specified semantics; e.g. *"I recommend the restaurant Beluga near the cathedral."*
- CLASSIC is a dataset of transcribed spoken user utterances from the CLASSiC project.[4] The utterances consist of user queries for restaurants, such as *"I need an Italian restaurant with a moderate price range."*

Our joint dataset consists of $1,361$ human utterances, 450 from the LIST, 334 from MAI, and 577 from CLASSIC. We asked users on the CrowdFlower crowdsourcing platform[5] to read utterances and rate their colloquialism, politeness and naturalness on a 1-5 scale (the higher the better). The following questions were asked.

- **Colloquialism:** The utterance is colloquial, i.e. could have been spoken.
- **Politeness:** The utterance is polite / friendly.
- **Naturalness:** The utterance is natural, i.e. could have been produced by a human.

The question on naturalness can be seen as a general quality check for our training set. We do not aim to generate unnatural utterances. 167 users took part in our rating study leading to a rated dataset of altogether $3,849$ utterances. All users were from the USA. The average ratings per dataset and stylistic dimension are summarised in Table 1. From this, we can see that LIST utterances were perceived as the least natural and

[2] http://www.list.co.uk/

[3] http://people.csail.mit.edu/francois/research/bagel/

[4] http://www.classic-project.org/

[5] http://crowdflower.com/

colloquial, but as the most polite. CLASSIC utterances were perceived as the most colloquial, but the least polite, and MAI utterances were rated as the most natural. Differences between ratings for each dimension and dataset are significant at $p<0.01$, using the Wilcoxon signed-rank test, except the naturalness for MAI and CLASSIC.

Since we are mainly interested in the lexical and syntactic features of utterances here, the fact that CLASSIC utterances are spoken, whereas the other two corpora are written, should not affect the quality of the resulting model. Similarly, some stylistic categories may seem closely related, such as *colloquialism* and *naturalness*, or orthogonal to each other, such as *politeness* and *colloquialism*. However, while ratings for *colloquialism* and *naturalness* are very close for the CLASSIC dataset, they vary significantly for the two other datasets ($p<0.01$). Also, the ratings for *colloquialim* and *politeness* show a weak positive correlation of 0.23, i.e. are not perceived as orthogonal by users. These results suggest that all in all our three stylistic categories are perceived as sufficiently different from each other and suitable for training to predict a spectrum of different styles.

Another interesting aspect is that individual user ratings vary significantly, leading to a high degree of variability for identical utterances. This will be the focus of the following sections.

4.2 Feature Estimation

Table 2 shows the feature set we will use in our regression experiments. We started from a larger subset including 45 lexical and syntactic features as well as unigrams and bigrams, all of which could be identified from the corpus without manual annotation. The only analysis tool we used was the Stanford Parser,[6] which identified certain types of words (pronouns, wh-words) or the depth of syntactic embedding. A step-wise regression analysis was then carried out to identify those features that contributed significantly (at $p<0.01$) to the overall regression equation obtained per stylistic dimension. Of all lexical features (unigrams and bigrams), the word *with* was the only contributor. A related feature was the average *tf-idf* score of the content words in an utterance.

[6] http://nlp.stanford.edu/software/lex-parser.shtml

Feature	Type
Length of utterance	num
Presence of personal pronouns	bool
Presence of WH words	bool
with cue word	bool
Presence of negation	bool
Average length of content words	num
Ave *tf-idf* score of content words	num
Depth of syntactic embedding	num

Table 2: Features used for regression, which were identified as significant contributors ($p<0.01$) from a larger feature set in a step-wise regression analysis.

4.3 Regression Experiments

Based on the features identified in Section 4.2, we train a separate regressor for each stylistic dimension. The task of the regressor is to predict, based on the extracted linguistic features of an utterance, a score in the range of 1-5 for colloquialism, politeness and naturalness. We compare: (1) a multivariate multiple regressor (MMR), (2) an M5P decision tree regressor, (3) a support vector machine (SVM) with linear kernel, and (4) a ZeroR classifier, which serves as a majority baseline. We used the *R* statistics toolkit[7] for the MMR and the Weka toolkit[8] for the remaining models.

