Trainable Generation of Big-Five Personality Styles through Data-driven Parameter Estimation

François Mairesse
Cambridge University Engineering Department
Trumpington Street
Cambridge, CB2 1PZ, United Kingdom
farm2@eng.cam.ac.uk

Marilyn Walker
Department of Computer Science
University of Sheffield
Sheffield, S1 4DP, United Kingdom
lynwalker@gmail.com

Abstract

Previous work on statistical language generation has primarily focused on grammaticality and naturalness, scoring generation possibilities according to a language model or user feedback. More recent work has investigated data-driven techniques for controlling linguistic style without overgeneration, by reproducing variation dimensions extracted from corpora. Another line of work has produced handcrafted *rule-based* systems to control specific stylistic dimensions, such as politeness and personality. This paper describes a novel approach that automatically learns to produce recognisable variation along a meaningful stylistic dimension—personality—without the computational cost incurred by overgeneration techniques. We present the first evaluation of a data-driven generation method that projects multiple personality traits *simultaneously* and on a *continuous* scale. We compare our performance to a rule-based generator in the same domain.

1 Introduction

Over the last 20 years, statistical language models (SLMs) have been used successfully in many tasks in natural language processing, and the data available for modeling has steadily grown (Lapata and Keller, 2005). Langkilde and Knight (1998) first applied SLMs to statistical natural language generation (SNLG), showing that high quality paraphrases can be generated from an underspecified representation of meaning, by first applying a very underconstrained, rule-based *overgeneration* phase, whose outputs are then ranked by an SLM *scoring* phase. Since then, research in SNLG has explored a range of models for both dialogue and text generation.

One line of work has primarily focused on grammaticality and naturalness, scoring the overgeneration phase with a SLM, and evaluating against a gold-standard corpus, using string or tree-match metrics (Langkilde-Geary, 2002; Bangalore and Rambow, 2000; Chambers and Allen, 2004; Belz, 2005; Isard et al., 2006).

Another thread investigates SNLG scoring models trained using higher-level linguistic features to replicate human judgments of utterance quality (Rambow et al., 2001; Nakatsu and White, 2006; Stent and Guo, 2005). The error of these scoring models approaches the gold-standard human ranking with a relatively small training set.

A third SNLG approach eliminates the overgeneration phase (Paiva and Evans, 2005). It applies factor analysis to a corpus exhibiting stylistic variation, and then learns which generation parameters to manipulate to correlate with factor measurements. The generator was shown to reproduce intended factor levels across several factors, thus modelling the stylistic variation as measured in the original corpus.

Our goal is a generation technique that can target multiple stylistic effects *simultaneously* and over a *continuous* scale, controlling stylistic dimensions that are commonly understood and thus *meaningful* to users and application developers. Our intended applications are output utterances for intelligent training or intervention systems, video game characters, or virtual environment avatars. In previous work, we presented PERSONAGE, a psychologically-informed rule-based generator based on the Big Five personality model, and we showed that PERSONAGE can project extreme personality on the extraversion scale, i.e. both introverted and extraverted personality types (Mairesse and Walker, 2007). We used the Big Five model to develop PERSONAGE for several reasons. First, the Big Five has been shown in psychology to ex-

Trait	High	Low
Extraversion	warm, assertive, sociable, excitement seeking, active, spontaneous, optimistic, talkative	shy, quiet, reserved, passive, solitary, moody
Emotional stability	calm, even-tempered, reliable, peaceful, confident	neurotic, anxious, depressed, self-conscious
Agreeableness	trustworthy, considerate, friendly, generous, helpful	unfriendly, selfish, suspicious, uncooperative, malicious
Conscientiousness	competent, disciplined, dutiful, achievement striving	disorganised, impulsive, unreliable, forgetful
Openness to experience	creative, intellectual, curious, cultured, complex	narrow-minded, conservative, ignorant, simple

Table 1: Example adjectives associated with extreme values of the Big Five trait scales.

plain much of the variation in human perceptions of personality differences. Second, we believe that the adjectives used to develop the Big Five model provide an intuitive, *meaningful* definition of linguistic style. Table 1 shows some of the trait adjectives associated with the extremes of each Big Five trait. Third, there are many studies linking personality to linguistic variables (Pennebaker and King, 1999; Mehl et al., 2006, *inter alia)*. See (Mairesse and Walker, 2007) for more detail.

In this paper, we further test the utility of basing stylistic variation on the Big Five personality model. The Big Five traits are represented by scalar values that range from 1 to 7, with values normally distributed among humans. While our previous work targeted extreme values of individual traits, here we show that we can target multiple personality traits *simultaneously* and over the *continuous* scales of the Big Five model. Section 2 describes a novel parameter-estimation method that automatically learns to produce recognisable variation for all Big Five traits, without overgeneration, implemented in a new SNLG called PERSONAGE-PE. We show that PERSONAGE-PE generates targets for multiple personality dimensions, using linear and non-linear parameter estimation models to predict generation parameters *directly* from the scalar targets. Section 3.2 shows that humans accurately perceive the intended variation, and Section 3.3 compares PERSONAGE-PE (trained) with PERSONAGE (rule-based; Mairesse and Walker, 2007). We delay a detailed discussion of related work to Section 4, where we summarize and discuss future work.

