Controlling User Perceptions of Linguistic Style: Trainable Generation of Personality Traits

François Mairesse *
University of Cambridge

Marilyn A. Walker **
University of California, Santa Cruz

Recent work in natural language generation has begun to take linguistic variation into account, developing algorithms that are capable of modifying the system's linguistic style based either on the user's linguistic style or other factors, such as personality or politeness. While stylistic control has traditionally relied on handcrafted rules, statistical methods are likely to be needed for generation systems to scale to the production of the large range of variation observed in human dialogues. Previous work on statistical natural language generation (SNLG) has shown that the grammaticality and naturalness of generated utterances can be optimized from data, however these data-driven methods have not been shown to produce stylistic variation that is perceived by humans in the way that the system intended. This paper describes PERSONAGE, *a highly parameterizable language generator whose parameters are based on psychological findings about the linguistic reflexes of personality. We present a novel SNLG method which uses parameter estimation models trained on personality-annotated data to predict the generation decisions required to convey any combination of scalar values along the five main dimensions of personality. A human evaluation shows that parameter estimation models produce recognizable stylistic variation along multiple dimensions, on a continuous scale, and without the computational cost incurred by overgeneration techniques.*

1. Introduction

While language can be seen as simply a method for exchanging information, it also has a social function (Goffman 1970; Labov 2006; Dunbar 1996). Speakers use linguistic cues to project social aspects of utterances, such as the speaker's personality, emotions, and social group, and hearers use these cues to infer properties about the speaker. While some cues appear to be produced through automatic cognitive processes (Levelt and Kelter 1982; Pickering and Garrod 2004), speakers may also *overload* their communicative intentions to try to satisfy multiple goals simultaneously (Jordan 2000; Pollack 1991; Stone and Webber 1998), such as projecting a specific image to the hearer while communicating information and minimizing communicative effort (Clark and Brennan

* This work was done at University of Sheffield. The author's present address is: Cambridge University Engineering Department, Trumpington street, Cambridge CB2 1PZ, United Kingdom, E-mail: f.mairesse@eng.cam.ac.uk.
** Baskin School of Engineering, 1156 High Street, SOE-3, Santa Cruz, CA 95064, E-mail: maw@soe.ucsc.edu.com

Submission received: 20 January 2009; revised submission received: 18 October 2010; accepted for publication: 30 November 2010.

1991; Brennan and Clark 1996). The combination of these pragmatic effects results in the large range of *linguistic variation* observed between individual speakers (Biber 1988).

Much of the research on generating utterances that manifest different linguistic styles has focused on text generation applications such as journalistic writing or instruction manuals (Hovy 1988; Green and DiMarco 1996; Paris and Scott 1994; Scott and de Souza 1990; Power, Scott, and Bouayad-Agha 2003; Bouayad-Agha, Scott, and Power 2000; Inkpen and Hirst 2004). Recent research in language generation for dialogue applications has also begun to take linguistic variation into account, developing algorithms to modify the system's linguistic style based on either the user's linguistic style, or other factors, such as the user's emotional state, her personality, or considerations of politeness strategies (Walker, Cahn, and Whittaker 1997; André et al. 2000; Lester, Towns, and Fitzgerald 1999; Lester, Stone, and Stelling 1999; Cassell and Bickmore 2003; Piwek 2003). There is growing evidence that dialogue systems, such as intelligent tutoring systems, are more effective if they can generate a range of different types of stylistic linguistic variation (McQuiggan, Mott, and Lester 2008; Porayska-Pomsta and Mellish 2004; Litman and Forbes-Riley 2004, 2006; Tapus and Mataric 2008; Wang et al. 2005). Most of this work uses either templates or handcrafted rules to generate utterances. This guarantees high quality, natural outputs, which is very useful for demonstrating the utility of stylistic variation.

However, handcrafted approaches mean that utterances have to be constructed by hand for each new application, leading to problems of portability and scalability (Rambow, Rogati, and Walker 2001). Statistical natural language generation (SNLG) has the potential to address such scalability issues, by relying on annotated data rather than manual parameter tuning. It also offers the promise of techniques for producing continuous stylistic variation over multiple stylistic factors, by automatically learning a model of the relation between stylistic factors and properties (parameters) of generated utterances (Paiva and Evans 2004, 2005). It is difficult to produce such continuous variation over multiple factors with a rule-based or template-based approach (but see (Bouayad-Agha, Scott, and Power 2000)). Moreover, to date, no-one has shown that humans correctly perceive the generated variation as the system intended, nor has anyone shown that an SNLG approach can produce outputs that are natural enough to be used in dialogue applications such as intelligent tutoring systems, interactive drama systems, and conversational agents, where some types of stylistic variation have already been shown to be useful.

In previous work, we argue that the Big Five model of personality provides a useful framework for modeling some types of stylistic linguistic variation. This model of human personality has become widely accepted in psychology over the last 50 years (Funder 1997). Table 1 tabulates each Big Five trait along with some of the important trait adjectives associated with the extremes of each trait. We believe that these trait adjectives provide an intuitive, meaningful definition of linguistic style. In previous work we describe a rule-based version of PERSONAGE, which here we will refer to as PERSONAGE-RB (Mairesse and Walker 2007; Mairesse 2008). In PERSONAGE-RB, generation parameters are implemented, and their values are set, based on correlations between linguistic cues and the Big Five traits that have been systematically documented in the psychology literature (Pennebaker and King 1999; Mehl, Gosling, and Pennebaker 2006; Scherer 1979; Furnham 1990). For example, parameters for the extraversion trait include verbosity, sentence length, and the production of positive content. We showed experimentally that humans perceive utterances generated by PERSONAGE-RB as conveying the extremes of all Big Five traits (e.g., neuroticism (**low**) vs. emotionally stable

Table 1
Example adjectives associated with the extremes of all Big Five traits.

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

(**high**), see Table 1). Our evaluation uses a validated perceptual questionnaire from the personality psychology literature (Gosling, Rentfrow, and Swann 2003).

However, while PERSONAGE-RB only generates 10 discrete personalities emphasizing either the high or the low end of one trait, psychologists measure personality traits on *continuous* scales (Norman 1963; Goldberg 1990; Marcus et al. 2006), and human language simultaneously manifests multiple personality traits. Some computational applications may require more than a small set of personality types, which suggests that systems adapting their linguistic style to the user would benefit from fine-grained personality models. We believe that the only way to robustly and efficiently learn such fine-grained variation is to model personality as a continuous variable, rather than using arbitrary discrete personality classes. Personality generation models should thus learn to map continuous target personality scores to discrete utterances. In order to achieve this, the handcrafted rule-based approach would require the manual examination of psycholinguistic findings, followed by testing in the application domain, to determine the appropriate range for each parameter value. Thus extending this approach to continuous variation that can project multiple traits simultaneously does not appear to be tractable.

The objective of this paper is to present and evaluate a language generator that is trained with a novel method, and which learns to generate stylistic variation expressing *multiple continuous stylistic dimensions* (in this case multiple personality traits). Before presenting our method, let us review existing paradigms for statistical language generation.

1.1 Previous Statistical Language Generation Methods

Previous work on SNLG has focused on three main approaches: (a) learning statistical language models (SLMs) from corpora in order to rerank a set of pre-generated utterances; (b) learning utterance reranking models from user feedback rather than corpora; and (c) learning generation parameters directly from data.

The first approach of prior work has used SLMs to rerank a large set of candidate utterances, and focused on grammaticality and naturalness (Langkilde-Geary 2002;

Bangalore and Rambow 2000; Chambers and Allen 2004; Nakatsu and White 2006). The seminal work of Langkilde and Knight (1998) in this area showed 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. The SLM scoring gives a low score (rank) to any ungrammatical output produced by the rule-based generator. We will refer to this as the **overgenerate and scoring (OS)** approach.

In a novel twist, Isard, Brockmann, and Oberlander (2006) applied this method to the generation of dialogues in which conversational agents with different personalities discuss movies. The SLM ranking model blends SLM's from blogs annotated with Big Five personality traits with SLMs from Switchboard, a much larger conversational dialogue corpus. Their CRAG-2 generator discretizes the blog personality ratings into three groups (low, medium and high), and models personality with three distinct SLM models for each trait. Each model estimates the likelihood of the utterance given the personality type. A cache model based on recently used linguistic forms can also be combined, in order to model recency effects and alignment (Pickering and Garrod 2004). This approach was integrated into a demonstrator, but it does not generate continuous variation (discretization of personality ratings), and to our knowledge it has never been evaluated to test whether the variation produced is perceivable by users.

A second approach to SNLG is a variant of the OS technique that trains the scoring phase to replicate human judgments rather than relying on the probabilities or frequencies of a SLM. This approach typically uses higher-level syntactic, semantic and discourse features rather than only n-grams, with typical results demonstrating that the performance of the scoring models approaches the gold-standard human ranking with a relatively small training set (Rambow, Rogati, and Walker 2001; Stent and Guo 2005; Nakatsu and White 2006). An advantage of this approach is that human judgments can be based on any aspect of the output, such as stylistic differences in politeness or personality. Walker et al. (2007) showed that this technique can be used to model individual preferences in rhetorical structure, syntactic form, and content ordering.

