PERSONAGE: Personality Generation for Dialogue

François Mairesse
Department of Computer Science
University of Sheffield
Sheffield, S1 4DP, United Kingdom
F.Mairesse@sheffield.ac.uk

Marilyn Walker
Department of Computer Science
University of Sheffield
Sheffield, S1 4DP, United Kingdom
M.A.Walker@sheffield.ac.uk

Abstract

Over the last fifty years, the "Big Five" model of personality traits has become a standard in psychology, and research has systematically documented correlations between a wide range of linguistic variables and the Big Five traits. A distinct line of research has explored methods for automatically generating language that varies along personality dimensions. We present PERSONAGE (PERSONAlity GEnerator), the first highly parametrizable language generator for extraversion, an important aspect of personality. We evaluate two personality generation methods: (1) direct generation with particular parameter settings suggested by the psychology literature; and (2) overgeneration and selection using statistical models trained from judge's ratings. Results show that both methods reliably generate utterances that vary along the extraversion dimension, according to human judges.

1 Introduction

Over the last fifty years, the "Big Five" model of personality traits has become a standard in psychology (extraversion, neuroticism, agreeableness, conscientiousness, and openness to experience), and research has systematically documented correlations between a wide range of linguistic variables and the Big Five traits (Mehl et al., 2006; Norman, 1963; Oberlander and Gill, 2006; Pennebaker and King, 1999). A distinct line of research has explored methods for automatically generating language that varies along personality dimensions, targeting applications such as computer gaming and educational virtual worlds (André et al., 2000; Isard et al., 2006; Loyall and Bates, 1997; Piwek, 2003; Walker et al., 1997) *inter alia*. Other work suggests a clear utility for generating language manifesting personality (Reeves and Nass, 1996). However, to date, (1) research in generation has not systematically exploited the psycholinguistic findings; and (2) there has been little evaluation showing that automatic generators can produce language with recognizable personality variation.

Alt	Realization	Extra
5	Err... it seems to me that Le Marais isn't as bad as the others.	1.83
4	Right, I mean, Le Marais is the only restaurant that is any good.	2.83
8	Ok, I mean, Le Marais is a quite french, kosher and steak house place, you know and the atmosphere isn't nasty, it has nice atmosphere. It has friendly service. It seems to me that the service is nice. It isn't as bad as the others, is it?	5.17
9	Well, it seems to me that I am sure you would like Le Marais. It has good food, the food is sort of rather tasty, the ambience is nice, the atmosphere isn't sort of nasty, it features rather friendly servers and its price is around 44 dollars.	5.83
3	I am sure you would like Le Marais, you know. The atmosphere is acceptable, the servers are nice and it's a french, kosher and steak house place. Actually, the food is good, even if its price is 44 dollars.	6.00
10	It seems to me that Le Marais isn't as bad as the others. It's a french, kosher and steak house place. It has friendly servers, you know but it's somewhat expensive, you know!	6.17
2	Basically, actually, I am sure you would like Le Marais. It features friendly service and acceptable atmosphere and it's a french, kosher and steak house place. Even if its price is 44 dollars, it just has really good food, nice food.	6.17

Table 1: Recommendations along the extraversion dimension, with the average extraversion rating from human judges on a scale from 1 to 7. Alt-2 and 3 are from the extravert set, Alt-4 and 5 are from the introvert set, and others were randomly generated.

Our aim is to produce a highly parameterizable generator whose outputs vary along personality dimensions. We hypothesize that such language can be generated by varying parameters suggested by psycholinguistic research. So, we must first map the psychological findings to parameters of a natural language generator (NLG). However, this presents several challenges: (1) The findings result from studies of genres of language, such as stream-of-consciousness essays (Pennebaker and King, 1999), and informal conversations (Mehl et al., 2006), and thus may not apply to fixed content domains used in NLG; (2) Most findings are based on self-reports of personality, but we want to affect observer's perceptions; (3) The findings consist of weak but significant correlations, so that individual parameters may not have a strong enough effect to produce recognizable variation within a single utterance; (4) There are many possible mappings of the findings to generation parameters; and (5) It is unclear whether only specific speech-act types manifest personality or whether all utterances do.