2 Parameter Estimation Models

The data-driven parameter estimation method consists of a development phase and a generation phase (Section 3). The development phase:

1. Uses a base generator to produce multiple utterances by randomly varying its parameters;
2. Collects human judgments rating the personality of each utterance;
3. Trains statistical models to predict the parameters from the personality judgments;

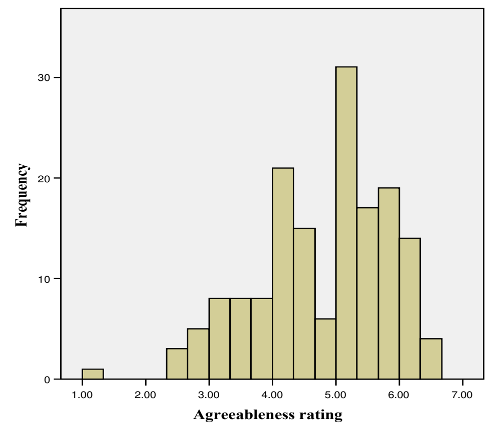

Figure 1: Distribution of average agreeableness ratings from the 2 expert judges for 160 random utterances.

4. Selects the best model for each parameter via cross-validation.

2.1 Base Generator

We make minimal assumptions about the input to the generator to favor domain independence. The input is a speech act, a potential content pool that can be used to achieve that speech act, and five scalar personality parameters (1. . .7), specifying values for the continuous scalar dimensions of each trait in the Big Five model. See Table 1. This requires a base generator that generates multiple outputs expressing the same input content by varying linguistic parameters related to the Big Five traits. We start with the PERSONAGE generator (Mairesse and Walker, 2007), which generates recommendations and comparisons of restaurants. We extend PERSONAGE with new parameters for a total of 67 parameters in PERSONAGE-PE. See Table 2. These parameters are derived from psychological studies identifying linguistic markers of the Big Five traits (Pennebaker and King, 1999; Mehl et al., 2006, *inter alia*). As PERSONAGE's input parameters are domain-independent, most parameters range continuously between 0 and 1, while pragmatic marker insertion parameters are binary, except for the SUBJECT IMPLICITNESS, STUTTERING and PRONOMI-

Parameters	Description
Content parameters:	
VERBOSITY	Control the number of propositions in the utterance
RESTATEMENTS	Paraphrase an existing proposition, e.g. *'Chanpen Thai has great service, it has fantastic waiters'*
REPETITIONS	Repeat an existing proposition
CONTENT POLARITY	Control the polarity of the propositions expressed, i.e. referring to negative or positive attributes
REPETITIONS POLARITY	Control the polarity of the restated propositions
CONCESSIONS	Emphasise one attribute over another, e.g. *'even if Chanpen Thai has great food, it has bad service'*
CONCESSIONS POLARITY	Determine whether positive or negative attributes are emphasised
POLARISATION	Control whether the expressed polarity is neutral or extreme
POSITIVE CONTENT FIRST	Determine whether positive propositions—including the claim—are uttered first
Syntactic template selection parameters:	
SELF-REFERENCES	Control the number of first person pronouns
CLAIM COMPLEXITY	Control the syntactic complexity (syntactic embedding)
CLAIM POLARITY	Control the connotation of the claim, i.e. whether positive or negative affect is expressed
Aggregation operations:	
PERIOD	Leave two propositions in their own sentences, e.g. *'Chanpen Thai has great service. It has nice decor.'*
RELATIVE CLAUSE	Aggregate propositions with a relative clause, e.g. *'Chanpen Thai, which has great service, has nice decor'*
WITH CUE WORD	Aggregate propositions using *with*, e.g. *'Chanpen Thai has great service, with nice decor'*
CONJUNCTION	Join two propositions using a conjunction, or a comma if more than two propositions
MERGE	Merge the subject and verb of two propositions, e.g. *'Chanpen Thai has great service and nice decor'*
ALSO CUE WORD	Join two propositions using *also*, e.g. *'Chanpen Thai has great service, also it has nice decor'*
CONTRAST - CUE WORD	Contrast two propositions using *while, but, however, on the other hand*, e.g. *'While Chanpen Thai has great service, it has bad decor'*, *'Chanpen Thai has great service, but it has bad decor'*
JUSTIFY - CUE WORD	Justify a proposition using *because, since, so*, e.g. *'Chanpen Thai is the best, because it has great service'*
CONCEDE - CUE WORD	Concede a proposition using *although, even if, but/though*, e.g. *'Although Chanpen Thai has great service, it has bad decor'*, *'Chanpen Thai has great service, but it has bad decor though'*
MERGE WITH COMMA	Restate a proposition by repeating only the object, e.g. *'Chanpen Thai has great service, nice waiters'*
CONJ. WITH ELLIPSIS	Restate a proposition after replacing its object by an ellipsis, e.g. *'Chanpen Thai has . . . , it has great service'*
Pragmatic markers:	
SUBJECT IMPLICITNESS	Make the restaurant implicit by moving the attribute to the subject, e.g. *'the service is great'*
NEGATION	Negate a verb by replacing its modifier by its antonym, e.g. *'Chanpen Thai doesn't have bad service'*
SOFTENER HEDGES	Insert syntactic elements (*sort of*, *kind of*, *somewhat*, *quite*, *around*, *rather*, *I think that*, *it seems that*, *it seems to me that*) to mitigate the strength of a proposition, e.g. *'Chanpen Thai has kind of great service'* or *'It seems to me that Chanpen Thai has rather great service'*
EMPHASIZER HEDGES	Insert syntactic elements (*really*, *basically*, *actually*, *just*) to strengthen a proposition, e.g. *'Chanpen Thai has really great service'* or *'Basically, Chanpen Thai just has great service'*
ACKNOWLEDGMENTS	Insert an initial back-channel (*yeah*, *right*, *ok*, *I see*, *oh*, *well*), e.g. *'Well, Chanpen Thai has great service'*
FILLED PAUSES	Insert syntactic elements expressing hesitancy (*like*, *I mean*, *err*, *mmhm*, *you know*), e.g. *'I mean, Chanpen Thai has great service, you know'* or *'Err... Chanpen Thai has, like, great service'*
EXCLAMATION	Insert an exclamation mark, e.g. *'Chanpen Thai has great service!'*
EXPLETIVES	Insert a swear word, e.g. *'the service is damn great'*
NEAR-EXPLETIVES	Insert a near-swear word, e.g. *'the service is darn great'*
COMPETENCE MITIGATION	Express the speaker's negative appraisal of the hearer's request, e.g. *'everybody knows that . . . '*
TAG QUESTION	Insert a tag question, e.g. *'the service is great, isn't it?'*
STUTTERING	Duplicate the first letters of a restaurant's name, e.g. *'Ch-ch-anpen Thai is the best'*
CONFIRMATION	Begin the utterance with a confirmation of the restaurant's name, e.g. *'did you say Chanpen Thai?'*
INITIAL REJECTION	Begin the utterance with a mild rejection, e.g. *'I'm not sure'*
IN-GROUP MARKER	Refer to the hearer as a member of the same social group, e.g. *pal*, *mate* and *buddy*
PRONOMINALIZATION	Replace occurrences of the restaurant's name by pronouns
Lexical choice parameters:	
LEXICAL FREQUENCY	Control the average frequency of use of each content word, according to BNC frequency counts
WORD LENGTH	Control the average number of letters of each content word
VERB STRENGTH	Control the strength of the selected verbs, e.g. *'I would suggest'* vs. *'I would recommend'*