In previous work, we also applied this method to scoring randomly produced outputs of PERSONAGE (Mairesse 2008). The resulting statistical generator is referred to as PERSONAGE-OS. We randomly varied PERSONAGE's non-deterministic decisions points to generate a large number of paraphrases. We then computed post-hoc features consisting of the actual generation decisions, surface word n-grams, and content-analysis features from the Linguistic Inquiry and Word Count (LIWC) tool (Pennebaker, Francis, and Booth 2001) and the MRC psycholinguistic database (Coltheart 1981). Example content-analysis features include the ratio of words related to positive emotions (e.g., *good*), social interactions (e.g, *pal*), or the average frequency of use of each word. Scoring models trained on personality ratings of random utterances (in-domain data) outperformed the mean value baseline for all Big Five traits, with the best results for agreeableness, extraversion, and emotional stability. The models for those traits predict the ratings of unseen utterances with correlations of $r = .52$, $r = .37$, and $r = .29$ respectively. We also trained models on out-of-domain data, i.e. 96 personality-annotated conversation extracts (without any generation decision features). Results show that the out-of-domain models perform worse for all traits, only outperforming the baseline for agreeableness and conscientiousness. We also explored several *hybrid* methods for training that mix and blend data from different sources. Inspired by recent work on domain adaptation, we tested whether the performance of the out-of-domain models can be improved when training includes a small amount of data from the *target* domain, by applying the method of (Daumé III 2007). While adding out-of-domain data improved

performance for some traits, we find that adding a single domain feature performs as well as Daume's method. The results showed that mixing randomly generated in-domain utterances with rule-based in-domain utterances improves performance; the rule-based utterances provide a way to incorporate knowledge from the personality psychology literature into an SNLG approach. Thus, personality scoring models can be effective, however the computational cost of the OS approach remains a major drawback.

The third SNLG approach estimates the generation parameters *directly* from data, without any overgeneration phase. If the language generator is constrained to be a generative SLM, the parameters can then be learned through standard maximum-likelihood estimation. While n-gram SLMs can only model local linguistic phenomena, Belz showed that a context-free grammar (PCFG) can successfully model individual differences in the production of weather reports (Belz 2005, 2008). This method provides a principled way to produce utterances matching the linguistic style of a specific corpus — e.g., of an individual author — without any overgeneration phase. However, standard PCFG generation methods require a treebank-annotated corpus, and they cannot model context-dependent generation decisions, such as the control of sentence length or the generation of referring expressions.

Paiva and Evans (2005) adopt a more general framework by learning a regression model mapping generation decisions to stylistic dimensions extracted from a corpus, independently of the language generation mechanism. Factors are identified by applying factor analysis to a corpus exhibiting stylistic variation, and expressed as a linear combination of linguistic features (Biber 1988). Textual outputs are generated with a rule-based generator in the target domain, that is allowed to randomly vary the generation parameters, while logging the parameter settings corresponding to each output. Then the same factors found in the original corpus are measured in the random outputs, and linear regression is applied to learn which generation parameters predict the factor measurements. The generation parameters can then be manipulated to hit multiple stylistic targets on a continuous scale (since factors are measured continuously), by searching for the parameter setting yielding the target stylistic scores according to the linear models. The generator of Paiva and Evans (2005), trained in this way, can reproduce intended factor levels across several factors, such as sentence length and type of referring expression, thus modeling the stylistic variation as measured in the original corpus. Again, it has not been shown that humans perceive the stylistic differences that this approach produces.

1.2 Parameter Estimation Models

In the previous sections, we referred to two existing methods for controlling the parameters of PERSONAGE to produce stylistic variation: PERSONAGE-RB uses handcrafted generation parameter values for every target style of interest, while PERSONAGE-OS uses a statistical rescoring model to rerank a set of randomly generated utterances. The following sections develop and evaluate PERSONAGE-PE, a trainable generator which uses a direct generation method inspired by Paiva and Evans' approach (2005), to produce the stylistic variation found in personality traits, without any overgeneration phase. While Paiva and Evans learn models predicting the target stylistic scores from the generation parameters, we train **parameter estimation models (PE)** to estimate the optimal generation parameters *given* target personality scores, which are then used by the base generator to produce the output utterance. As parameter estimation models learn the reverse relationship to Paiva and Evans' regression models, there is no need

to search for the optimal generation parameter values at generation time. We evaluate the PE approach using the PERSONAGE base generator, whose parameters, architecture and capabilities are described in Section 2. Our experimental method is described in Section 3, together with an analysis of the data required to train our models. Section 4 analyzes some of the learned models, and evaluates the quality of the generated outputs using human judges, to compare our approach with the handcrafted PERSONAGE-RB generator of our previous work. Finally, Section 5 discusses the implications of our results and suggests many areas of future work.

This paper makes several contributions. First, we present a novel method for training an SNLG system that can produce multiple stylistic dimensions simultaneously, over continuous dimensions, without overgeneration or search. In order to evaluate our approach, we present the first empirical results showing that humans correctly perceive the stylistic variations (of any kind based on any utterance dimensions) that a statistical language generator intended to produce. Our evaluation of PERSONAGE-RB is the only other result that we know of for non-statistical generators (Mairesse and Walker 2007). Our experiments show that PERSONAGE-PE produces utterances perceived by humans as portraying different personalities, while maintaining a reasonable naturalness level (4.0 on a scale of 1 to 7). We do not know of any other human evaluation of an SNLG system that produces stylistic variation. Additionally, we test a wide range of machine learning algorithms to determine the best model for each generation decision in Section 4.1. We are not aware of any other work on SNLG to test such a wide range of algorithms.

2. The PERSONAGE Base Generator

The architecture of the PERSONAGE base generator is shown in Figure 1; it is discussed in detail in (Mairesse 2008)and (Mairesse and Walker 2010), and is only briefly summarized here.

The PERSONAGE architecture (Figure 1) builds on a standard natural language generation (NLG) pipeline architecture as described in (Reiter and Dale 2000; Kittredge, Korelsky, and Rambow 1991; Walker and Rambow 2002). We assume that the inputs to the generator are (1) a high-level communicative goal; (2) a content pool that can be used to achieve that goal, and (3) a set of generation parameter values. In a dialogue system, the communicative goal is provided by the dialogue manager. Two types of communicative goals are currently supported by PERSONAGE: *recommendation* and *comparison* of restaurants. PERSONAGE's content pool is based on a database of restaurants in New York City, with associated scalar values representing evaluative ratings for six attributes: *food quality*, *service*, *cuisine*, *location*, *price* and *atmosphere*.

The first component of the architecture shown in Figure 1 is the **content planner** which specifies the structure of the information to be conveyed. The resulting content plan tree is then processed by the **sentence planner**, which selects syntactic templates for expressing individual propositions, and aggregates them to produce the utterance's full syntactic structure. The pragmatic marker insertion component then modifies the syntactic structure locally to produce various pragmatic effects, depending on the markers' insertion constraints. The lexical choice component selects the most appropriate lexeme for each content word, given the lexical selection parameters. Finally, the RealPro **surface realizer** (Lavoie and Rambow 1997) converts the final syntactic structure into a string by applying surface grammatical rules, such as morphological inflection and function word insertion. When integrated into a dialogue system, the output of the realizer is annotated for prosodic information by the prosody assigner, before being sent

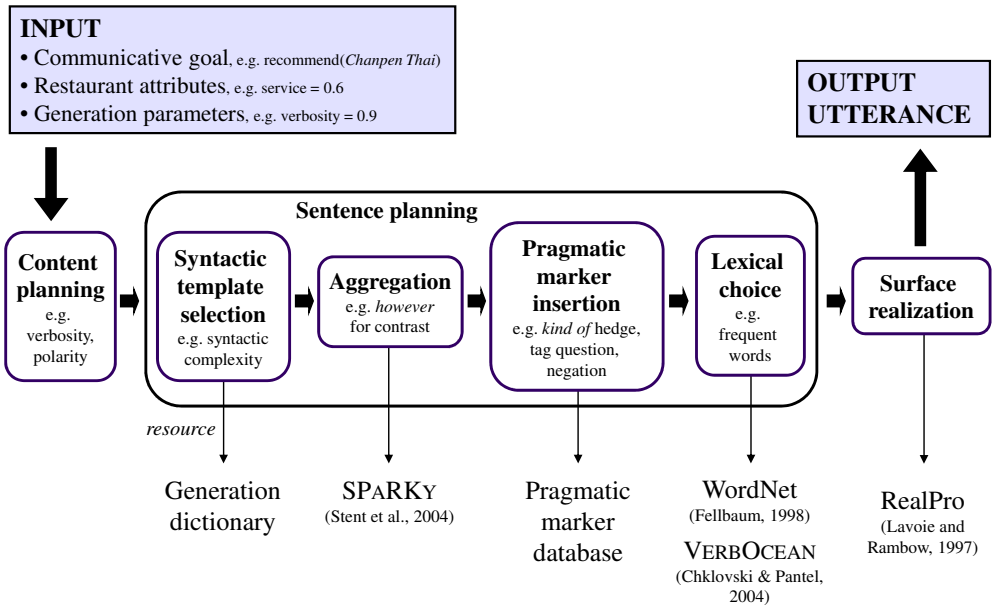


Figure 1
The architecture of the PERSONAGE base generator.

to the text-to-speech engine to be converted into an acoustic signal. PERSONAGE does not currently express personality through prosody, although there are studies that could be used to develop such parameters (Scherer 1979; Furnham 1990).

Figure 1 also indicates the modules in which PERSONAGE introduces parameters to produce and control personality-based linguistic variation. The generation parameters are shown in Table 2 and organized into blocks that correspond to the modules of the architecture in Figure 1. Compare Figure 1 to Table 2. As mentioned above, all of PERSONAGE's parameters are motivated by findings in the personality psychology literature. However the mapping from a finding to parameters represents a set of hypotheses about how the finding can be implemented, as discussed in more detail in (Mairesse and Walker 2007; Mairesse 2008).

Table 2 includes a description for each parameter that explains what the parameter does and often includes an example. For example, there are 12 **content planning** parameters shown in the first block of Table 2; these control aspects of utterances such their verbosity, rhetorical structure, content selection parameters such as positive content, and the level of redundancy and restatement (Walker 1993). Table 2 also includes 13 **pragmatic marker** parameters, which we believe to be completely novel. These include the introduction of HEDGES and TAG QUESTIONS. We are not aware of any other generators that produce the range of pragmatic variation illustrated here. Note also that Figure 1 indicates that the **lexical choice** parameters in Table 2 make use of multiple online lexical resources such as WordNet and VERBOCEAN to support lexical variation. The LEXICAL FREQUENCY parameter is calculated with respect to a corpus.