Thus this paper makes several contributions. First, Section 2 summarizes the linguistic reflexes of extraversion, organized by the modules in a standard NLG system, and propose a mapping from these findings to NLG parameters. To our knowledge this is the first attempt to put forward a systematic framework for generating language manifesting personality. We start with the extraversion dimension because it is an important personality factor, with many associated linguistic variables. We believe that our framework will generalize to the other dimensions in the Big Five model. Second, Sections 3 and 4 describe the PERSONAGE (PERSONAlity GEnerator) generator and its 29 parameters. Table 1 shows examples generated by PERSONAGE for recommendations in the restaurant domain, along with human extraversion judgments. Third, Sections 5 and 6 describe experiments evaluating two generation methods. We first show that (1) the parameters generate utterances that vary significantly on the extraversion dimension, according to human judgments; and (2) we can train a statistical model that matches human performance in assigning extraversion ratings to generation outputs produced with random parameter settings. Section 7 sums up and discusses future work.

2 Psycholinguistic Findings and PERSONAGE Parameters

We hypothesize that personality can be made manifest in evaluative speech acts in any dialogue domain, i.e. utterances responding to requests to RECOMMEND or COMPARE domain entities, such as restaurants or movies (Isard et al., 2006; Stent et al., 2004). Thus, we start with the SPaRKy generator[1], which produces evaluative recommendations and comparisons in the restaurant domain, for a database of restaurants in New York City. There are eight attributes for each restaurant: the name and address, scalar attributes for *price*, *food quality*, *atmosphere*, and *service* and categorical attributes for *neighborhood* and *type of cuisine*. SPaRKy is based on the standard NLG architecture (Reiter and Dale, 2000), and consists of the following modules:

1. Content Planning: refine communicative goals, select and structure content;
2. Sentence planning; choose linguistic resources (lexicon, syntax) to achieve goals;
3. Realization: use grammar (syntax, morphology) to generate surface utterances.

Given the NLG architecture, speech-act types, and domain, the first step then is to summarise psychological findings on extraversion and map them to this architecture. The column **NLG modules** of Table 2 gives the proposed mapping. The first row specifies findings for the content planning module and the other rows are aspects of sentence planning. Realization is achieved with the RealPro surface realizer (Lavoie and Rambow, 1997). An examination of the introvert and extravert findings in Table 2 highlights the challenges above, i.e. exploiting these findings in a systematic way within a parameterizable NLG system.

The column **Parameter** in Table 2 proposes parameters (explained in Sections 3 and 4) that are manipulated within each module to realize the findings in the other columns. Each parameter varies continuously from 0 to 1, where end points are meant to produce extreme but plausible output. Given the challenges above, it is important to note that these parameters represent *hypotheses* about how a finding can be mapped into *any* NLG system. The **Intro** and **Extra** columns at the right hand side of the **Parameter** column indicate a range of settings for this parameter, suggested by the psychological findings, to produce introverted vs. extraverted language.

SPaRKy produces content plans for restaurant recommendations and comparisons that are modified by the parameters. The sample content plan for a recommendation in Figure 1 corresponds to the outputs in Table 1. While Table 1 shows that PERSONAGE's parameters have various pragmatic effects, they preserve the meaning at the Gricean intention level (dialogue goal). Each content plan contains a claim (nucleus) about the overall quality of