Table 2: The 67 generation parameters whose target values are learned. Aggregation cue words, hedges, acknowledgments and filled pauses are learned individually (as separate parameters), e.g. *kind of* is modeled differently than *somewhat* in the SOFTENER HEDGES category. Parameters are detailed in previous work (Mairesse and Walker, 2007).

NALIZATION parameters.

2.2 Random Sample Generation and Expert Judgments

We generate a sample of 160 *random* utterances by varying the parameters in Table 2 with a uniform distribution. This sample is intended to provide enough training material for estimating all 67 parameters for each personality dimension. Following Mairesse and Walker (2007), two expert judges (not the authors) familiar with the Big Five adjectives (Table 1) evaluate the personality of each utterance using the Ten-Item Personality Inventory (TIPI; Gosling et al., 2003), and also judge the utterance's naturalness. Thus 11 judgments were made for each utterance for a total of 1760 judgments. The TIPI outputs a rating on a scale from 1 (low) to 7 (high) for each Big Five trait. The expert judgments are approximately normally distributed; Figure 1 shows the distribution for agreeableness.

2.3 Statistical Model Training

Training data is created for each generation parameter—i.e. the output variable—to train statistical models predicting the optimal parameter value from the target personality scores. The models are thus based on the simplifying assumption that the generation parameters are independent. Any personality trait whose correlation with a generation decision is below 0.1 is removed from the training data. This has the effect of removing parameters that do not correlate strongly with any trait, which are set to a constant default value at generation time. Since the input parameter values may not be satisfiable depending on the input content, the actual generation decisions made for each utterance are recorded. For example, the CONCESSIONS decision value is the actual number of concessions produced in the utterance. To ensure that the models' output can control the generator, the generation decision values are normalized to match the input range (0...1) of PERSONAGE-PE. Thus the dataset consists of 160 utterances and the corresponding generation decisions, each associated with 5 personality ratings averaged over both judges.

Parameter estimation models are trained to predict either continuous (e.g. VERBOSITY) or binary (e.g. EXCLAMATION) generation decisions. We compare various learning algorithms using the Weka toolkit (with default values unless specified; Witten and Frank, 2005). Continuous parameters are modeled with a linear regression model (LR), an M5' model tree (M5), and a model based on support vector machines with a linear kernel (SVM). As regression models can extrapolate beyond the $[0, 1]$ interval, the output parameter values are truncated if needed—at generation time—before being sent to the base generator. Binary parameters are modeled using classifiers that predict whether the parameter is *enabled* or *disabled*. We test a Naive Bayes classifier (NB), a j48 decision tree (J48), a nearest-neighbor classifier using one neighbor (NN), a Java implementation of the RIPPER rule-based learner (JRIP), the AdaBoost boosting algorithm (ADA), and a support vector machines classifier with a linear kernel (SVM).