Average User Ratings The regressors were first trained to predict the average user ratings of an utterance and evaluated in a 10-fold cross validation experiment. Table 3 shows the results. Here, r denotes the Pearson correlation coefficient, which indicates the correlation between the predicted and the actual user scores; R^2 is the coefficient of determination, which provides a measure of how well the learnt model fits the data; and $RMSE$ refers to the Root Mean Squared Error, the error between the predicted and actual user ratings.

We can observe that MMR achieves the best performance for predicting colloquialism and naturalness, whereas M5P best predicts politeness. Unfortunately, all regressors achieve at best a moderate correlation with human ratings. Based on these results, we ran a correlation analysis for all utterances for which more than 20 original user ratings were available. The purpose was to find out to what extent human raters agree with each other. The results showed that user agreement in fact ranges from a high positive corre-

[7] http://www.r-project.org/

[8] http://www.cs.waikato.ac.nz/ml/weka/

	Model	r	R^2	$RMSE$
Colloquial	MMR	**0.50**	**0.25**	**0.85**
	SVM	0.47	0.22	0.86
	M5P	0.48	0.23	0.85
	ZeroR	-0.08	0.006	0.97
Natural	MMR	**0.30**	**0.09**	**0.78**
	SVM	0.24	0.06	0.81
	M5P	0.27	0.07	0.78
	ZeroR	-0.09	0.008	0.81
Polite	MMR	0.33	0.11	0.71
	SVM	0.31	0.09	0.73
	M5P	**0.42**	**0.18**	**0.69**
	ZeroR	-0.09	0.008	0.76

Table 3: Comparison of regression models per dimension *using average user ratings*. The best model is indicated in bold-face for the correlation coefficient.

	Model	r	R^2	$RMSE$
Colloquial	MMR	**0.61**	**0.37**	**1.05**
	SVM	0.36	0.13	1.3
	M5P	0.56	0.31	1.07
	ZeroR	-0.06	0.004	1.3
Natural	MMR	**0.55**	**0.30**	**0.96**
	SVM	0.36	0.13	1.13
	M5P	0.49	0.24	0.99
	ZeroR	-0.08	0.06	1.13
Polite	MMR	0.69	0.48	0.76
	SVM	0.54	0.30	0.92
	M5P	**0.71**	**0.50**	**0.73**
	ZeroR	-0.04	0.002	1.04

Table 4: Comparison of regression models per dimension *using individual user ratings.* The best model is indicated in bold-face for the correlation coefficient.

lation of 0.79 to a moderate negative correlation of -0.55. The average is 0.04 (SD=0.95), i.e. indicating *no correlation between user ratings*, even for the same utterance. This observation is partially in line with related work that has found high diversity in subjective user ratings. Yeh and Mellish (1997) report only 70% agreement of human judges on the best choice of referring expression. Amatriain et al. (2009) report inconsistencies in user ratings in recommender systems with an RMSE range of 0.55 to 0.81 and argue that this constitutes a lower bound for system performance. This inconsistency is exacerbated by raters recruited via crowdsourcing platforms as in our study (Koller et al., 2010; Rieser et al., 2011). However, while crowdsourced data have been shown to contain substantially more noise than data collected in a lab environment, they do tend to reflect the general tendency of their more controlled counterparts (Gosling et al., 2004).

Individual User Ratings Given that individual preferences exist for surface realisation (Walker et al., 2007), we included the *user's ID* as a regression feature and re-ran the experiments. The hypothesis was that if users differ in their preferences for realisation candidates, they may also differ in terms of their perceptions of linguistic styles. The results shown in Table 4 support this: the obtained correlations are significantly higher ($p<0.001$, using the Fisher r-to-z transformation) than those without the user's ID (though we are still not able to model the full variation observed in ratings). Importantly, this shows that user ratings are *intrinsically coherent* (not random) and that variation exists mainly for inter-user agreement. This model performs satisfactorily for a known population of users. However, it does not allow the prediction of ratings of *unknown* users, who we mostly encounter in generation.