[1] Available for download from
www.dcs.shef.ac.uk/cogsys/sparky.html

NLG modules	Introvert findings	Extravert findings	Parameter	Intro	Extra
Content selection and structure	Single topic	Many topics	VERBOSITY	low	high
	Strict selection	Think out loud*	RESTATEMENTS	low	high
			REPETITIONS	low	low
	Problem talk, dissatisfaction	Pleasure talk, agreement, compliment	CONTENT POLARITY	low	high
			REPETITIONS POLARITY	low	high
			CLAIM POLARITY	low	high
			CONCESSIONS	avg	avg
			CONCESSIONS POLARITY	low	high
			POLARISATION	low	high
			POSITIVE CONTENT FIRST	low	high
Syntactic templates selection	Few self-references	Many self-references	SELF-REFERENCES	low	high
	Elaborated constructions	Simple constructions*	CLAIM COMPLEXITY	high	low
	Many articles	Few articles			
Aggregation Operations	Many words per sentence/clause	Few words per sentence/clause	RELATIVE CLAUSES	high	low
			WITH CUE WORD	high	low
			CONJUNCTION	low	high
	Many unfilled pauses	Few unfilled pauses	PERIOD	high	low
			...		
Pragmatic transformations	Many nouns, adjectives, prepositions (explicit)	Many verbs, adverbs, pronouns (implicit)	SUBJECT IMPLICITNESS	low	high
	Many negations	Few negations	NEGATION INSERTION	high	low
	Many tentative words	Few tentative words	DOWNTONER HEDGES:		
			·SORT OF, SOMEWHAT, QUITE, RATHER, ERR, I THINK THAT, IT SEEMS THAT, IT SEEMS TO ME THAT, I MEAN	high	low
			·AROUND	avg	avg
	Formal	Informal	·KIND OF, LIKE	low	high
			ACKNOWLEDGMENTS:		
			·YEAH	low	high
			·RIGHT, OK, I SEE, WELL	high	low
	Realism	Exaggeration*	EMPHASIZER HEDGES:		
			·REALLY, BASICALLY, ACTUALLY, JUST HAVE, JUST IS, EXCLAMATION	low	high
			·YOU KNOW	low	high
	No politeness form	Positive face redressment*	TAG QUESTION INSERTION	low	high
	Lower word count	Higher word count	HEDGE VARIATION	low	avg
			HEDGE REPETITION	low	low
Lexical choice	Rich	Poor	LEXICON FREQUENCY	low	high
	Few positive emotion words	Many positive emotion words	*see polarity parameters*		
	Many negative emotion words	Few negative emotion words	*see polarity parameters*		

Table 2: Summary of language cues for extraversion, based on Dewaele and Furnham (1999); Furnham (1990); Mehl et al. (2006); Oberlander and Gill (2006); Pennebaker and King (1999), as well as PERSONAGE's corresponding generation parameters. Asterisks indicate hypotheses, rather than results. For details on aggregation parameters, see Section 4.2.

Relations:	JUSTIFY (nuc:1, sat:2); JUSTIFY (nuc:1, sat:3); JUSTIFY (nuc:1, sat:4); JUSTIFY (nuc:1, sat:5); JUSTIFY (nuc:1, sat:6)
Content:	1. assert(best (*Le Marais*))
	2. assert(is (*Le Marais*, cuisine (*French*)))
	3. assert(has (*Le Marais*, food-quality (*good*)))
	4. assert(has (*Le Marais*, service (*good*)))
	5. assert(has (*Le Marais*, decor (*decent*)))
	6. assert(is (*Le Marais*, price (*44 dollars*)))

Figure 1: A content plan for a recommendation.

the selected restaurant(s), supported by a set of satellite content items describing their attributes. See Table 1. Claims can be expressed in different ways, such as *RESTAURANT_NAME is the best*, while the attribute satellites follow the pattern *RESTAURANT_NAME has MODIFIER ATTRIBUTE_NAME*, as in *Le Marais has good food*. Recommendations are characterized by a JUSTIFY rhetorical relation associating the claim with all other content items, which are linked together through an INFER relation. In comparisons, the attributes of multiple restaurants are compared using a CONTRAST relation. An optional claim about the quality of all restaurants can also be expressed as the nucleus of an ELABORATE relation, with the rest of the content plan tree as a satellite.

3 Content Planning

Content planning selects and structures the content to be communicated. Table 2 specifies 10 parameters hypothesized to affect this process which are explained below.

Content size: Extraverts are more talkative than introverts (Furnham, 1990; Pennebaker and King, 1999), although it is not clear whether they actually produce more content, or are just redundant and wordy. Thus various parameters relate to the amount and type of content produced. The VERBOSITY parameter controls the number of content items selected from the content plan. For example, Alt-5 in Table 1 is terse, while Alt-2 expresses all the items in the content plan. The REPETITION parameter adds an exact repetition: the content item is duplicated and linked to the original content by a RESTATE rhetorical relation. In a similar way, the RESTATEMENT parameter adds paraphrases of content items to the plan, that are obtained from the initial handcrafted generation dictionary (see Section 4.1) and by automatically substituting content words with the most frequent WordNet synonym (see Section 4.4). Alt-9 in Table 1 contains restatements for the food quality and the atmosphere attributes.