Figures 2, 3 and 4 show the models learned for the EXCLAMATION (binary), STUTTERING (continuous), and CONTENT POLARITY (continuous) parameters in Table 2. The models predict generation parameters from input personality scores; note that sometimes the best performing model is non-linear. Given input trait values, the AdaBoost model in Figure 2 outputs the class yielding the largest sum of weights for the rules returning that class. For example, the first rule of the EXCLAMATION model shows that an extraversion score above 6.42 out of 7 would increase the weight of the *enabled* class by 1.81. The fifth rule indicates that a target agreeableness above 5.13 would further increase the weight by .42. The STUTTERING model tree in Figure 4 lets us calculate that a low emotional stability (1.0) together with a neutral conscientiousness and openness to experience (4.0) yield a parameter value of .62 (see LM2), whereas a neutral emotional stability decreases the value down to .17. Figure 4 also shows how personality traits that do not affect the parameter are removed, i.e. emotional stability, conscientiousness and openness to experience are the traits that affect stuttering. The linear model in Figure 3 shows that agreeableness has a strong effect on the CONTENT POLARITY parameter (.97 weight), but emotional stability, conscientiousness and openness to experience also have an effect.

```
Condition                         Class      Weight
---------                         -----      ------
if extraversion > 6.42  then 1 else 0    1.81
if extraversion > 4.42  then 1 else 0    0.38
if extraversion <= 6.58 then 1 else 0    0.22
if extraversion > 4.71  then 1 else 0    0.28
if agreeableness > 5.13 then 1 else 0    0.42
if extraversion <= 6.58 then 1 else 0    0.14
if extraversion > 4.79  then 1 else 0    0.19
if extraversion <= 6.58 then 1 else 0    0.17
```

Figure 2: AdaBoost model predicting the EXCLAMATION parameter. Given input trait values, the model outputs the class yielding the largest sum of weights for the rules returning that class. Class 0 = *disabled*, class 1 = *enabled*.

```
(normalized) Content polarity  =
   0.054
 - 0.102 * (normalized) emotional stability
 + 0.970 * (normalized) agreeableness
 - 0.110 * (normalized) conscientiousness
 + 0.013 * (normalized) openness to
                        experience
```

Figure 3: SVM model with a linear kernel predicting the CONTENT POLARITY parameter.

2.4 Model Selection

The final step of the development phase identifies the best performing model(s) for each generation parameter via cross-validation. For continuous pa-

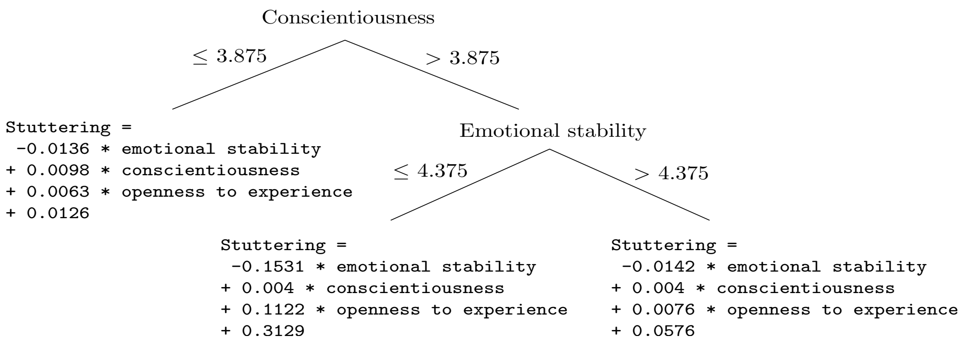

Figure 4: M5' model tree predicting the STUTTERING parameter.

Continuous parameters	**LR**	**M5**	**SVM**
Content parameters:			
VERBOSITY	0.24	**0.26**	0.21
RESTATEMENTS	0.14	**0.14**	0.04
REPETITIONS	**0.13**	0.13	0.08
CONTENT POLARITY	0.46	0.46	**0.47**
REPETITIONS POLARITY	0.02	**0.15**	0.06
CONCESSIONS	**0.23**	0.23	0.12
CONCESSIONS POLARITY	-0.01	**0.16**	0.07
POLARISATION	0.20	**0.21**	0.20
Syntactic template selection:			
CLAIM COMPLEXITY	0.10	**0.33**	0.26
CLAIM POLARITY	0.04	0.04	**0.05**
Aggregation operations:			
INFER - WITH CUE WORD	0.03	**0.03**	0.01
INFER - ALSO CUE WORD	**0.10**	0.10	0.06
JUSTIFY - SINCE CUE WORD	0.03	**0.07**	0.05
JUSTIFY - SO CUE WORD	0.07	**0.07**	0.04
JUSTIFY - PERIOD	**0.36**	0.35	0.21
CONTRAST - PERIOD	**0.27**	0.26	0.26
RESTATE - MERGE WITH COMMA	0.18	**0.18**	0.09
CONCEDE - ALTHOUGH CUE WORD	**0.08**	0.08	0.05
CONCEDE - EVEN IF CUE WORD	0.05	**0.05**	0.03
Pragmatic markers:			
SUBJECT IMPLICITNESS	**0.13**	0.13	0.04
STUTTERING INSERTION	0.16	**0.23**	0.17
PRONOMINALIZATION	**0.22**	0.20	0.17
Lexical choice parameters:			
LEXICAL FREQUENCY	0.21	**0.21**	0.19
WORD LENGTH	**0.18**	0.18	0.15