5 Clustering User Rating Behaviour

5.1 Spectral Clustering

The goal of this section is to find a number of k clusters which partition our data set of user ratings in a way that users in one cluster rate utterances with particular linguistic properties most similarly to each other, while rating them most dissimilarly to users in other clusters. We assume a set of n data points $x_1 \dots x_n$, which in our case correspond to an individual user or group of users, characterised in terms of word bigrams, POS tag bigrams, and assigned ratings of the utterance they rated. An example is *Beluga*_NNP *serves*_VBZ *Italian*_JJ *food*_NN*; [col=5.0, nat=5.0, pol=4.0].* Features were chosen as a subset of relevant features from the larger set used for regression above.

Using spectral clustering (von Luxburg, 2007), clusters can be identified from a set of eigenvectors of an affinity matrix S derived from pair-wise similarities between data points $s_{ij} = s(x_i, x_j)$ using a symmetric and non-negative similarity function. To do that, we use a cumulative similarity based on the Kullback-Leibler divergence,

$$D(P,Q) = \frac{\sum_i p_i log_2(\frac{p_i}{q_i}) + \sum_j q_j log_2(\frac{q_j}{p_j})}{2},$$

where P is a distribution of words, POS tags or ratings in data point x_i; and Q a similar distribution in data point x_j. The lower the cumulative di-

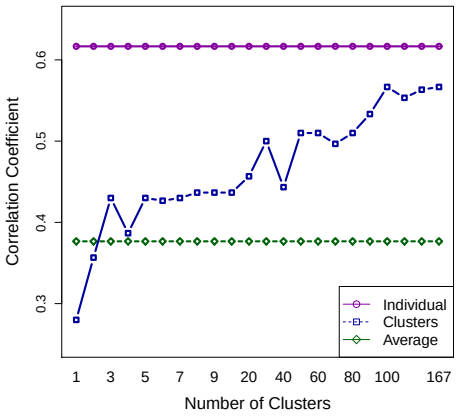

Figure 2: Average correlation coefficient for different numbers of clusters. For comparison, results from *average* and *individual* user ratings are also shown.

vergence between two data sets, the more similar they are. To find clusters of similar users from the affinity matrix S, we use the algorithm described in Ng et al. (2001). It derives clusters by choosing the k largest eigenvectors $u_1, u_2, \ldots, u_k$ from the Laplacian matrix $L = D^{1/2} - SD^{1/2}$ (where D is a diagonal matrix), arranging them into columns in a matrix $U = [u_1 u_2 \ldots u_k]$ and then normalising them for length. The result is a new matrix T, obtained through $t_{ij} = u_{ij}/(\sum_k u_{ik}^2)^{1/2}$. The set of clusters $C_1, \ldots C_k$ can then be obtained from T using the K-means algorithm, where each row in T serves as an individual data point. Finally, each original data point x_i (row i of T) is assigned to a cluster C_j. In comparison to other clustering algorithms, experiments by Ng et al. (2001) show that spectral clustering is robust for convex and non-convex data sets. The authors also demonstrate why using K-means only is often not sufficient.

The main clusters obtained describe surface realisation preferences by particular groups of users. An example is the realisation of the location of a restaurant as a prepositional phrase or as a relative clause as in *restaurant in the city centre* vs. *restaurant located in the city centre*; or the realisation of the food type as an adjective, *an Italian restaurant*, vs. a clause, *this restaurant serves Italian food*. Clusters can then be characterised as different combinations of such preferences.

5.2 Results: Predicting Stylistic Ratings

Figure 2 shows the average correlation coefficient r across dimensions in relation to the number of clusters, in comparison to the results obtained with *average* and *individual* user ratings. We can see that the baseline without user information is outperformed with as few as three clusters. From 30 clusters on, a medium correlation is obtained until another performance jump occurs around 90 clusters. Evidently, the best performance would be achieved by obtaining one cluster per user, i.e. 167 clusters, but nothing would be gained in this way, and we can see that useful generalisations can be made from much fewer clusters. Based on the clusters found, we will now predict the ratings of known and unknown users.

Known Users For known users, first of all, Figure 3 shows the correlations between the predicted and actual ratings for colloquialism, politeness and naturalness based on 90 user clusters. Correlation coefficients were obtained using an MMR regressor. We can see that a medium correlation is achieved for naturalness and (nearly) strong correlations are achieved for politeness and colloquialism. This confirms that clustering users can help to better predict their ratings than based on shallow linguistic features alone, but that more generalisation is achieved than based on individual user ratings that include the user's ID as a regression feature. The performance gain in comparison to predicting average ratings is significant ($p<0.01$) from as few as three clusters onwards.