Polarity: Extraverts tend to be more positive; introverts are characterized as engaging in more 'problem talk' and expressions of dissatisfaction (Thorne, 1987). To control for polarity, content items are defined as positive or negative based on the scalar value of the corresponding attribute. The *type of cuisine* and *neighborhood* attributes have neutral polarity. There are multiple parameters associated with polarity. The CONTENT POLARITY parameter controls whether the content is mostly negative (e.g. *X has mediocre food*), neutral (e.g. *X is a Thai restaurant*), or positive. From the filtered set of content items, the POLARISATION parameter determines whether the final content includes items with extreme scalar values (e.g. *X has fantastic staff*).

In addition, polarity can also be implied more subtly through rhetorical structure. The CONCESSIONS parameter controls how negative and positive information is presented, i.e. whether two content items with different polarity are presented objectively, or if one is foregrounded and the other backgrounded. If two opposed content items are selected for a concession, a CONCESS rhetorical relation is inserted between them. While the CONCESSIONS parameter captures the tendency to put information into perspective, the CONCESSION POLARITY parameter controls whether the positive or the negative content is concessed, i.e. marked as the satellite of the CONCESS relation. The last sentence of Alt-3 in Table 1 illustrates a positive concession, in which the good food quality is put before the high price.

Content ordering: Although extraverts use more positive language (Pennebaker and King, 1999; Thorne, 1987), it is unclear how they position the positive content within their utterances. Additionally, the position of the claim affects the persuasiveness of an argument (Carenini and Moore, 2000): starting with the claim facilitates the hearer's understanding, while finishing with the claim is more effective if the hearer disagrees. The POSITIVE CONTENT FIRST parameter therefore controls whether positive content items – including the claim – appear first or last, and the order in which the content items are aggregated. However, some operations can still impose a specific ordering (e.g. BECAUSE cue word to realize the JUSTIFY relation, see Section 4.2).

4 Sentence Planning

Sentence planning chooses the linguistic resources from the lexicon and the syntactic and discourse structures to achieve the communicative goals specified in the input content plan. Table 2 specifies four sets of findings and parameters for different aspects of sentence planning discussed below.

4.1 Syntactic template selection

PERSONAGE's input generation dictionary is made of 27 Deep Syntactic Structures (DSyntS): 9 for the recommendation claim, 12 for the comparison claim, and one per attribute. Selecting a DSyntS requires assigning it automatically to a point in a three dimensional space described below. All parameter values are normalized over all the DSyntS, so the DSyntS closest to the target value can be computed.

Syntactic complexity: Furnham (1990) suggests that introverts produce more complex constructions: the CLAIM COMPLEXITY parameter controls the depth of the syntactic structure chosen to represent the claim, e.g. the claim *X is the best* is rated as less complex than *X is one of my favorite restaurants*.

Self-references: Extraverts make more self-references than introverts (Pennebaker and King, 1999). The SELF-REFERENCE parameter controls whether the claim is made in the first person, based on the speaker's own experience, or whether the claim is reported as objective or information obtained elsewhere. The self-reference value is obtained from the syntactic structure by counting the number of first person pronouns. For example, the claim of Alt-2 in Table 1, i.e. *I am sure you would like Le Marais*, will be rated higher than *Le Marais isn't as bad as the others* in Alt-5.

Polarity: While polarity can be expressed by content selection and structure, it can also be directly associated with the DSyntS. The CLAIM POLARITY parameter determines the DSyntS selected to realize the claim. DSyntS are manually annotated for polarity. For example, Alt-4's claim in Table 1, i.e. *Le Marais is the only restaurant that is any good*, has a lower polarity than Alt-2.

4.2 Aggregation operations

SPaRKy aggregation operations are used (See Stent et al. (2004)), with additional operations for concessions and restatements. See Table 2. The probability of the operations biases the production of complex clauses, periods and formal cue words for introverts, to express their preference for complex syntactic constructions, long pauses and rich vocabulary (Furnham, 1990). Thus, the introvert parameters favor operations such as RELATIVE CLAUSE for the INFER relation, PERIOD HOWEVER CUE WORD for CONTRAST, and ALTHOUGH ADVERBIAL CLAUSE for CONCESS, that we hypothesize to result in more formal language. Extravert aggregation produces longer sentences with simpler constructions and informal cue words. Thus extravert utterances tend to use operations such as a CONJUNCTION to realize the INFER and RESTATE relations, and the EVEN IF ADVERBIAL CLAUSE for CONCESS relations.

4.3 Pragmatic transformations

This section describes the insertion of markers in the DSyntS to produce various pragmatic effects.