Table 3: Pearson's correlation between parameter model predictions and continuous parameter values, for different regression models. Parameters that do not correlate with any trait are omitted. Aggregation operations are associated with a rhetorical relation (e.g. INFER). Results are averaged over a 10-fold cross-validation.

Binary parameters	**NB**	**J48**	**NN**	**ADA**	**SVM**
Pragmatic markers:					
SOFTENER HEDGES					
kind of	0.00	0.00	**0.16**	0.11	0.10
rather	0.00	0.00	**0.02**	0.01	0.01
quite	**0.14**	0.08	0.09	0.07	0.06
EMPHASIZER HEDGES					
basically	0.00	0.00	**0.02**	0.01	0.01
ACKNOWLEDGMENTS					
yeah	0.00	0.00	**0.04**	0.03	0.03
ok	**0.13**	0.07	0.06	0.05	0.05
FILLED PAUSES					
err	**0.32**	0.20	0.24	0.22	0.19
EXCLAMATION	0.23	0.34	0.36	**0.38**	0.34
EXPLETIVES	**0.27**	0.18	0.24	0.17	0.15
IN-GROUP MARKER	**0.40**	0.31	0.31	0.24	0.21
TAG QUESTION	**0.32**	0.21	0.21	0.15	0.13
CONFIRMATION	0.00	0.00	**0.07**	0.04	0.04

Table 4: F-measure of the *enabled* class for classification models of binary parameters. Parameters that do not correlate with any trait are omitted. Results are averaged over a 10-fold cross-validation. JRIP models are not shown as they never perform best.

rameters, Table 3 evaluates modeling accuracy by comparing the correlations between the model's predictions and the actual parameter values in the test folds. Table 4 reports results for binary parameter classifiers, by comparing the F-measures of the *enabled* class. Best performing models are identified in bold; parameters that do not correlate with any trait or that produce a poor modeling accuracy are omitted.

The CONTENT POLARITY parameter is modeled the most accurately, with the SVM model in Figure 3 producing a correlation of .47 with the true parameter values. Models of the PERIOD aggregation operation also perform well, with a linear regression model yielding a correlation of .36 when realizing a justification, and .27 when contrasting two propositions. CLAIM COMPLEXITY and VERBOSITY are also modeled successfully, with correlations of .33 and .26 using a model tree. The model tree controlling the STUTTERING parameter illustrated in Figure 4 produces a correlation of .23. For binary parameters, Table 4 shows that the Naive Bayes classifier is generally the most accurate, with F-measures of .40 for the IN-GROUP MARKER parameter, and .32 for both the insertion of filled pauses (*err*) and tag questions. The AdaBoost algorithm best predicts the EXCLAMATION parameter, with an F-measure of .38 for the model in Figure 2.

#	Traits	End	Rating	Nat	Output utterance
1.a	Extraversion Agreeableness	high high	4.42 4.94	4.79	Radio Perfecto's price is 25 dollars but Les Routiers provides adequate food. I imagine they're alright!
1.b	Emotional stability Conscientiousness	high high	5.35 5.21	5.04	Let's see, Les Routiers and Radio Perfecto... You would probably appreciate them. Radio Perfecto is in the East Village with kind of acceptable food. Les Routiers is located in Manhattan. Its price is 41 dollars.
2.a	Extraversion Agreeableness	low low	3.65 4.02	3.21	Err... you would probably appreciate Trattoria Rustica, wouldn't you? It's in Manhattan, also it's an italian restaurant. It offers poor ambience, also it's quite costly.
2.b	Emotional stability Openness to experience	low low	4.13 3.85	4.50	Trattoria Rustica isn't as bad as the others. Err... even if it's costly, it offers kind of adequate food, alright? It's an italian place.

Table 5: Example outputs controlled by the parameter estimation models for a comparison (#1) and a recommendation (#2), with the average judges' ratings (*Rating*) and naturalness (*Nat*). Ratings are on a scale from 1 to 7, with 1 = very low (e.g. neurotic or introvert) and 7 = very high on the dimension (e.g. emotionally stable or extraverted).

3 Evaluation Experiment

The generation phase of our parameter estimation SNLG method consists of the following steps:

1. Use the best performing models to predict parameter values from the desired personality scores;
2. Generate the output utterance using the predicted parameter values.

We then evaluate the output utterances using naive human judges to rate their perceived personality and naturalness.

3.1 Evaluation Method

Given the best performing model for each generation parameter, we generate 5 utterances for each of 5 recommendation and 5 comparison speech acts. Each utterance targets an extreme value for two traits (either 1 or 7 out of 7) and neutral values for the remaining three traits (4 out of 7). The goal is for each utterance to project *multiple* traits on a *continuous* scale. To generate a range of alternatives, a Gaussian noise with a standard deviation of 10% of the full scale is added to each target value.