Furthermore, while some parameters primarily have a linear effect on an utterance (e.g., verbosity), other parameters are highly non-linear (e.g., the effect of inserting two expletives rather than one is not as strong as the effect of inserting one expletive rather than none). Parameters are therefore modeled as having either **continuous (C)** or **binary**

Table 2
PERSONAGE's generation parameters. The **Type** column indicates whether the stylistic effect is modeled as a continuous (C) or binary (B) parameter (i.e., resulting in continuous or binary parameter estimation models). Aggregation operation parameters are selection probabilities (C).

Parameter	Type	Description
Content planning:		
VERBOSITY	C	Control the number of propositions in the utterance
RESTATEMENTS	C	Paraphrase an existing proposition, e.g. *'Chanpen Thai has great service, it has fantastic waiters'*
REPETITIONS	C	Repeat an existing proposition
CONTENT POLARITY	C	Control the polarity of the propositions expressed, i.e. referring to negative or positive attributes
REPETITIONS POLARITY	C	Control the polarity of the restated propositions
CONCESSIONS		Emphasize one attribute over another, e.g. *'even if Chanpen Thai has great food, it has bad service'*
CONCESSIONS POLARITY	C	Determine whether positive or negative attributes are emphasized
POLARIZATION	C	Control whether the expressed polarity is neutral or extreme
POSITIVE CONTENT FIRST	C	Determine whether positive propositions — including the claim — are uttered first
REQUEST CONFIRMATION	B	Begin the utterance with a confirmation of the restaurant's name, e.g. *'did you say Chanpen Thai?'*
INITIAL REJECTION	B	Begin the utterance with a mild rejection, e.g. *'I'm not sure'*
COMPETENCE MITIGATION	B	Express the speaker's negative appraisal of the hearer's request, e.g. *'everybody knows that ...'*
Syntactic template selection:		
SELF-REFERENCES	C	Control the number of first person pronouns
SYNTACTIC COMPLEXITY	C	Control the syntactic complexity (syntactic embedding)
TEMPLATE POLARITY	C	Control the connotation of the claim, i.e. whether positive or negative affect is expressed
Aggregation operations:		
PERIOD	C	Leave two propositions in their own sentences, e.g. *'Chanpen Thai has great service. It has nice decor.'*
RELATIVE CLAUSE	C	Aggregate propositions with a relative clause, e.g. *'Chanpen Thai, which has great service, has nice decor'*
WITH CUE WORD	C	Aggregate propositions using *with*, e.g. *'Chanpen Thai has great service, with nice decor'*
CONJUNCTION	C	Join two propositions using a conjunction, or a comma if more than two propositions
MERGE	C	Merge the subject and verb of two propositions, e.g. *'Chanpen Thai has great service and nice decor'*
ALSO CUE WORD	C	Join two propositions using *also*, e.g. *'Chanpen Thai has great service, also it has nice decor'*
CONTRAST - CUE WORD	C	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	C	Justify a proposition using *because, since, so*, e.g. *'Chanpen Thai is the best, because it has great service'*
CONCEDE - CUE WORD	C	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	C	Restate a proposition by repeating only the object, e.g. *'Chanpen Thai has great service, nice waiters'*
OBJECT ELLIPSIS	C	Restate a proposition after replacing its object by an ellipsis, e.g. *'Chanpen Thai has ..., it has great service'*
Pragmatic markers:		
SUBJECT IMPLICITNESS	C	Make the restaurant implicit by moving the attribute to the subject, e.g. *'the service is great'*
STUTTERING	C	Duplicate the first letters of a restaurant's name, e.g. *'Ch-ch-anpen Thai is the best'*
PRONOMINALIZATION	C	Replace occurrences of the restaurant's name by pronouns
NEGATION	B	Negate a verb by replacing its modifier by its antonym, e.g. *'Chanpen Thai doesn't have bad service'*
SOFTENER HEDGES	B	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	B	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	B	Insert an initial back-channel (*yeah*, *right*, *ok*, *I see*, *oh*, *well*), e.g. *'Well, Chanpen Thai has great service'*
FILLED PAUSES	B	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	B	Insert an exclamation mark, e.g. *'Chanpen Thai has great service!'*
EXPLETIVES	B	Insert a swear word, e.g. *'the service is damn great'*
NEAR EXPLETIVES	B	Insert a near-swear word, e.g. *'the service is darn great'*
TAG QUESTION	B	Insert a tag question, e.g. *'the service is great, isn't it?'*
IN-GROUP MARKER	B	Refer to the hearer as a member of the same social group, e.g. *pal*, *mate* and *buddy*
Lexical choice:		
LEXICON FREQUENCY	C	Control the average frequency of use of each content word, according to BNC frequency counts
LEXICON WORD LENGTH	C	Control the average number of letters of each content word
VERB STRENGTH	C	Control the strength of the verbs, e.g. *'I would suggest'* vs. *'I would recommend'*

(B) values, as illustrated in column **Type** of Table 2. The models for continuous and binary parameters are trained using different algorithms. Section 3 below will provide examples of learned models of both types.

In addition, since generation decisions can be non-deterministic, some continuous parameter values are generation decision probabilities, e.g. the input to aggregation parameters such as CONJUNCTION is the probability that the aggregation operation is selected to combine any pair of propositions in the utterance (e.g., CONJUNCTION aggregates two propositions with the conjunction *and*). If the propositions cannot be aggregated because of syntactic constraints, another aggregation operation is sampled until the aggregation is successful. Complete details on the implementation of individual parameters can be found in (Mairesse 2008) and in (Mairesse and Walker 2010).

To make PERSONAGE as domain-independent as possible, the input parameter values are normalized between 0 and 1 for continuous parameters, and to 0 or 1 for binary parameters. For example, a VERBOSITY parameter of 1 maximizes the utterance's verbosity given the input, regardless of the actual number of propositions expressed. In order to ensure naturalness over the full parameter range, the maximum value of some continuous parameters is associated with an input-independent threshold (e.g., there cannot be more than two repeated propositions per utterance). While the goal of the base generator is to satisfy its input parameters, it cannot guarantee that all input parameter values will be reflected in the utterance due to constraints on the input content plan and other parameters. A consequence is that non-deterministic decision points are introduced to satisfy these naturalness constraints (e.g., if too many pragmatic marker parameters are enabled, only a random subset will appear in the utterance). Therefore, the only assumption we make regarding the impact of parameter values on the generation process is that they affect the likelihood of observing their intended effect over a large set of utterances.

3. Generation of Personality through Data-driven Parameter Estimation

While PERSONAGE-RB uses handcrafted parameter settings to convey different personality traits, PERSONAGE-PE relies on parameter estimation models to estimate the parameter values in Table 2 from target personality scores. At training time, our method requires the following steps:

1. Use a base generator to produce multiple utterances by randomly varying its parameters (see Section 3.1);
2. Ask human subjects to evaluate (rate) the personality/style of each utterance;
3. Train statistical models predicting the parameter values from the personality ratings (see Section 4.1);
4. Select the best model for each parameter via cross-validation (see Section 4.2).

At generation time, the models are used to predict the optimal set of generation parameters given a set of target personality scores, and the base generator is called once with the predicted parameter values. The architecture for the PE method is shown in Figure 2.

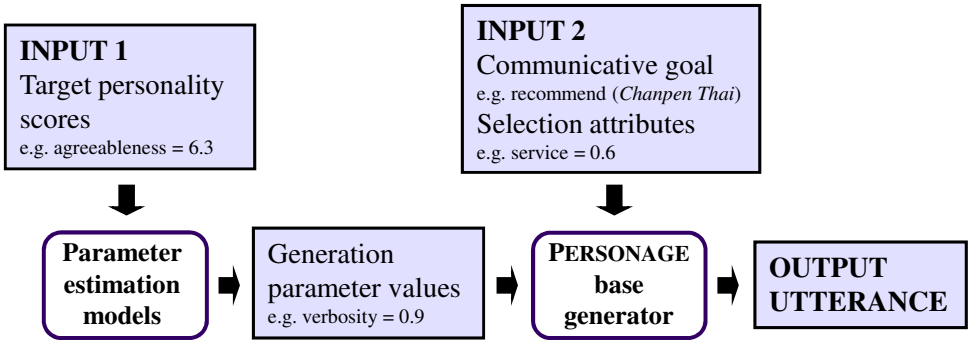


Figure 2
PERSONAGE-PE's parameter estimation framework.

In contrast with the overgenerate-score (OS) method discussed in Section 1.1, parameter estimation models predict generation decisions directly from input personality scores, in the spirit of the approach of Paiva and Evans (2005). However, whereas Paiva and Evans' approach *searches* for the generation decisions that will yield the optimal target scores according to their model, our PE method does not involve any search, as generation decisions are assumed to be conditionally independent given the target personality, and treated as dependent variables in individual models.