Unknown Users We initially sort unknown users into the majority cluster and then aim to make more accurate cluster allocations as more information becomes available. For example, after a user has assigned their first rating, we can take it into account to re-estimate their cluster more accurately. Clusters are re-estimated with each new rating, based on our trained regression model. While estimating a user cluster based on linguistic features alone yields an average correlation of 0.38, an estimation based on linguistic features and a single rating alone already yields an average correlation of 0.45. From around 30 ratings, the average correlation coefficients achieved are as good as for known users. More importantly, though, estimations based on a single rating alone significantly outperform ratings based on the av-

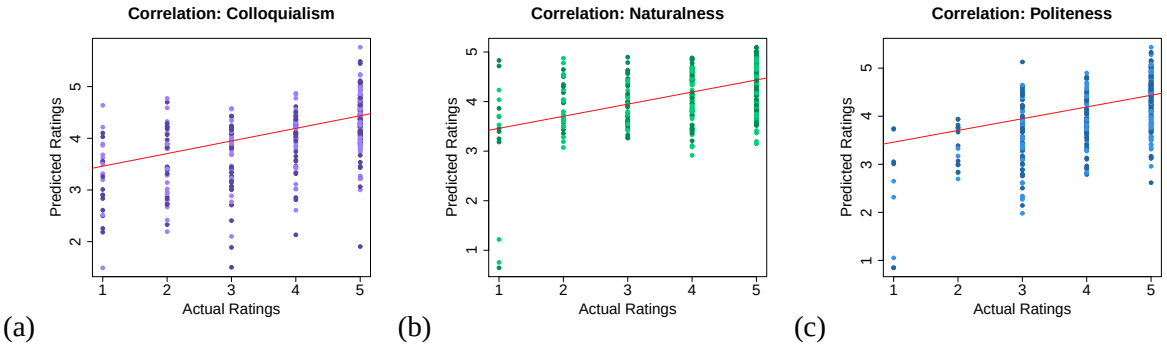

Figure 3: Correlations per dimension between actual and predicted user ratings based on 90 user clusters: (a) Colloquialism (r = 0.57, p<0.001), (b) Naturalness (r = 0.49, p<0.001) and (c) Politeness (r = 0.59, p<0.001).

erage population of users (p<0.001). Fig. 4 shows this process. It shows the correlation between predicted and actual user ratings for unknown users over time. This is useful in interactive scenarios, where system behaviour is refined as more information becomes available (Cuayáhuitl and Dethlefs, 2011; Gašić et al., 2011), or for incremental systems (Skantze and Hjalmarsson, 2010; Dethlefs et al., 2012b; Dethlefs et al., 2012a).

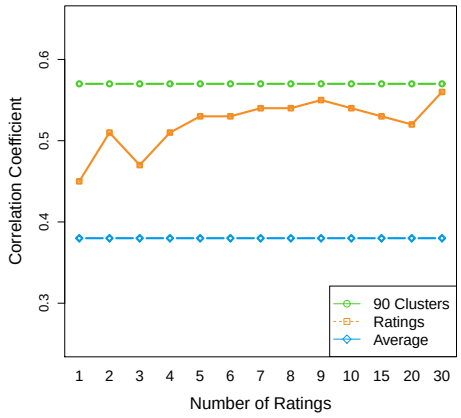

Figure 4: Average correlation coefficient for *unknown* users with an increasing number of ratings. Results from *90 clusters* and *average* ratings are also shown.

6 Evaluation: Stylistically-Aware Surface Realisation

To evaluate the applicability of our regression model for stylistically-adaptive surface realisation, this section describes work that compares four different surface realisers, which were not originally developed to produce stylistic variation. To do that, we first obtain the cluster for each input sentence s: $c^* = \arg\min_{c \in C} \sum_x D(P_s^x | Q_c^x)$, where x refers to n-grams, POS tags or ratings (see Section 5.1); P refers to a discrete probability distribution of sentence s; and Q refers to a discrete probability distribution of cluster c. The best cluster is used to compute the style score of sentence s using: $score(s) = \sum_i^n \theta_i f_i(s), c^* \in F$, where θ_i are the weights estimated by the regressor, and f_i are the features of sentence s; see Table 2. The idea is that if well-phrased utterances can be generated, whose stylistic variation is recognisable to human judges, then our regressor can be used in combination with any statistical surface realiser. Note however that the stylistic variation observed depends on the stylistic spectrum that each realiser covers. Here, our goal is mainly to show that whatever stylistic variation exists in a realiser can be recognised by our model.