Subjects were 24 native English speakers (12 male and 12 female graduate students from a range of disciplines from both the U.K. and the U.S.). Subjects evaluate the naturalness and personality of each utterance using the TIPI (Gosling et al., 2003). To limit the experiment's duration, only the two traits with extreme target values are evaluated for each utterance. Subjects thus answered 5 questions for 50 utterances, two from the TIPI for each extreme trait and one about naturalness (250 judgments in total per subject). Subjects were not told that the utterances were intended to manifest extreme trait values. Table 5 shows several sample outputs and the mean personality ratings from the human judges. For example, utterance 1.a projects a high extraversion through the insertion of an exclamation mark based on the model in Figure 2, whereas utterance 2.a conveys introversion by beginning with the filled pause *err*. The same utterance also projects a low agreeableness by focusing on negative propositions, through a low CONTENT POLARITY parameter value as per the model in Figure 3. This evaluation addresses a number of open questions discussed below.

Q1: Is the personality projected by models trained on ratings from a few expert judges recognised by a larger sample of naive judges? (Section 3.2)

Q2: Can a *combination* of multiple traits within a single utterance be detected by naive judges? (Section 3.2)

Q3: How does PERSONAGE-PE compare to PERSONAGE, a psychologically-informed rule-based generator for projecting extreme personality? (Section 3.3)

Q4: Does the parameter estimation SNLG method produce natural utterances? (Section 3.4)

3.2 Parameter Estimation Evaluation

Table 6 shows that extraversion is the dimension modeled most accurately by the parameter estimation models, producing a .45 correlation with the subjects' ratings ($p < .01$). Emotional stability, agreeableness, and openness to experience ratings also correlate strongly with the target scores, with correlations of .39, .36 and .17 respectively ($p < .01$). Additionally, Table 6 shows that the magnitude of the correlation increases when considering the perception of a hypothetical average subject, i.e. smoothing individual variation by averaging the ratings over all 24 judges, producing a correlation r_{avg} up to .80 for extraversion. These correlations are unexpectedly high; in corpus analyses, significant correlations as low as .05 to .10 are typically observed between personality and linguistic markers (Pennebaker and King, 1999; Mehl et al., 2006).

Conscientiousness is the only dimension whose ratings do not correlate with the target scores. The comparison with rule-based results in Section 3.3 suggests that this is not because conscientiousness cannot be exhibited in our domain or manifested in a single utterance, so perhaps this arises from differing perceptions of conscientiousness between the expert and naive judges.

Trait	r	r_{avg}	e
Extraversion	.45 •	.80 •	1.89
Emotional stability	.39 •	.64 •	2.14
Agreeableness	.36 •	.68 •	2.38
Conscientiousness	-.01	-.02	2.79
Openness to experience	.17 •	.41 •	2.51

• statistically significant correlation
$p < .05$, • $p = .07$ (two-tailed)

Table 6: Pearson's correlation coefficient r and mean absolute error e between the target personality scores and the 480 judges' ratings (20 ratings per trait for 24 judges); r_{avg} is the correlation between the personality scores and the average judges' ratings.

Table 6 shows that the mean absolute error varies between 1.89 and 2.79 on a scale from 1 to 7. Such large errors result from the decision to ask judges to answer just the TIPI questions for the two traits that were the extreme targets (See Section 3.1), because the judges tend to use the whole scale, with approximately normally distributed ratings. This means that although the judges make distinctions leading to high correlations, they do so on a compressed scale. This explains the large correlations despite the magnitude of the absolute error.

Table 7 shows results evaluating whether utterances targeting the extremes of a trait are perceived differently. The ratings differ significantly for all traits but conscientiousness ($p \leq .001$). Thus parameter estimation models can be used in applications that only require discrete binary variation.

Trait	Low	High
Extraversion	3.69	5.06 •
Emotional stability	3.75	4.75 •
Agreeableness	3.42	4.33 •
Conscientiousness	4.16	4.15
Openness to experience	3.71	4.06 •

• statistically significant difference
$p \leq .001$ (two-tailed)

Table 7: Average personality ratings for the utterances generated with the low and high target values for each trait on a scale from 1 to 7.

It is important to emphasize that generation parameters were predicted based on 5 target personality values. Thus, the results show that *individual* traits are perceived even when utterances project other traits as well, confirming that the Big Five theory models independent dimensions and thus provides a useful and meaningful framework for modeling variation in language. Additionally, although we do not directly evaluate the perception of mid-range values of personality target scores, the results suggest that mid-range personality is modeled correctly because the neutral target scores do not affect the perception of extreme traits.