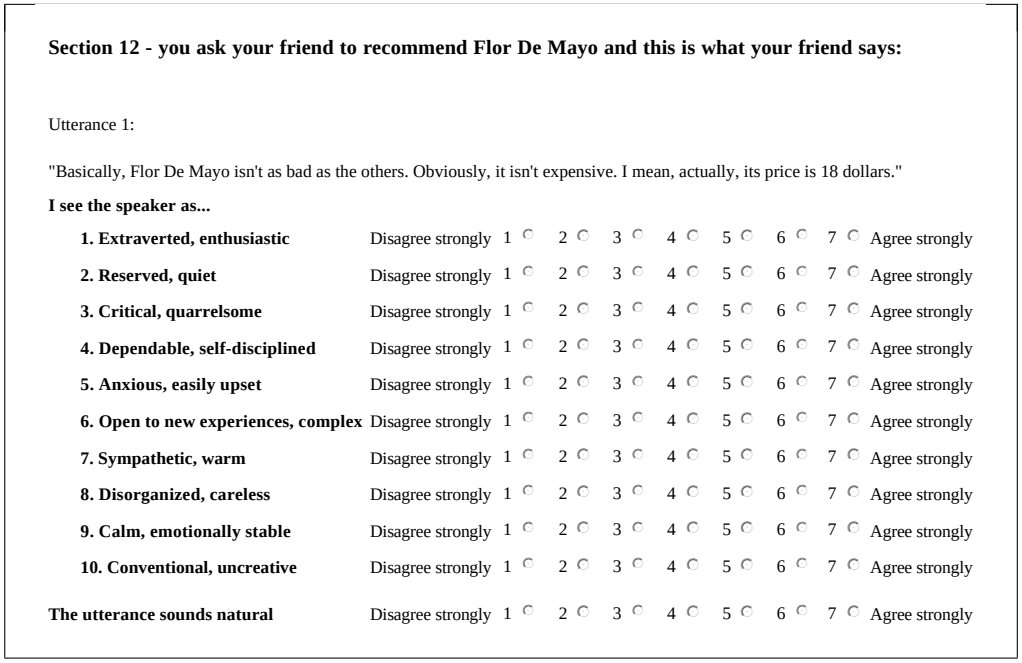
This section further details the steps required for training parameter estimation models. We first explain in Section 3.1 how we collect the judge's ratings for our training set. Then Section 3.2 analyzes the coverage and naturalness of the collected data. Finally, Section 3.3 describes how the models are trained.

3.1 Collecting Judgments of Random Sample

In order to train the parameter estimation models, the first step is to collect a dataset mapping generation decisions to personality ratings. This involves the following substeps:

1. Generate a sample of random utterances that produces examples covering the full range of all of the 67 PERSONAGE parameters as shown in Table 2.
2. Log the generation decisions that were made to produce each utterance.
3. Judges rate the random sample with a standard personality test shown in Figure 3, based on (Gosling, Rentfrow, and Swann 2003). This results in each utterance in the sample being labelled with five scalar values, one for each of the Big Five traits.

To be the basis for training a high performing statistical generator, the random sample must satisfy two properties. First, it must cover the full range of scalar values for each Big Five trait or there will not be enough training data to learn how to produce utterances manifesting those values. Second, the randomly produced utterances must be natural enough to produce stable personality judgments. The only way to verify that the random sample satisfies these properties is by first generating the random



Section 12 - you ask your friend to recommend Flor De Mayo and this is what your friend says:

Utterance 1:

"Basically, Flor De Mayo isn't as bad as the others. Obviously, it isn't expensive. I mean, actually, its price is 18 dollars."

I see the speaker as...

1. Extraverted, enthusiastic	Disagree strongly	1	2	3	4	5	6	7	Agree strongly
2. Reserved, quiet	Disagree strongly	1	2	3	4	5	6	7	Agree strongly
3. Critical, quarrelsome	Disagree strongly	1	2	3	4	5	6	7	Agree strongly
4. Dependable, self-disciplined	Disagree strongly	1	2	3	4	5	6	7	Agree strongly
5. Anxious, easily upset	Disagree strongly	1	2	3	4	5	6	7	Agree strongly
6. Open to new experiences, complex	Disagree strongly	1	2	3	4	5	6	7	Agree strongly
7. Sympathetic, warm	Disagree strongly	1	2	3	4	5	6	7	Agree strongly
8. Disorganized, careless	Disagree strongly	1	2	3	4	5	6	7	Agree strongly
9. Calm, emotionally stable	Disagree strongly	1	2	3	4	5	6	7	Agree strongly
10. Conventional, uncreative	Disagree strongly	1	2	3	4	5	6	7	Agree strongly
The utterance sounds natural	Disagree strongly	1	2	3	4	5	6	7	Agree strongly

Figure 3
The Ten Item Personality Inventory used in our experiments to calculate values for the Big Five traits, as modified for our experimental setting.

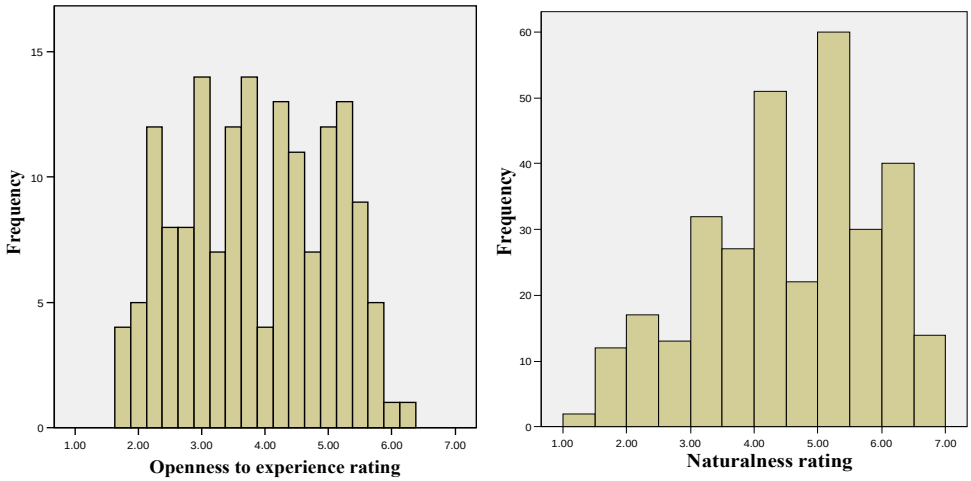
sample and then analyzing the judge's ratings. We generated 160 random utterances to constitute our random sample. Table 3 shows examples of random utterances and the scalar ratings for each trait that result from the judgment collection process.

A major advantage of the Big Five framework is that it offers standard validated questionnaires (John, Donahue, and Kentle 1991; Costa and McCrae 1992; Gosling, Rentfrow, and Swann 2003). Figure 3 shows the Ten Item Personality Inventory (TIPI) that we used to collect the personality judgments (Gosling, Rentfrow, and Swann 2003), adapted to our domain and task. The TIPI produces a scalar rating for each of the Big Five traits ranging from 1 (e.g., highly neurotic) to 7 (e.g., very stable), and it was shown to correlate well with longer questionnaires such as the Big Five Inventory, with convergent correlations of .87, .70, .75, .81, and .65 for extraversion, emotional stability, agreeableness, conscientiousness and openness to experience, respectively (Gosling, Rentfrow, and Swann 2003). While the TIPI has mostly been used as a self-report measure of personality, it has also been used to assess personality perceptions of observers, e.g. based on short social interactions (Srivastava, Guglielmo, and Beer 2010) or social networking websites (Gosling, Gaddis, and Vazire 2007). The judges in our experiment were researchers and postgraduate students in psychology, history and anthropology, who were familiar with the Big Five trait theory, but not with natural language generation. They were all native speakers of English. As illustrated in Figure 3, the judges were asked to rate each utterance in the random sample using the TIPI scale. They were instructed to rate the utterance as if it had been uttered by a friend responding in a dialogue to a request to recommend restaurants. Each judge rated the same sets of utterances corresponding to 20 communicative goals, 16 utterances per

Table 3
Example outputs of PERSONAGE with random parameter settings (random utterances), and scalar personality trait values after collection of TIPI judgments. *Extra* = extraversion, *ems* = emotional stability, *agree* = agreeableness, *consc* = conscientiousness, and *open* = openness to experience.

#	Communicative goal	PERSONAGE's output	Ratings
1.a	**compare (Kin Khao, Tossed)**	Kin Khao and Tossed are bloody outstanding. Kin Khao just has rude staff. Tossed features sort of unmannered waiters, even if the food is somewhat quite adequate.	extra=4.75 ems=6.00 agree=3.75 consc=4.00 open=3.25
1.b	**compare (Kin Khao, Tossed)**	Err... I am not really sure. Tossed offers kind of decent food. Mmhm... however, Kin Khao, which has quite ad-ad-adequate food, is a thai place. You would probably enjoy these restaurants.	extra=4.00 ems=4.00 agree=5.50 consc=4.75 open=5.00
2.a	**compare (Le Rivage, Pintaile's Pizza)**	Ok, right, Pintaile's Pizza is rather cheap. The atmosphere is bad, isn't it? I guess Le Rivage is in Manhattan, also its price is... it's expensive, you know.	extra=1.75 ems=3.50 agree=5.25 consc=4.00 open=2.75
2.b	**compare (Le Rivage, Pintaile's Pizza)**	Yeah, I thought everybody knew that Le Rivage's price is around 40 dollars, it's located in Manhattan with kind of poor atmosphere, Pintaile's Pizza is located in the Upper East Side and its price is 14 dollars. They're... I would advise these restaurants, alright?	extra=5.00 ems=4.25 agree=4.25 consc=5.75 open=3.00
3.a	**recommend (Chimi-churri Grill)**	Obviously, oh God Chimichurri Grill has like, quite adequate food, so it's the only re-restaurant I would consider, and it's a latin american place and located in Midtown West with ra-ra-rather friendly staff.	extra=4.50 ems=3.75 agree=5.25 consc=4.00 open=4.50
3.b	**recommend (Chimi-churri Grill)**	I see, oh Chimichurri Grill is a latin american place with sort of poor atmosphere. Although it doesn't have rather nasty food, its price is 41 dollars. I suspect it's kind of alright.	extra=2.50 ems=4.50 agree=3.50 consc=4.75 open=4.25
4.a	**recommend (Cent anni)**	Did you say Ce-Cent'anni? I see, I mean, I would consider it because it has friendly staff and tasty food, you know buddy.	extra=4.75 ems=5.00 agree=6.25 consc=6.25 open=5.25
4.b	**recommend (Cent anni)**	I am not sure. Cent'anni is... it's located in Manhattan, also the atmosphere is somewhat bloody poor, but it features tasty food though. Actually, this eating house, which provides quite acceptable service, is an italian restaurant. It's sort of the best eating place of its kind.	extra=4.25 ems=4.50 agree=4.25 consc=4.25 open=5.75

goal, one set at a time. The order of the sets and the order of the utterances within each set were both randomized. The judges were asked to read all the utterances in a set before rating them. Eight utterances out of sixteen were randomly generated for each communicative goal. The remaining utterances were generated using the handcrafted parameter settings of PERSONAGE-RB for each end of each Big Five trait (Mairesse and



(a) Distribution of openness scores on the random sample

(b) Distribution of naturalness scores on the random sample

Figure 4
The distribution of training data samples for openness personality judgments and naturalness.