6.1 Overview of Surface Realisers

In a human rating study, we compare four surface realisers (ordered alphabetically), all of which are able to return a ranked list of candidate realisations for a semantic input. Please refer to the references given for details of each system. The BAGEL and SPaRKy realisers were compared based on published ranked output lists.[9]

- **BAGEL** is a surface realiser based on dynamic Bayes Nets originally trained using Active Learning by Mairesse et al. (2010). It was shown to generate well-phrased utterances from unseen semantic inputs.
- **CRF (global)** treats surface realisation as a

[9] Available from `http://people.csail.mit.edu/francois/research/bagel` and `http://users.soe.ucsc.edu/~maw/downloads.html`.

System	Utterance
BAGEL	*Beluga is a moderately priced restaurant in the city centre area.* **Col** = 4.0, **Pol** = 4.0, **Nat** = 4.0
CRF (global)	*Set in the city centre, Beluga is a moderately priced location for the celebration of the Italian spirit.* **Col** = 2.0, **Pol** = 5.0, **Nat** = 2.0
p**CRU**	*Beluga is located in the city centre and serves cheap Italian food.* **Col** = 4.0, **Pol** = 3.0, **Nat** = 5.0
SPaRKy	*Beluga has the best overall quality among the selected restaurants since this Italian restaurant has good decor, with good service.* **Col** = 3.0, **Pol** = 4.0, **Nat** = 5.0

Table 5: Example utterances for the **BAGEL**, **CRF (global)**, p**CRU** and **SPaRKy** realisers shown to users. Sample ratings from individual users are also shown.

sequence labelling task: given a set of (observed) linguistic features, it aims to find the best (hidden) sequence of phrases realising a semantic input (Dethlefs et al., 2013).

- p**CRU** is based on probabilistic context-free grammars and generation is done using Viterbi search, sampling (used here), or random search. It is based on Belz (2008).
- **SPaRKy** is based on a rank-and-boost approach. It learns a mapping between the linguistic features of a target utterance and its predicted user ratings and ranks candidates accordingly (Walker et al., 2007).

6.2 Results: Recognising Stylistic Variation

242 users from the USA took part in a rating study on the CrowdFlower platform and rated altogether $1,702$ utterances, from among the highest-ranked surface realisations above. For each utterance they read, they rated the *colloquialism, naturalness* and *politeness* based on the same questions as in Section 4.1, used to obtain the training data. Based on this, we compare the perceived strength of each stylistic dimension in an utterance to the one predicted by the regressor. Example utterances and ratings are shown in Table 5. Results are shown in Table 6 and confirm our observations: ratings for *known* users can be estimated with a medium (or high) correlation based on clusters of users who assign similar ratings to utterances with similar linguistic features. We can also see that such estimations do not depend on a particular data set or realiser.

System	Colloquial	Polite	Natural
BAGEL	0.78	0.66	0.69
CRF global	0.58	0.63	0.63
pCRU	0.67	0.42	0.77
SPaRKy	0.87	0.56	0.81

Table 6: Correlation coefficients between subjective user ratings and ratings predicted by the regressor for *known* users across data-driven surface realisers.

A novel aspect of our technique in comparison to previous work on stylistic realisation is that it does not depend on the time- and resource-intensive design of a hand-coded generator, as in Paiva and Evans (2005) and Mairesse and Walker (2011). Instead, it can be applied in conjunction with any system designer's favourite realiser and preserves the realiser's original features by re-ranking only its top n (e.g. 10) output candidates. Our method is therefore able to strike a balance between highly-ranked and well-phrased utterances and stylistic adaptation. A current limitation of our model is that some ratings can still not be predicted with a high correlation with human judgements. However, even the medium correlations achieved have been shown to be significantly better than estimations based on the average population of users (Section 5.2).