3.3 Comparison with Rule-Based Generation

PERSONAGE is a rule-based personality generator based on handcrafted parameter settings derived from psychological studies. Mairesse and Walker (2007) show that this approach generates utterances that are perceptibly different along the extraversion dimension. Table 8 compares the mean ratings of the utterances generated by PERSONAGE-PE with ratings of 20 utterances generated by PERSONAGE for each extreme of each Big Five scale (40 for extraversion, resulting in 240 handcrafted utterances in total). Table 8 shows that the handcrafted parameter settings project a significantly more extreme personality for 6 traits out of 10. However, the learned parameter models for neuroticism, disagreeableness, unconscientiousness and openness to experience do not perform significantly worse than the handcrafted generator. These findings are promising as we discuss further in Section 4.

Method	**Rule-based**		**Learned parameters**	
Trait	**Low**	**High**	**Low**	**High**
Extraversion	2.96	5.98	3.69 ∘	5.05 ∘
Emotional stability	3.29	5.96	3.75	4.75 ∘
Agreeableness	3.41	5.66	3.42	4.33 ∘
Conscientiousness	3.71	5.53	4.16	4.15 ∘
Openness to experience	2.89	4.21	3.71 ∘	4.06

•,∘ significant increase or decrease of the variation range over the average rule-based ratings ($p < .05$, two-tailed)

Table 8: Pair-wise comparison between the ratings of the utterances generated using PERSONAGE-PE with extreme target values (*Learned Parameters*), and the ratings for utterances generated with Mairesse and Walker's rule-based PERSONAGE generator, (*Rule-based*). Ratings are averaged over all judges.

3.4 Naturalness Evaluation

The naive judges also evaluated the naturalness of the outputs of our trained models. Table 9 shows that the average naturalness is 3.98 out of 7, which is significantly lower (p $<$ .05) than the naturalness of handcrafted and randomly generated utterances reported by Mairesse and Walker (2007). It is possible that the differences arise from judgments of utterances targeting multiple traits, or that the naive judges are more critical.

Trait	Rule-based	Random	Learned
All	4.59	4.38	3.98

Table 9: Average naturalness ratings for utterances generated using (1) PERSONAGE, the rule-based generator, (2) the random utterances (expert judges) and (3) the outputs of PERSONAGE-PE using the parameter estimation models (*Learned*, naive judges). The means differ significantly at the $p < .05$ level (two-tailed independent sample t-test).

4 Conclusion

We present a new method for generating linguistic variation projecting multiple personality traits continuously, by combining and extending previous research in statistical natural language generation (Paiva and Evans, 2005; Rambow et al., 2001; Isard et al., 2006; Mairesse and Walker, 2007). While handcrafted rule-based approaches are limited to variation along a small number of discrete points (Hovy, 1988; Walker et al., 1997; Lester et al., 1997; Power et al., 2003; Cassell and Bickmore, 2003; Piwek, 2003; Mairesse and Walker, 2007; Rehm and André, in press), we learn models that predict parameter values for any arbitrary value on the variation dimension scales. Additionally, our data-driven approach can be applied to any dimension that is meaningful to human judges, and it provides an elegant way to project multiple dimensions simultaneously, by including the relevant dimensions as features of the parameter models' training data.

Isard et al. (2006) and Mairesse and Walker (2007) also propose a personality generation method, in which a data-driven personality model selects the best utterance from a large candidate set. Isard et al.'s technique has not been evaluated, while Mairesse and Walker's overgenerate and score approach is inefficient. Paiva and Evans' technique does not overgenerate (2005), but it requires a search for the optimal generation decisions according to the learned models. Our approach does not require any search or overgeneration, as parameter estimation models predict the generation decisions directly from the target variation dimensions. This technique is therefore beneficial for real-time generation. Moreover the variation dimensions of Paiva and Evans' data-driven technique are extracted from a corpus: there is thus no guarantee that they can be easily interpreted by humans, and that they generalise to other corpora. Previous work has shown that modeling the relation between personality and language is far from trivial (Pennebaker and King, 1999; Argamon et al., 2005; Oberlander and Nowson, 2006; Mairesse et al., 2007), suggesting that the control of personality is a harder problem than the control of data-driven variation dimensions.

We present the first human perceptual evaluation of a data-driven stylistic variation method. In terms of our research questions in Section 3.1, we show that models trained on expert judges to project multiple traits in a single utterance generate utterances whose personality is recognized by naive judges. There is only one other similar evaluation of an SNLG (Rambow et al., 2001). Our models perform only slightly worse than a handcrafted rule-based generator in the same domain. These findings are promising as (1) parameter estimation models are able to target any combination of traits over the full range of the Big Five scales; (2) they do not benefit from psychological knowledge, i.e. they are trained on randomly generated utterances.

This work also has several limitations that should be addressed in future work. Even though the parameters of PERSONAGE-PE were suggested by psychological studies (Mairesse and Walker, 2007), some of them are not modeled successfully by our approach, and thus omitted from Tables 3 and 4. This could be due to the relatively small development dataset size (160 utterances to optimize 67 parameters), or to the implementation of some parameters. The strong parameter-independence assumption could also be responsible, but we are not aware of any state of the art implementation for learning multiple dependent variables, and this approach could further aggravate data sparsity issues.