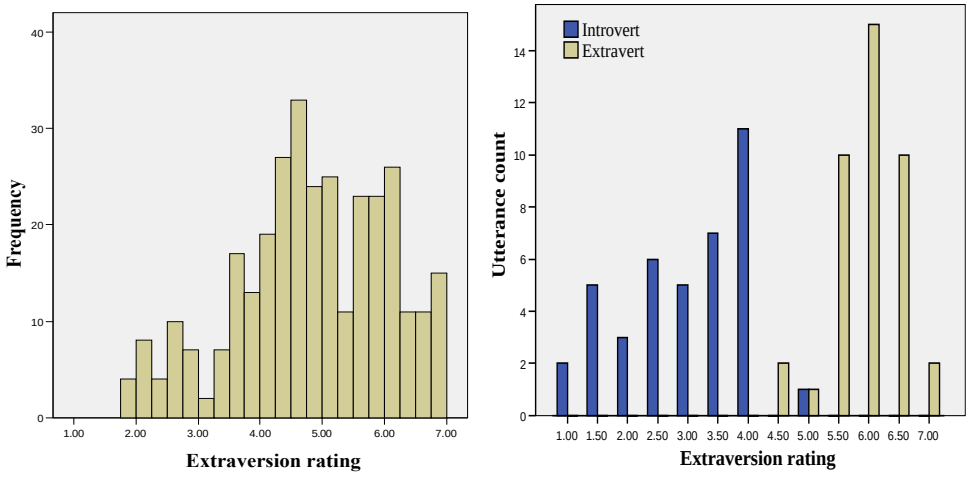
Walker 2007). The rule-based utterances are used as a comparison point, not for training the models. The same methodology was used to collect additional extraversion ratings for another set of 160 random and 80 rule-based utterances in a separate experiment, resulting in 320 random and 240 rule-based utterances for that trait, and 160 random utterances and 40 rule-based utterance for each of the other four traits. Examples of the resulting scalar ratings are shown in Table 3. The judges also evaluated the naturalness of each utterance on the same scale.

3.2 Generation Range and Naturalness

Analysis of the collected ratings of the random utterances shows that 67.8% of the utterances were rated as natural (rating above or equal to 4), with an average naturalness rating of 4.38 out of 7. Figure 4 shows the distributions of openness to experience and naturalness ratings. Figure 4a illustrates that most randomly generated utterances are not perceived as projecting an extreme personality. Table 4 examines whether randomly generated utterances can hit the extreme ends of each trait scale, by tabulating the most extreme ratings obtained from the 8 random utterances generated for each communicative goal with the ratings of the rule-based utterance generated from the same goal. This comparison provides useful information regarding (a) the potential of data-driven models to outperform handcrafted methods, and (b) whether our training corpus is large enough to capture the range of behavior we intend to convey. Paired t-tests over 20 communicative goals show that on average the most extreme random utterance is significantly more extreme for the positive end of the extraversion, emotional stability and agreeableness scales, and significantly more extreme for *both* ends of the conscientiousness and openness to experience scales ($p < .05$, two-tailed). However, random utterances are not perceived as as introverted as those generated

Table 4
For each communicative goal, the most extreme rating of the random utterances (*Random*) is compared with the ratings obtained for rule-based utterances (*Rule-based*). Ratings are averaged over 20 content plans and over all judges. • = the ratings of the random utterances are significantly more extreme (○ = more moderate) than the ratings of rule-based utterances ($p < .05$, two-tailed).

Method	**Rule-based**		**Random**	
Trait	**Low**	**High**	**Lowest**	**Highest**
Extraversion	2.96	5.98	3.60 ○	6.23 •
Emotional stability	3.29	5.96	3.05	6.25 •
Agreeableness	3.41	5.66	3.26	6.01 •
Conscientiousness	3.71	5.53	3.11 •	5.93 •
Openness to experience	2.89	4.21	2.28 •	5.48 •
Naturalness	4.59		4.38	



(a) Distribution of extraversion scores on the random sample

(b) Distribution of extraversion scores on the rule-based sample

Figure 5
The distribution of extraversion judgments for utterances generated using random parameters (*Random*) and handcrafted parameters derived from psychology studies (*Rule-based*).

using the introvert parameter settings (see **Rule-based/Low** column for extraversion). Compare the distributions of judgments for the rule-based extraversion utterances with the judgments on the random sample shown in Figure 5. Nevertheless, these results suggest that randomizing PERSONAGE's parameters produces a wide range of variation with an utterance sample of less than 10 utterances, for any communicative goal.

The bottom row of Table 4 also compares the *naturalness* of the random utterances with the naturalness of the rule-based utterances produced by PERSONAGE-RB (Mairesse and Walker 2007; Mairesse 2008). Results suggest that the random utterances

Table 5
Average inter-rater correlation for the rule-based and random utterances. All correlations are significant at the $p < .05$ level (two-tailed).

Parameter set	Rule-based	Random
Extraversion	.73	.30
Emotional stability	.67	.33
Agreeableness	.54	.40
Conscientiousness	.42	.26
Openness to experience	.44	.28

are less natural than the rule-based utterances, and this difference is close to significance ($p = .075$, two-tailed t-test).

It is also important to quantify the quality of the annotations by evaluating the *inter-rater agreement* between the judges. Table 5 shows that the judges agree significantly on the ratings of random utterances for all Big Five traits ($p < .05$, two-tailed), with correlations ranging from .26 (conscientiousness) to .40 (agreeableness), which are high correlations for human perceptual judgements. However the agreement is lower than on the rule-based utterances. A possible explanation of both the naturalness differences and rater agreement is that the random generation decisions sometimes produce utterances with inconsistent personality cues, which can be interpreted in different ways by the judges. For example, the utterance *'Err... I am sure you would like Chanpen Thai!'* expresses markers of both introversion (filled pause) and extraversion (exclamation mark).

3.3 Training Parameter Estimation Models

Parameter estimation requires a series of pre-processing steps, in order to ensure that the models' output is re-usable by the PERSONAGE base generator. The initial dataset includes the random sample annotated with the generation decision features shown in Table 2, together with the average judges' ratings along each Big Five dimension, as described in Section 3.1. The following transformations are performed before the learning phase:

- Reverse input and output: As parameter estimation models map from personality scores to generation parameters, the generation decisions are set as the dataset's output variables and the averaged personality ratings as the input features.
- Predict parameters individually: A new dataset is created for each output variable — i.e. generation parameter — as the statistical models we use only predict one output. We thus make the simplifying assumption that PERSONAGE's generation parameters are independent.[1]
- Map output variables into PERSONAGE's input space: The generation decisions made when generating each utterance in the random sample

1 While this assumption is violated by the internal constraints of PERSONAGE's generation process, Section 4.3 investigates the extent to which this violation affects the models' accuracy.

were recorded. In order to ensure that the parameter estimation models' output is re-usable by the base generator, the generation decision space is mapped to PERSONAGE's input parameter space. The conversion is dependent on the type of generation parameter:

  - **Continuous parameters:** Generation decision values are normalized over all random utterances, resulting in values between 0 and 1. E.g. a VERBOSITY parameter value of 1 indicates the utterance with the largest number of propositions in the utterance set.
  - **Aggregation operation probabilities:** Frequency counts of aggregation operations realizing a specific rhetorical relation are divided by the number of occurrences of the rhetorical relation in the utterance. This ratio is the maximum likelihood estimate of the conditional probability of the aggregation operation given the rhetorical relation. E.g. if out of four INFER relations in the utterance, only one is realized using the MERGE operation, the value for the INFER - MERGE parameter is .25 for that utterance.
  - **Binary parameters:** No processing is required as generation decisions are already boolean. E.g. if an exclamation mark was inserted in the utterance, the EXCLAMATION parameter value is set to 1 rather than 0.

- Feature selection: Personality traits that do not correlate with a generation parameter with a Pearson's correlation coefficient above .1 are removed from that parameter's dataset. 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.

Once the data is partitioned into datasets mapping the relevant personality dimensions (the features) to each generation parameter (the dependent variable), it can be used to train parameter estimation models predicting the most appropriate parameter value given target personality scores. Parameters are estimated using either regression or classification models, depending on whether they are continuous (e.g., VERBOSITY) or binary (e.g., EXCLAMATION). Recall that Table 2 indicated for each parameter whether it is continous (C) or binary (B). In order to identify what model should be used for each parameter, we compare various learning algorithms using the Weka toolbox (Witten and Frank 2005).

Continuous parameters in Table 2 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. Regression models are evaluated using the correlation between the model's predictions and the actual parameter values in the test data.

Binary parameters in Table 2 are modeled using classifiers that predict whether the parameter should be *enabled* or *disabled*. We test a Naive Bayes classifier (NB), a C4.5 decision tree (J48), a nearest neighbour classifier using one neighbour (NN), the Ripper rule-based learner (JRIP), the AdaBoost boosting algorithm (ADA) and a support vector machines classifier with a linear kernel (SVM). Unless specified, the learning algorithms use Weka's default parameter values.