7 Conclusion and Future Work

We have presented a model of stylistic realisation that is able to adapt its output along several stylistic dimensions. Results show that the variation is recognisable by humans and that user ratings can be predicted for *known* as well as *unknown* users. A model which clusters individual users based on their ratings of linguistically similar utterances achieves significantly higher performance than a model trained on the average population of ratings. These results may also play a role in other domains in which users display variability in their subjective ratings, e.g. recommender systems, sentiment analysis, or emotion generation. Future work may explore the use of additional clustering features as a more scalable alternative to re-ranking. It also needs to determine how user feedback can be obtained during an interaction, where asking users for ratings may be disruptive. Possibilities include to infer user ratings from their next dialogue move, or from multimodal information such as hesitations or eye-tracking.

Acknowledgements This research was funded by the EC FP7 programme FP7/2011-14 under grant agreements no. 270019 (SPACEBOOK) and no. 287615 (PARLANCE).

References

Xavier Amatriain, Josep M. Pujol, and Nuria Oliver. 2009. I like It... I Like It Not: Evaluating User Ratings Noise in Recommender Systems. In *In the 17th International Conference on User Modelling, Adaptation, and Personalisation (UMAP)*, pages 247–258, Trento, Italy. Springer-Verlag.

Anja Belz. 2008. Automatic Generation of Weather Forecast Texts Using Comprehensive Probabilistic Generation-Space Models. *Natural Language Engineering*, 14(4):431–455.

Penelope Brown and Stephen Levinson. 1987. *Some Universals in Language Usage*. Cambridge University Press, Cambridge, UK.

Heriberto Cuayáhuitl and Nina Dethlefs. 2011. Optimizing Situated Dialogue Management in Unknown Environments. In *INTERSPEECH*, pages 1009–1012.

Robert Dale and Jette Viethen. 2009. Referring Expression Generation Through Attribute-Based Heuristics. In *Proceedings of the 12th European Workshop on Natural Language Generation (ENLG)*, Athens, Greece.

Nina Dethlefs, Helen Hastie, Verena Rieser, and Oliver Lemon. 2012a. Optimising Incremental Dialogue Decisions Using Information Density for Interactive Systems. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP-CoNLL)*, Jeju, South Korea.

Nina Dethlefs, Helen Hastie, Verena Rieser, and Oliver Lemon. 2012b. Optimising Incremental Generation for Spoken Dialogue Systems: Reducing the Need for Fillers. In *Proceedings of the International Conference on Natural Language Generation (INLG)*, Chicago, Illinois, USA.

Nina Dethlefs, Helen Hastie, Heriberto Cuayáhuitl, and Oliver Lemon. 2013. Conditional Random Fields for Responsive Surface Realisation Using Global Features. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (ACL)*, Sofia, Bulgaria.

Michael Fleischman and Eduard Hovy. 2002. Emotional Variation in Speech-Based Natural Language Generation. In *Proceedings of the 2nd International Natural Language Generation Conference*.

Milica Gašić, Filip Jurčíček, Blaise Thomson, Kai Yu, and Steve Young. 2011. On-Line Policy Optimisation of Spoken Dialogue Systems via Interaction with Human Subjects. In *Proceedings of the IEEE Automatic Speech Recognition and Understanding (ASRU) Workshop*.

Alastair Gill and Jon Oberlander. 2002. Taking Care of the Linguistic Features of Extraversion. In *Proceedings of the 24th Annual Conference of the Cognitive Science Society*, pages 363–368, Fairfax, VA.

Samuel Gosling, Simine Vazire, Sanjay Srivastava, and Oliver John. 2004. Should We Trust Web-Based Studies? A Comparative Analysis of Six Preconceptions About Internet Questionnaires. *American Psychologist*, 59(2):93–104.

Swati Gupta, Marilyn Walker, and Daniela Romano. 2007. How Rude Are You? Evaluating Politeness and Affect in Interaction. In *Proceedings of the 2nd International Conference on Affective Computing and Intelligent Interaction*.