In addition, it is unclear why PERSONAGE performs better for projecting extreme personality and produces more natural utterances, and why PERSONAGE-PE fails to project conscientiousness correctly. It might be possible to improve the parameter estimation models with a larger sample of random utterances at development time, or with additional extreme data generated using the rule-based approach. Such hybrid models are likely to perform better for extreme target scores, as they are trained on more uniformly distributed ratings (e.g. compared to the normal distribution in Figure 1). In addition, we have only shown that personality can be expressed by information presentation speech-acts in the restaurant domain; future work should assess the extent to which the parameters derived from psychological findings are culture, domain, and speech act dependent.

References

S. Argamon, S. Dhawle, M. Koppel, and J. Pennebaker. Lexical predictors of personality type. In *Proceedings of the Joint Annual Meeting of the Interface and the Classification Society of North America*, 2005.

S. Bangalore and O. Rambow. Exploiting a probabilistic hierarchical model for generation. In *Proceedings of the 18th International Conference on Computational Linguistics (COLING)*, pages 42–48, 2000.

A. Belz. Corpus-driven generation of weather forecasts. In *Proceedings of the 3rd Corpus Linguistics Conference*, 2005.

J. Cassell and T. Bickmore. Negotiated collusion: Modeling social language and its relationship effects in intelligent agents. *User Modeling and User-Adapted Interaction*, 13:89–132, 2003.

N. Chambers and J. Allen. Stochastic language generation in a dialogue system: Toward a domain independent generator. In *Proceedings 5th SIGdial Workshop on Discourse and Dialogue*, 2004.

S. D. Gosling, P. J. Rentfrow, and W. B. Swann. A very brief measure of the big five personality domains. *Journal of Research in Personality*, 37:504–528, 2003.

E. Hovy. *Generating Natural Language under Pragmatic Constraints*. Lawrence Erlbaum Associates, 1988.

A. Isard, C. Brockmann, and J. Oberlander. Individuality and alignment in generated dialogues. In *Proceedings of the 4th International Natural Language Generation Conference (INLG)*, pages 22–29, 2006.

I. Langkilde and K. Knight. Generation that exploits corpus-based statistical knowledge. In *Proceedings of the 36th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 704–710, 1998.

I. Langkilde-Geary. An empirical verification of coverage and correctness for a general-purpose sentence generator. In *Proceedings of the 1st International Conference on Natural Language Generation*, 2002.

M. Lapata and F. Keller. Web-based models for natural language processing. *ACM Transactions on Speech and Language Processing*, 2:1–31, 2005.

J. Lester, S. Converse, S. Kahler, S. Barlow, B. Stone, and R. Bhogal. The persona effect: affective impact of animated pedagogical agents. *Proceedings of the SIGCHI conference on Human factors in computing systems*, pages 359–366, 1997.

F. Mairesse and M. A. Walker. PERSONAGE: Personality generation for dialogue. In *Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 496–503, 2007.

F. Mairesse, M. A. Walker, M. R. Mehl, and R. K. Moore. Using linguistic cues for the automatic recognition of personality in conversation and text. *Journal of Artificial Intelligence Research (JAIR)*, 30:457–500, 2007.

M. R. Mehl, S. D. Gosling, and J. W. Pennebaker. Personality in its natural habitat: Manifestations and implicit folk theories of personality in daily life. *Journal of Personality and Social Psychology*, 90:862–877, 2006.

C. Nakatsu and M. White. Learning to say it well: Reranking realizations by predicted synthesis quality. In *Proceedings of the 44th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 1113–1120, 2006.

J. Oberlander and S. Nowson. Whose thumb is it anyway? classifying author personality from weblog text. In *Proceedings of the 44th Annual Meeting of the Association for Computational Linguistics (ACL)*, 2006.

D. S. Paiva and R. Evans. Empirically-based control of natural language generation. In *Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 58–65, 2005.

J. W. Pennebaker and L. A. King. Linguistic styles: Language use as an individual difference. *Journal of Personality and Social Psychology*, 77:1296–1312, 1999.

P. Piwek. A flexible pragmatics-driven language generator for animated agents. In *Proceedings of Annual Meeting of the European Chapter of the Association for Computational Linguistics (EACL)*, 2003.

R. Power, D. Scott, and N. Bouayad-Agha. Generating texts with style. In *Proceedings of the 4th International Conference on Intelligent Text Processing and Computational Linguistics*, 2003.

O. Rambow, M. Rogati, and M. A. Walker. Evaluating a trainable sentence planner for a spoken dialogue travel system. In *Proceedings of the 39th Annual Meeting of the Association for Computational Linguistics (ACL)*, 2001.

M. Rehm and E. André. From annotated multimodal corpora to simulated human-like behaviors. In I. Wachsmuth and G. Knoblich, editors, *Modeling Communication with Robots and Virtual Humans*. Springer, Berlin, Heidelberg, in press.

A. Stent and H. Guo. A new data-driven approach for multimedia presentation generation. In *Proc. EuroIMSA*, 2005.

M. A. Walker, J. E. Cahn, and S. J. Whittaker. Improvising linguistic style: Social and affective bases for agent personality. In *Proceedings of the 1st Conference on Autonomous Agents*, pages 96–105, 1997.

I. H. Witten and E. Frank. *Data Mining: Practical machine learning tools and techniques*. Morgan Kaufmann, San Francisco, CA, 2005.