Hedges: Hedges correlate with introversion (Pennebaker and King, 1999) and affect politeness (Brown and Levinson, 1987). Thus there are parameters for inserting a wide range of hedges, both affective and epistemic, such as *kind of*, *sort of*, *quite*, *rather*, *somewhat*, *like*, *around*, *err*, *I think that*, *it seems that*, *it seems to me that*, and *I mean*. Alt-5 in Table 1 shows hedges *err* and *it seems to me that*.

To model extraverts use of more social language, agreement and backchannel behavior (Dewaele and Furnham, 1999; Pennebaker and King, 1999), we use informal acknowledgments such as *yeah*, *right*, *ok*. Acknowledgments that may affect introversion are *I see*, expressing self-reference and cognitive load, and the *well* cue word implying reservation from the speaker (see Alt-9).

To model social connection and emotion we added mechanisms for inserting emphasizers such as *you know, basically, actually, just have, just is*, and exclamations. Alt-3 in Table 1 shows the insertion of *you know* and *actually*.

Although similar hedges can be grouped together, each hedge has a unique pragmatic effect. For example, *you know* implies positive-face redressment, while *actually* doesn't. A parameter for each hedge controls the likelihood of its selection.

To control the general level of hedging, a HEDGE VARIATION parameter defines how many different hedges are selected (maximum of 5), while the frequency of an individual hedge is controlled by a HEDGE REPETITION parameter, up to a maximum of 2 identical hedges per utterance.

The syntactic structure of hedges are defined as well as constraints on their insertion point in the utterance's syntactic structure. Each time a hedge is selected, it is randomly inserted at one of the insertion points respecting the constraints, until the specified frequency is reached. For example, a constraint on the hedge *kind of* is that it modifies adjectives.

Tag questions: Tag questions are also politeness markers (Brown and Levinson, 1987). They redress the hearer's positive face by claiming common ground. A TAG QUESTION INSERTION parameter leads to negating the auxiliary of the verb and pronominalizing the subject, e.g. *X has great food* results in the insertion of *doesn't it?*, as in Alt-8.

Negations: Introverts use significantly more negations (Pennebaker and King, 1999). Although the content parameters select more negative polarity content items for introvert utterances, we also manipulate negations, while keeping the content constant, by converting adjectives to the negative of their antonyms, e.g. *the atmosphere is nice* was transformed to *not nasty* in Alt-9 in Table 1.

Subject implicitness: Heylighen and Dewaele (2002) found that extraverts use more implicit language than introverts. To control the level of implicitness, the SUBJECT IMPLICITNESS parameter determines whether predicates describing restaurant attributes are expressed with the restaurant in the subject, or with the attribute itself (e.g., *it has good food* vs. *the food is tasty* in Alt-9).

4.4 Lexical choice

Introverts use a richer vocabulary (Dewaele and Furnham, 1999), so the LEXICON FREQUENCY parameter selects lexical items by their normalized frequency in the British National Corpus. WordNet synonyms are used to obtain a pool of synonyms, as well as adjectives extracted from a corpus of restaurant reviews for all levels of polarity (e.g. the adjective *tasty* in Alt-9 is a high polarity modifier of the food attribute). Synonyms are manually checked to make sure they are interchangeable. For example, the content item expressed originally as *it has decent service* is transformed to *it features friendly service* in Alt-2, and to *the servers are nice* in Alt-3.

5 Experimental Method and Hypotheses

Our primary hypothesis is that language generated by varying parameters suggested by psycholinguistic research can be recognized as extravert or introvert. To test this hypothesis, three expert judges evaluated a set of generated utterances as if they had been uttered by a friend responding in a dialogue to a request to recommend restaurants. These utterances had been generated to systematically manipulate extraversion/introversion parameters.