Helen Hastie, Marie-Aude Aufaure, Panos Alexopoulos, Heriberto Cuayhuitl, Nina Dethlefs, James Henderson Milica Gasic, Oliver Lemon, Xingkun Liu, Peter Mika, Nesrine Ben Mustapha, Verena Rieser, Blaise Thomson, Pirros Tsiakoulis, Yves Vanrompay, Boris Villazon-Terrazas, and Steve Young. 2013. Demonstration of the PARLANCE System: A Data-Driven, Incremental, Spoken Dialogue System for Interactive Search. In *Proceedings of the 14th Annual Meeting of the Special Interest Group on Discourse and Dialogue (SIGdial)*.

Eduard Hovy. 1988. *Generating Natural Language under Pragmatic Constraints*. Lawrence Erlbaum Associates, Hillsdale, NJ.

Amy Isard, Carsten Brockmann, and Jon Oberlander. 2006. Individuality and Alignment in Generated Dialogues. In *Proceedings of the 4th International Natural Language Generation Conference (INLG)*, Sydney, Australia.

Srini Janarthanam and Oliver Lemon. 2014. Adaptive generation in dialogue systems using dynamic user modeling. *Computational Linguistics*. (in press).

Pamela Jordan and Marilyn Walker. 2005. Learning Content Selection Rules for Generating Object Descriptions in Dialogue. *Journal of Artificial Intelligence Research*, 24:157–194.

Alexander Koller, Kristina Striegnitz, Donna Byron, Justine Cassell, Robert Dale, and Johanna Moore. 2010. The First Challenge on Generating Instructions in Virtual Environments. In M. Theune and E. Krahmer, editors, *Empirical Methods in Natural Language Generation*, pages 337–361. Springer Verlag, Berlin/Heidelberg.

François Mairesse and Marilyn Walker. 2011. Controlling User Perceptions of Linguistic Style: Trainable Generation of Personality Traits. *Computational Linguistics*, 37(3):455–488, September.

François Mairesse, Milica Gašić, Filip Jurčíček, Simon Keizer, Blaise Thomson, Kai Yu, and Steve Young. 2010. Phrase-based statistical language generation using graphical models and active learning. In *Proceedings of the Annual Meeting of the*

Association for Computational Linguistics (ACL), pages 1552–1561.

Andrew Ng, Michael Jordan, and Yair Weiss. 2001. On Spectral Clustering: Analysis and an Algorithm. In *Advances in Neural Information Processing Systems*, pages 849–856. MIT Press.

Michael O'Mahony, Neil Hurley, and Guénolé Silvestre. 2006. Detecting Noise in Recommender System Databases. In *Proceedings of the International Conference on Intelligent User Interfaces (IUI)s*. ACM Press.

Daniel Paiva and Roger Evans. 2005. Empirically-Based Control of Natural Language Generation. In *Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics (ACL)*, Ann Arbor, Michigan, USA.

Bo Pang and Lillian Lee. 2005. Seeing Stars: Exploiting Class Relationships for Sentiment Categorization with Respect to Rating Scales. In *Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics (ACL)*.

Kaska Porayska-Pomsta and Chris Mellish. 2004. Modelling Politness in Natural Language Generation. In *Proceedings of the 3rd International Natural Language Generation Conference (INLG)*, Brighton, UK.

Verena Rieser, Simon Keizer, Xingkun Liu, and Oliver Lemon. 2011. Adaptive Information Presentation for Spoken Dialogue Systems: Evaluation with Human Subjects. In *Proceedings of the 13th European Workshop on Natural Language Generation (ENLG)*, Nancy, France.

Gabriel Skantze and Anna Hjalmarsson. 2010. Towards Incremental Speech Generation in Dialogue Systems. In *Proceedings of the 11th Annual SigDial Meeting on Discourse and Dialogue*, Tokyo, Japan.

Ulrike von Luxburg. 2007. A Tutorial on Spectral Clustering. *Statistics and Computing*, 17(4).

Marilyn Walker, Amanda Stent, François Mairesse, and Rashmi Prasad. 2007. Individual and Domain Adaptation in Sentence Planning for Dialogue. *Journal of Artificial Intelligence Research*, 30(1):413–456.

Ching-long Yeh and Chris Mellish. 1997. An Empirical Study on the Generation of Anaphora in Chinese. *Computational Linguistics*, 23:169–190.