4. Evaluation

This section first details some of the parameter estimation models trained on the data collected in Section 3. The models' predictive power is then evaluated by doing a 10-

Rules				Weight
if extraversion	$>$	6.42	then enabled else disabled	1.81
if extraversion	$>$	4.42	then enabled else disabled	0.38
if extraversion	$\leq$	6.58	then enabled else disabled	0.22
if extraversion	$>$	4.71	then enabled else disabled	0.28
if agreeableness	$>$	5.13	then enabled else disabled	0.42
if extraversion	$\leq$	6.58	then enabled else disabled	0.14
if extraversion	$>$	4.79	then enabled else disabled	0.19
if extraversion	$\leq$	6.58	then enabled else disabled	0.17

Figure 6
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.

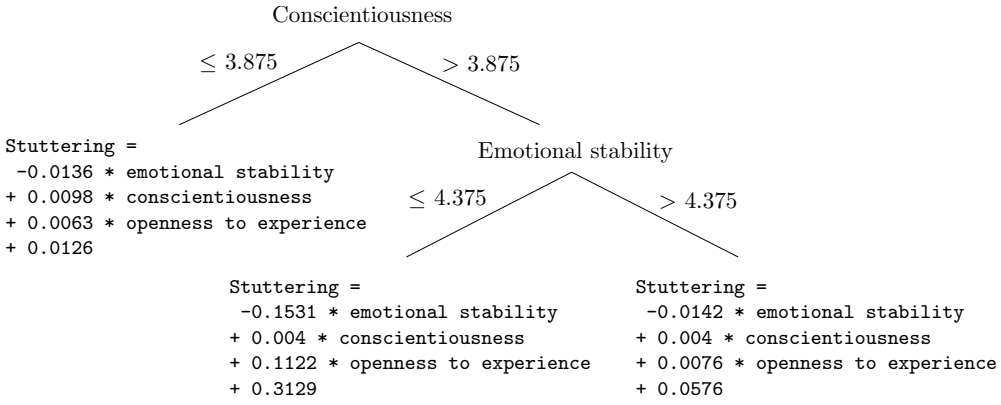


Figure 7
M5′ model tree predicting the STUTTERING parameter.

fold cross-validation in Section 4.2. Finally, Section 4.3 evaluates human perceptions of utterances generated using the models.

4.1 Qualitative Model Evaluation

Before discussing our quantitative results, we use Figures 6, 7 and 9 to illustrate how the learned models predict generation parameters from input personality scores. Note that sometimes the best performing model is non-linear. For example, given input trait values, the AdaBoost model in Figure 6 outputs the class yielding the largest sum of weights for the rules returning that class. The first rule of the EXCLAMATION model in Figure 6 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. Figure 6 also illustrates how personality traits that do not have an effect on the parameter are removed, i.e. extraversion and agreeableness are the traits that affect the use of exclamation marks. The STUTTERING model tree in Figure 7 lets us calculate that a low emotional stability (1.0) together with a neutral conscientiousness (4.0) and openness to experience (4.0) yield a parameter value of .62 (see bottom-left linear model), whereas a neutral emotional stability decreases the

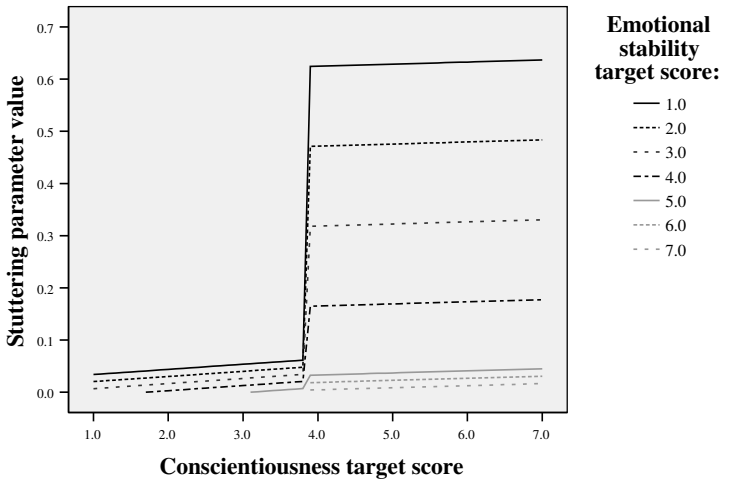


Figure 8
Variation of the predicted STUTTERING parameter value based on the model in Figure 7 for different emotional stability and conscientiousness target scores. All other trait scores are set to 4.0 out of 7.

Content polarity =

-0.102 · emotional stability +
0.970 · agreeableness +
- 0.110 · conscientiousness +
0.013 · openness to experience +
0.054

Figure 9
SVM model with a linear kernel predicting the CONTENT POLARITY parameter.

value down to .17. The full parameter range obtained when varying both emotional stability and conscientiousness is illustrated in Figure 8, which shows that the .5 cut-off point can be reached for low emotional stability scores and mid-range conscientiousness scores. The linear model in Figure 9 shows that agreeableness has a strong effect on the CONTENT POLARITY parameter (.97 weight), but emotional stability, conscientiousness and openness to experience also influence the parameter value.

Inspection of the learned models provides interesting information about whether findings in the psychology literature carry over to our domain. However, in order to optimize the overall generation performance, we rely on a quantitative analysis for selecting individual models.

4.2 Cross-validation on Corpus of Expert Judgments

We identify the best performing model(s) for each generation parameter via a 10-fold cross-validation. For continuous parameters, Table 6 evaluates modeling accuracy by comparing the correlations between the model's predictions and the actual parameter values in the test folds. Table 7 reports results for binary parameter classifiers, by

Table 6
Pearson's correlation coefficient between parameter model predictions and continuous parameter values, for different regression models. Parameters that do not correlate with any trait are omitted. Results are averaged over a 10-fold cross-validation, and the best result for each parameter is in bold.

Continuous parameters	LR	M5	SVM
Content planning:			
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**
REPETITION POLARITY	0.02	**0.15**	0.06
CONCESSIONS	**0.23**	0.23	0.12
CONCESSION POLARITY	-0.01	**0.16**	0.07
POLARIZATION	0.20	**0.21**	0.20
Syntactic template selection:			
SYNTACTIC COMPLEXITY	0.10	**0.33**	0.26
TEMPLATE POLARITY	0.04	0.04	**0.05**
Aggregation operations:			
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	0.16	**0.23**	0.17
PRONOMINALIZATION	**0.22**	0.20	0.17
Lexical choice:			
LEXICON FREQUENCY	0.21	**0.21**	0.19
LEXICON WORD LENGTH	**0.18**	0.18	0.15

comparing the F-measures of the *enabled* class. The F-measure measures how well the models predict the enabled class given the small proportion of instances labelled as *enabled* in the training utterances, i.e. it is less sensitive to class-imbalance than classification accuracy. Models producing the best cross-validation results are identified in bold for each parameter; parameters that produce a poor modeling accuracy are omitted. Because of the large number of parameters tested simultaneously in each training utterance, many reported accuracies are relatively low. As our training approach aims at including all parameters that can potentially convey personality, we include models with correlations or F-measures above .05 in our system, and let individual models learn the extent to which their parameter will affect the trained system.

Table 6 shows that the CONTENT POLARITY parameter is modeled the most accurately, with the SVM model in Figure 9 producing a correlation of .47 with the true parameter values in Table 6. 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. The SYNTACTIC COMPLEXITY and VERBOSITY parameters are also modeled successfully, with correlations of .33 and .26 using a model tree. The model tree controlling the STUTTERING parameter illustrated

Table 7
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, and the best result for each parameter is in bold.

Binary parameters	NB	J48	NN	JRIP	ADA	SVM
Content planning:						
REQUEST CONFIRMATION	0.00	0.00	**0.07**	0.05	0.04	0.04
Pragmatic markers:						
SOFTENER HEDGES						
kind of	0.00	0.00	**0.16**	0.13	0.11	0.10
quite	**0.14**	0.08	0.09	0.09	0.07	0.06
ok	**0.13**	0.07	0.06	0.05	0.05	0.05
FILLED PAUSES						
err	**0.32**	0.20	0.24	0.24	0.22	0.19
EXCLAMATION	0.23	0.34	0.36	0.37	**0.38**	0.34
EXPLETIVES	**0.27**	0.18	0.24	0.20	0.17	0.15
IN-GROUP MARKER	**0.40**	0.31	0.31	0.26	0.24	0.21
TAG QUESTION	**0.32**	0.21	0.21	0.17	0.15	0.13

in Figure 7 produces a correlation of .23. Concerning binary parameters, Although differences between the best performing models are not significant, Table 7 suggests 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 results suggest that the AdaBoost learning algorithm performs best for predicting the EXCLAMATION parameter, with an F-measure of .38 for the model in Figure 6.

These results show that there are large variations between model performance for a given parameter (e.g., CONCESSION POLARITY). This suggests that exploring different learning algorithms for individual parameters is beneficial. While the overall modeling accuracy can seem relatively low — e.g., there are only 4 parameters with correlations above $r = .25$ in Table 6 — it is important to keep in mind that the utterances were randomly generated, hence the effect of a specific parameter is likely to be affected by other random parameter values. The next section therefore evaluates whether the models are good enough to produce reliable effects on user perceptions.