The judges rated each utterance for perceived extraversion, by answering the two questions measuring that trait from the Ten-Item Personality Inventory, as this instrument was shown to be psychometrically superior to a 'single item per trait' questionnaire (Gosling et al., 2003). The answers are averaged to produce an extraversion rating ranging from 1 (highly introvert) to 7 (highly extravert). Because it was unclear whether the generation parameters in Table 2 would produce natural sounding utterances, the judges also evaluated the naturalness of each utterance on the same scale. The judges rated 240 utterances, grouped into 20 sets of 12 utterances generated from the same content plan. They rated one randomly ordered set at a time, but viewed all 12 utterances in that set before rating them. The utterances were generated to meet two experimental goals. First, to test the direct control of the perception of extraversion. 2 introvert utterances and 2 extravert utterances were generated for each content plan (80 in total) using the parameter values in Table 2. Multiple outputs were generated with both parameter settings normally distributed with a 15% standard deviation. Second, 8 utterances for each content plan (160 in total) were generated with random parameter values. These random utterances make it possible to: (1) improve PERSONAGE's direct output by calibrating its parameters more precisely; and (2) build a statistical model that selects utterances matching input personality values after an overgeneration phase (see Section 6.2). The interrater agreement for extraversion between the judges over all 240 utterances (average Pearson's correlation of 0.57) shows that the magnitude of the differences of perception between judges is almost constant ($\sigma = .037$). A low agreement can yield a high correlation (e.g. if all values differ by a constant factor), so we also compute the intraclass correlation coefficient r based on a two-way random effect model. We obtain a r of 0.79, which is significant at the $p < .001$ level (reliability of average measures, identical to Cronbach's alpha). This is comparable to the agreement of judgments of personality in Mehl et al. (2006) (mean $r = 0.84$).

6 Experimental Results

6.1 Hypothesized parameter settings

Table 1 provides examples of PERSONAGE's output and extraversion ratings. To assess whether PERSONAGE generates language that can be recognized as introvert and extravert, we did a independent sample t-test between the average ratings of the 40 introvert and 40 extravert utterances (parameters with 15% standard deviation as in Table 2). Table 3 shows that introvert utterances have an average rating of 2.96 out of 7 while extravert utterances have an average rating of 5.98. These ratings are significantly different at the $p < .001$ level (two-tailed). In addition, if we divide the data into two equal-width bins around the neutral extravert rating (4 out of 7), then PERSONAGE's utterance ratings fall in the bin predicted by the parameter set 89.2% of the time. Extravert utterance are also slightly more natural than the introvert ones ($p < .001$).

Rating	**Introvert**	**Extravert**	**Random**
Extraversion	2.96	5.98	5.02
Naturalness	4.93	5.78	4.51

Table 3: Average extraversion and naturalness ratings for the utterances generated with introvert, extravert, and random parameters.

Table 3 also shows that the 160 random parameter utterances produce an average extraversion rating of 5.02, both significantly higher than the introvert set and lower than the extravert set ($p < .001$). Interestingly, the random utterances, which may combine linguistic variables associated with both introverts and extraverts, are less natural than the introvert ($p = .059$) and extravert sets ($p < .001$).

6.2 Statistical models evaluation

We also investigate a second approach: overgeneration with random parameter settings, followed by ranking via a statistical model trained on the judges' feedback. This approach supports generating utterances for any input extraversion value, as well as determining which parameters affect the judges' perception.

We model perceived personality ratings $(1 \ldots 7)$ with regression models from the Weka toolbox (Witten and Frank, 2005). We used the full dataset of 160 averaged ratings for the random parameter utterances. Each utterance was associated with a feature vector with the generation decisions for each parameter in Section 2. To reduce data sparsity, we select features that correlate significantly with the ratings ($p < .10$) with a coefficient higher than 0.1.

Regression models are evaluated using the mean absolute error and the correlation between the predicted score and the actual average rating. Table 4 shows the mean absolute error on a scale from 1 to 7 over ten 10-fold cross-validations for the 4 best regression models: Linear Regression (LR), M5' model tree (M5), and Support Vector Machines (i.e. SMOreg) with linear kernels (SMO_1) and radial-basis function kernels (SMO$_r$). All models significantly outperform the baseline (0.83 mean absolute error, $p < .05$), but surprisingly the linear model performs the best with a mean absolute error of 0.65. The best model produces a correlation coefficient of 0.59 with the judges' ratings, which is higher than the correlations between pairs of judges, suggesting that the model performs as well as a human judge.

Metric	**LR**	**M5**	**SMO$_1$**	**SMO$_r$**
Absolute error	**0.65**	0.66	0.72	0.70
Correlation	**0.59**	0.56	0.54	0.57

Table 4: Mean absolute regression errors (scale from 1 to 7) and correlation coefficients over ten 10-fold cross-validations, for 4 models: Linear Regression (LR), M5' model tree (M5), Support Vector Machines with linear kernels (SMO$_1$) and radial-basis function kernels (SMO$_r$). All models significantly outperform the mean baseline (0.83 error, $p < .05$).