4.3 Evaluation with Naive Subjects

The evaluation presented in Section 4.2 measures how well the PE models predict the parameters, using parameter settings in the random sample and expert judge ratings as the predictors. We relied on a small number of expert judges to minimize rating inconsistencies and facilitate the learning process. In order to test whether the PE method produces high quality outputs manifesting personality, we ran an experiment with 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.). We produced a set of 50 utterances for this experiment using the *best* performing models for each generation parameter shown in Tables 6 and 7. Given this model, we generate 5 utterances for each of 10 input communicative goals. 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. Here, we test whether

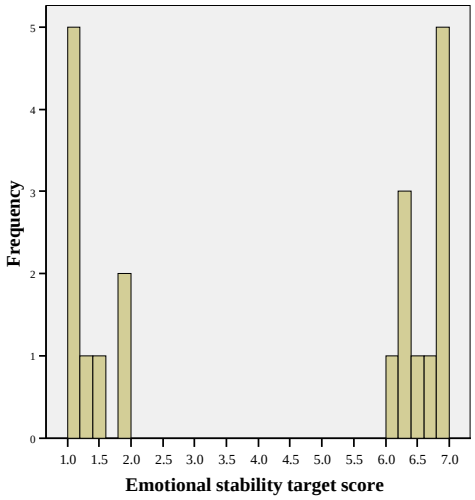


Figure 10
Distribution of the 20 emotional stability target scores, normally distributed over both extremes with a standard deviation of 10% of the full scale.

PE models can convey personality in extreme regions of the Big Five space. In order to generate a range of alternatives for each input communicative goal, all target scores are randomized around their initial value according to a normal probability distribution with a standard deviation of 10% of the full scale (see Figure 10).

All 50 utterances were evaluated by the 24 naive subjects using the Ten-Item Personality Inventory in Figure 3 (Gosling, Rentfrow, and Swann 2003). As in the training data collection (Section 3.1), the subjects rated one set of 5 utterances at a time, one for each communicative goal. The communicative goals and the utterances were presented in random order. To limit the experiment's duration, only the two traits with extreme target values are evaluated for each utterance. As a result, 20 utterances are evaluated for each trait, 10 of which were generated to convey the low end of that trait, and 10 of which target the high end of that trait. Each utterance was also evaluated for its naturalness as before. 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 8 shows several sample outputs and the mean personality ratings from the naive subjects for two communicative goals. For example, utterance 1.a projects a high extraversion through the insertion of an exclamation mark based on the model in Figure 6, 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 produced by the model in Figure 9.

4.3.1 Naturalness and Inter-rater Agreement. Figure 11 shows the distribution of the naturalness ratings. A Shapiro-Wilk test confirms that the ratings are normally distributed ($W = .97, p = .31$), with only one utterance out of 50 rated below 2.5 out of 7 on average. The average naturalness is 3.98 out of 7, with a rating of 4 indicating

Table 8
Example outputs from PERSONAGE-PE for a comparison (#1) and a recommendation (#2), with the average judges' personality (*Rating*) and naturalness (*Nat*) scores. Ratings are on a scale from 1 to 7, with 1 = very low (e.g. introvert) and 7 = very high (e.g., extravert).

#	Trait and target		Rating	Nat	Output utterance
1.a	Extraversion Agreeableness	high 6.90 high 7.00	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 7.00 high 6.21	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.
1.c	Agreeableness Openness to experience	low 1.28 high 7.00	4.04 4.52	4.58	Les Routiers and Radio Perfecto are outstanding. Radio Perfecto's price is 25 dollars, isn't it? The food is acceptable, also it's located in the East Village. Err... although Les Routiers offers adequate food, it's damn costly.
2.a	Extraversion Agreeableness	low 1.92 low 1.73	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 1.97 low 1.00	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.
2.c	Agreeableness Openness to experience	low 1.00 low 1.33	3.27 3.94	3.58	Trattoria Rustica is the only eating place that is any good. Err... it's located in Manhattan. This restaurant is an italian place with poor ambience. It's bloody costly, even if this eating house has friendly waiters you see?

neither agreement nor disagreement. This is lower than the naturalness scores obtained in Section 3.2 for the random training utterances collected using a small number of expert judges. The differences in naturalness judgements could possibly be due to (a) the different set of judges; (b) the fact that utterances conveying extreme personality are likely to be perceived as less natural; or (c) the fact that the expert judges made a very large number of judgements, and thus became accustomed to judging the outputs.

Table 9 reports the inter-rater correlation over all personality ratings, averaged over the 276 pairs of judges. The level of agreement between the naive subjects reflects the difficulty of the personality recognition task for humans, thus providing an upper bound on the performance to be expected from a model trained on human data. The judges agree modestly, with correlations ranging from .17 (openness to experience) to .41 (emotional stability). This agreement is lower than that for rule-based utterances, which could be due to the nature of the personality cues conveyed by PERSONAGE-RB's handcrafted parameters. However, this difference could also result from the use of *naive* judge, which we believe are less consistent in their personality judgments.

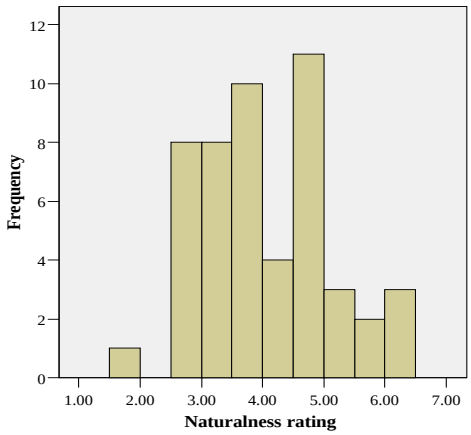


Figure 11
Distribution of naturalness ratings over the 50 utterances, averaged over all 24 judges. The mean naturalness rating is 3.98, with a standard deviation of 1.07.

Table 9
Average and standard deviation of the inter-rater correlations over the 276 pairs of judges.

Trait	$\bar{r}_{inter}$	σ_{inter}
Extraversion	.33	.22
Emotional stability	.41	.17
Agreeableness	.28	.23
Conscientiousness	.34	.18
Openness to experience	.17	.25
All	.34	.10

4.3.2 Modeling Accuracy. Modeling accuracy is measured using both the correlation and mean absolute error (on a scale from 1 to 7) between the target personality scores and the judges ratings. Given the relatively small number of distinct utterances being evaluated (20 per trait), it is important to take the non-determinism of PERSONAGE into account when evaluating the correlation between the target scores and the judges ratings. Evaluating correlations over the 480 ratings of each judge and utterance pair is likely to result in inflated significance level, as it does not account for the possibility that a specific outcome of PERSONAGE's random generation decisions could produce the intended personality, rather than accurate personality modeling. In order to address this issue, we report correlations r_{avg} between the target personality scores and the 20 personality ratings averaged over all 24 judges, i.e. the reported significance levels do take the number of distinct test utterances into account. Table 10 shows that extraversion is the dimension modeled the most accurately by the parameter estimation models, producing a .80 correlation between the target extraversion and the average subjects' ratings ($p < .001$). Emotional stability, agreeableness ratings also correlate strongly with the target scores, with correlations of .64 and .68, respectively ($p < .005$ and $p < .001$). The correlation for openness to experience is also relatively strong (.41), although it is not significant at the $p < .05$ level ($p = .07$). These correlations are unexpectedly high;

Table 10
Pearson's correlation coefficient r_{avg}, correlation significance level p, and absolute error e between the target personality scores and the mean utterance ratings averaged over 24 judges.

Trait	r_{avg}	p	e
Extraversion	.80 •	$p < .001$	1.89
Emotional stability	.64 •	$p = .002$	2.14
Agreeableness	.68 •	$p < .001$	2.38
Conscientiousness	-.02		2.79
Openness to experience	.41 •	$p = .07$	2.51

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

in corpus analyses, significant correlations as low as .05 to .15 are typically observed between averaged personality ratings and linguistic markers (Pennebaker and King 1999; Mehl, Gosling, and Pennebaker 2006). Although each utterance is used to test two hypotheses (i.e., rated for two traits), results for extraversion, emotional stability and agreeableness remain largely significant even after applying Bonferroni correction ($p < .001$, $p < .005$ and $p < .005$ respectively).

Conscientiousness is the only dimension whose ratings do not correlate with the target scores. The comparison with rule-based results in Table 11 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. It is also possible that inconsistencies in the training data prevented the models from learning accurate cues for conscientiousness, as Table 5 shows that the judges disagreed the most over that trait when rating the training utterances.

Table 10 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, because the judges tend to use the whole scale, with normally distributed ratings (Shapiro-Wilk tests, $p > .05$). This means that although the judges make distinctions leading to high correlations, the averaged ratings result in a compressed scale. This explains the large correlations despite the magnitude of the absolute error.

It is important to emphasize that generation parameter values were predicted based on five target personality scores. Thus, the results show that *individual* traits are perceived even when utterances project other traits as well, as would be expected according to the Big Five theory.

Table 11 compares the mean ratings of the utterances generated by PERSONAGE-PE with ratings of 20 utterances generated with the rule-based parameter settings for each extreme of each Big Five trait (40 for extraversion, resulting in 240 rule-based utterances in total). Although rating differences could be due to the different set of judges, Table 11 suggests that the handcrafted parameter settings project a more extreme personality for 6 traits out of 10. However, the parameter estimation models are not significantly worse than the rule-based generator for neuroticism, disagreeableness, unconscientiousness and openness to experience. Moreover, the differences between **low** and **high** average ratings for the parameter models shown in the right-hand-side of Table 11 are significant for all traits but conscientiousness ($p \leq .001$). Thus parameter estimation models can be used in applications that only require discrete binary variation.

Table 11
Comparison between the ratings of PERSONAGE-PE's utterances with extreme target values (*Param models*) and the expert judges' ratings for utterances generated using the *Rule-based* method of our previous work.

Method	**Rule-based**		**Param models**	
Trait	**Low**	**High**	**Low**	**High**
Extraversion	2.96	5.98	3.69 ○	5.06 ○
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 12
Pearson's correlation coefficient between the target personality scores and averaged ratings (r_{avg}) for each group of extreme targets, as well as the target score range.