The M5' regression tree in Figure 2 assigns a rating given the features. Verbosity plays the most important role: utterances with 4 or more content items are modeled as more extravert. Given a low verbosity, lexical frequency and restatements determine the extraversion level, e.g. utterances with less than 4 content items and infrequent words are perceived as very introverted (rating of 2.69 out of 7). For verbose utterances, the *you know* hedge indicates extraversion, as well as concessions, restatements, self-references, and positive content. Although relatively simple, these models are useful for identifying new personality markers, as well as calibrating parameters in the direct generation model.

7 Discussion and Conclusions

We present and evaluate PERSONAGE, a parameterizable generator that produces outputs that vary along the extraversion personality dimension. This paper makes four contributions:

1. We present a systematic review of psycholinguistic findings, organized by the NLG reference architecture;
2. We propose a mapping from these findings to generation parameters for each NLG module and a real-time implementation of a generator using these parameters[2]. To our knowledge this is the first attempt to put forward a systematic framework for generating language that manifests personality;
3. We present an evaluation experiment showing that we can control the parameters to produce recognizable linguistic variation along the extraversion personality dimension. Thus, we show that the weak correlations reported in other genres of language, and for self-reports rather than observers, carry over to the production of single evaluative utterances with recognizable personality in a restricted domain;
4. We present the results of a training experiment showing that given an output, we can train a model that matches human performance in assigning an extraversion rating to that output.

Some of the challenges discussed in the introduction remain. We have shown that evaluative utterances in the restaurant domain can manifest personality, but more research is needed on which speech acts recognisably manifest personality in a restricted domain. We also showed that the mapping we hypothesised of findings to generation parameters was effective, but there may be additional parameters that the psycholinguistic findings could be mapped to.

Our work was partially inspired by the ICONOCLAST and PAULINE parameterizable generators (Bouayad-Agha et al., 2000; Hovy, 1988), which vary the style, rather than the personality, of the generated texts. Walker et al. (1997) describe a generator intended to affect perceptions of personality, based on Brown and Levinson's theory of politeness (Brown and Levinson, 1987), that uses some of the linguistic constructions implemented here, such as tag questions and hedges, but it was never evaluated. Research by André et al. (2000); Piwek (2003) uses personality variables to affect the linguistic behaviour of conversational agents, but they did not systematically manipulate parameters, and their generators were not evaluated. Reeves and Nass (1996) demonstrate that manipulations of personality affect many aspects of user's perceptions, but their experiments use handcrafted utterances, rather than generated utterances. Cassell and Bickmore (2003) show that extraverts prefer systems utilizing discourse plans that include small talk. Paiva and Evans' trainable generator (2005) produces outputs that correspond to a set of linguistic variables measured in a corpus of target texts. Their method is similar to our statistical method using regression trees, but provides direct control. The method reported in Mairesse and Walker (2005) for training individualized sentence planners ranks the outputs produced by an overgeneration phase, rather than directly predicting a scalar value, as we do here. The closest work to ours is probably Isard et al.'s CRAG-2 system (2006), which overgenerates and ranks using ngram language models trained on a corpus labelled for all Big Five personality dimensions. However, CRAG-2 has no explicit parameter control, and it has yet to be evaluated.

[2] An online demo is available at www.dcs.shef.ac.uk/cogsys/personage.html

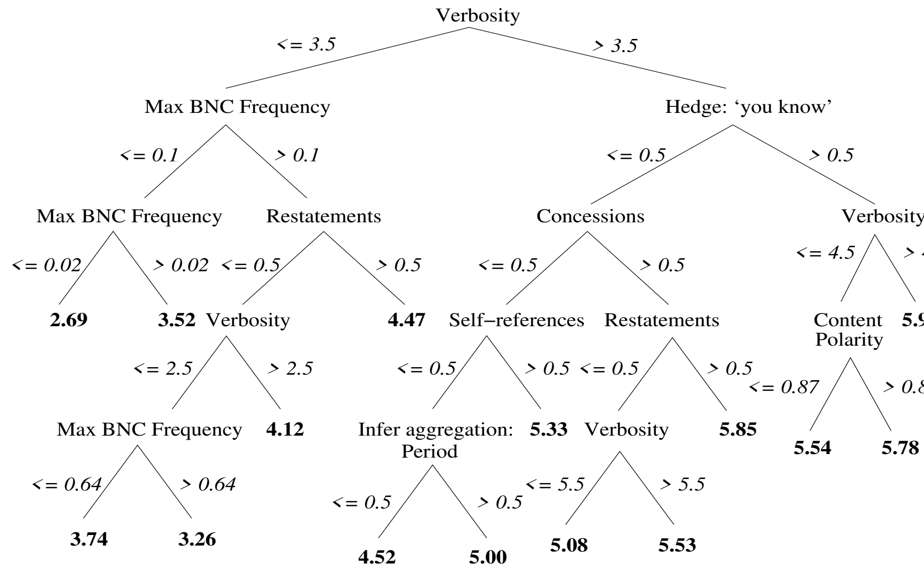

Figure 2: M5' regression tree. The output ranges from 1 to 7, where 7 means strongly extravert.

In future work, we hope to directly compare the direct generation method of Section 6.1 with the overgenerate and rank method of Section 6.2, and to use these results to refine PERSONAGE's parameter settings. We also hope to extend PERSONAGE's generation capabilities to other Big Five traits, identify additional features to improve the model's performance, and evaluate the effect of personality variation on user satisfaction in various applications.

References

E. André, T. Rist, S. van Mulken, M. Klesen, and S. Baldes. 2000. The automated design of believable dialogues for animated presentation teams. In *Embodied conversational agents*, p. 220–255. MIT Press, Cambridge, MA.

N. Bouayad-Agha, D. Scott, and R. Power. 2000. Integrating content and style in documents: a case study of patient information leaflets. *Information Design Journal*, 9:161–176.

P. Brown and S. Levinson. 1987. *Politeness: Some universals in language usage*. Cambridge University Press.

G. Carenini and J. D. Moore. 2000. A strategy for generating evaluative arguments. In *Proc. of International Conference on Natural Language Generation*, p. 47–54.

J. Cassell and T. Bickmore. 2003. Negotiated collusion: Modeling social language and its relationship effects in intelligent agents. *User Modeling and User-Adapted Interaction*, 13 (1-2):89–132.

J-M. Dewaele and A. Furnham. 1999. Extraversion: the unloved variable in applied linguistic research. *Language Learning*, 49(3):509–544.

A. Furnham. 1990. Language and personality. In *Handbook of Language and Social Psychology*. Winley.

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

F. Heylighen and J-M. Dewaele. 2002. Variation in the contextuality of language: an empirical measure. *Context in Context, Foundations of Science*, 7(3):293–340.

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

A. Isard, C. Brockmann, and J. Oberlander. 2006. Individuality and alignment in generated dialogues. In *Proc. of INLG*.

B. Lavoie and O. Rambow. 1997. A fast and portable realizer for text generation systems. In *Proc. of ANLP*.

A. Loyall and J. Bates. 1997. Personality-rich believable agents that use language. In *Proc. of the First International Conference on Autonomous Agents*, p. 106–113.

F. Mairesse and M. Walker. 2005. Learning to personalize spoken generation for dialogue systems. In *Proc. of the Interspeech - Eurospeech*, p. 1881–1884.

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

W. T. Norman. 1963. Toward an adequate taxonomy of personality attributes: Replicated factor structure in peer nomination personality rating. *Journal of Abnormal and Social Psychology*, 66:574–583.

J. Oberlander and A. Gill. 2006. Language with character: A stratified corpus comparison of individual differences in e-mail communication. *Discourse Processes*, 42:239–270.

D. Paiva and R. Evans. 2005. Empirically-based control of natural language generation. In *Proc. of ACL*.

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

P. Piwek. 2003. A flexible pragmatics-driven language generator for animated agents. In *Proc. of EACL*.

B. Reeves and C. Nass. 1996. *The Media Equation*. University of Chicago Press.

E. Reiter and R. Dale. 2000. *Building Natural Language Generation Systems*. Cambridge University Press.

A. Stent, R. Prasad, and M. Walker. 2004. Trainable sentence planning for complex information presentation in spoken dialog systems. In *Proc. of ACL*.

A. Thorne. 1987. The press of personality: A study of conversations between introverts and extraverts. *Journal of Personality and Social Psychology*, 53:718–726.

M. Walker, J. Cahn, and S. Whittaker. 1997. Improvising linguistic style: Social and affective bases for agent personality. In *Proc. of the Conference on Autonomous Agents*.

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