	Correlation r_{avg}		**Range**	
Trait	**Low**	**High**	**Low**	**High**
Extraversion	.01	.05	.92	1.31
Emotional stability	.46	.55 ●	.97	.98
Agreeableness	.20	-.17	1.13	.70
Conscientiousness	.03	-.32	.97	.79
Openness to experience	.08	.46	.84	1.52

● statistically significant correlation
$p < .05$, ● $p = .08$ (two-tailed)

Additionally, we evaluate whether the naive subjects perceive the small differences *within* each group of extreme utterances. At the beginning of this section, we mentioned that the target scores used in the evaluation experiment were randomized according to a normal distribution around 1 or 7, with a standard deviation of 10% of the full scale (.60). Figure 10 shows the distribution of the target scores for emotional stability. We compute the correlation between the target scores and the average judges' ratings over each group of extreme target scores.[2] Table 12 show that the naive subjects failed to detect the small variation within each group, however results are close to significance for the positive end of the emotional stability scale, with a correlation of .55 for the emotionally stable group ($p = .08$). This suggests that parameter estimation models could model that trait with a high granularity given more training samples, as all that is needed is for our models to learn to trigger relevant personality markers within extreme regions of the stylistic space. For example, Figure 8 shows that the STUTTERING parameter value predicted by the model in Figure 7 varies considerably with emotional stability. Given an input conscientiousness and openness to experience scores of 4 out of 7, the

2 Because target scores are truncated to fit PERSONAGE-PE's input range (between 1 and 7), approximately half of the values in each group are either 1.0 or 7.0.

STUTTERING parameter becomes enabled (i.e., above .5) for emotional stability scores below 1.81 (see middle leaf node in Figure 7).

The low correlation observed for other traits — e.g. extraversion — shows that the high accuracy reported in Table 10 is due to the successful modeling of large variations between each end of the scale, rather than the small-scale variations within one side of the dimension. While these results are promising, future work should evaluate the granularity of parameter estimation models over the full range of the Big Five scales.

5. Discussion

This article presents a novel SNLG method based on **parameter estimation models** that estimate optimal generation parameters given target stylistic scores, which are then used directly by the base generator to produce the output utterance. We believe that dialogue applications such as spoken dialogue systems, interactive drama systems and intelligent tutoring systems would benefit from taking individual stylistic differences into account, and that theories of human personality traits represent an appropriate set of dimensions for a computational model of this adaptation. Starting from a communicative goal, we show how personality affects all phases of the language generation process, and that certain parameters such as the polarity of the content selected have a strong effect on the perception of personality. We present, to our knowledge, the first results of a human evaluation experiment using the perceptions of naive subjects to evaluate the stylistic variation of a language generator. The parameter estimation method is a computationally tractable generation method that does not require any overgeneration phase, and our results show that naive judges can recognize the intended system personality for most traits.

While most previous work on SNLG is based on the overgenerate and scoring (OS) framework (Langkilde and Knight 1998; Walker, Rambow, and Rogati 2002; Isard, Brockmann, and Oberlander 2006), our results show that direct parameter optimization is a viable alternative for stylistic control. Given the cost of an overgeneration phase, parameter estimation methods might be the only alternative for real-time generation, which is necessary for spoken dialogue interaction. The fact that the accuracy of OS methods depends on the complexity of the overgeneration phase prevents them from scaling to the large generation space required for stylistic variation. For example, an exhaustive search over PERSONAGE's output space would require more than 2^{67} parameter combinations, for each of which an utterance would have to be generated and scored.

On the other hand, the tractability of OS methods could be improved by pruning candidates throughout the generation process, as well as by using compact data structures — such as in the OpenCCG chart realizer [White, Rajkumar, and Martin 2007] — although we do not know of any SNLG system using this approach. The main advantage of the OS approach is that it can model global utterance features reflecting multiple generation decisions as well as input dependencies (e.g., sentence length). The predictive accuracy of scoring models trained on our data discussed in the introduction suggests that OS methods could be useful for stylistic SNLG when limited to a mild data-driven overgeneration phase, e.g. to rerank candidate utterances produced by stochastic parameter estimation models. The OS approach therefore can be seen as a way to relax the parameter independence assumptions required to learn robust models.

Parameter estimation models are trained on utterances generated in the application domain. In contrast with out-of-domain corpus based methods (Isard, Brockmann, and Oberlander 2006), in-domain personality annotations not only reduce data sparsity by

constraining the range of outputs, they also allow us to explicitly model the relation between stylistic factors and generation decisions. The parameter estimation approach is inspired by previous work by Paiva and Evans (2005), who present a data-driven method for stylistic control that does not require an overgeneration phase. We extend this work in multiple ways. First, we focus on the control of the speaker's personality, rather than stylistic dimensions extracted from corpora. Second, we present a method for learning parameter estimation models predicting generation decisions directly from input personality scores, whereas the method of Paiva and Evans (2005) requires a search for the optimal generation decision over the model's input space. Third, we present a perceptual evaluation of the generated stylistic variation, using naive human judges.

We modeled stylistic variation in terms of the Big Five personality traits, because these traits are widely accepted as the most important dimensions of behavioral variation within human beings (Norman 1963; Goldberg 1990). Collecting reliable personality annotations for in-domain utterances is a non-trivial task, as reflected by the moderate inter-rater agreement reported in Section 4.3.1. The difficulty is increased by the use of random generation decisions, which is necessary to ensure that the learned relation between personality ratings and generation parameters is not due to artificial correlations between generation decisions. One way to increase inter-rater agreement would be to reduce the number of parameters varied simultaneously for each utterance, thereby decreasing the chances of observing inconsistent markers. While this would require collecting a larger training set to model the same number of parameters, one could filter out irrelevant parameters in a first data collection phase, to focus on data quality in a second phase. A second way to improve the quality of the annotations would be to increase the size of each text sample, by simultaneously rating multiple utterances generated from the same parameters, rather than a single utterance. This is likely to produce models that are more robust to input variation, given a small increase in annotation effort. Despite the difficulty of the annotation task, our experiments show that parameter estimation models are able to detect a large range personality markers in our training set. The fact that those markers are successfully recognized by naive judges suggests that our approach is robust to ambiguities in the data.

While the Big Five traits are widely accepted in the psychology literature, they differ in terms of their impact on linguistic production. Results suggest that perceptions of agreeableness and extraversion are easier to model, whereas conscientiousness and openness to experience are more difficult. A possible explanation is that these traits are not conveyed well in PERSONAGE's narrow domain. However, previous personality recognition results suggest that observed openness to experience is difficult to model using general conversational data as well (Mairesse et al. 2007). It is therefore possible that this trait may not be expressed through spoken language as clearly as other traits.

The evaluation using naive judges in Table 10 shows that the average ratings correlate strongly with the target personality scores for all traits but conscientiousness, with correlation coefficients ranging from .41 (openness to experience) up to .80 (extraversion). It is important to note that the magnitudes of the model correlations are high compared to traditional correlations between personality and linguistic markers, which typically range from .05 to .10 (Pennebaker and King 1999; Mehl, Gosling, and Pennebaker 2006). This is possibly due to our experimental method, which simultaneously tests a small number of personality markers in a controlled experiment, whereas such markers are harder to extract from the varied language samples used in psychology studies.

In terms of limitations, even though PERSONAGE's parameters were suggested by psychological studies, some of them are not modeled successfully by the parameter estimation approach, and thus were omitted from Tables 6 and 7. This could be due to the relatively small training set size (160 utterances to optimize 67 parameters). However, even with a larger training set, it is possible that some of the parameters in Table 2 are not perceivable in our domain. Additionally, the parameter-independence assumption of the PE method could be responsible for the poor accuracy of some models. While this issue could possibly be resolved by training statistical models that simultaneously predict multiple dependent variables (e.g. structured prediction methods), this would exponentially increase the size of the parameter prediction space and further aggravate data sparsity issues.

We also find that parameter estimation models perform slightly worse when projecting extreme traits than the rule-based generator in the same domain (see Section 4.3.2). We believe this is due to the lack of extreme utterances in the training data to learn from. Future work should consider whether it is possible to flatten the distribution of training data, perhaps using more knowledge or active learning methods for NLG (Mairesse et al. 2010).

Another limitation of the current approach is that we assume the existence of a generation dictionary containing syntactic templates that, in some cases, express various pragmatic effects, even though most of PERSONAGE's variation is generated automatically. Our dictionary is currently handcrafted, but future work could build on research on methods for extracting the generation dictionary from data (Higashinaka, Walker, and Prasad 2007; Barzilay and Lee ; Snyder and Barzilay 2007).

Finally, it is likely that both the accuracy and coverage of parameter estimation models could be improved with a larger sample of judges and random utterances at development time. A larger number of judges would also smooth out rating inconsistencies and individual differences in personality perception, thus allowing the direct modeling of laypeople's perceptions by removing the need for expert judges.

Another important area for future work is to explore the interface between the language generator and the dialogue manager and text-to-speech (TTS) engine, since personality also affects aspects of dialogue such as initiative and prosodic realization (Vogel and Vogel 1986; Scherer 1979). It might be possible to apply a similar methodology to the parameterization of a dialogue manager and a TTS component, in order to project a consistent personality to the user, or to use a reinforcement learning approach to train the dialogue manager (Walker, Fromer, and Narayanan 1998; Singh et al. 2002), with personality judgments as the objective function. Our results suggest that personality can be recognized by manipulating the linguistic cues of a single utterance, but it is likely that additional cues — e.g. characterizing the system's dialogue strategy — could only make the user's perception of the system's personality more robust.

Nevertheless it is clear that the PE SNLG approach offers many advantages over handcrafted methods. Firstly, SNLG methods can scale more easily to other domains and other types of stylistic variation, as collecting data requires less effort than tuning a large number parameters. In order to apply our method to a new domain or task, the base generator would have to be modified. Although PERSONAGE was implemented with domain independence in mind by using general parameters, handling a new communicative goal (e.g., a user request) would require modifying the syntactic aggregation operations as well as the pragmatic marker insertion rules. Porting PERSONAGE to a new information presentation domain would only require modifying the syntactic templates in the generation dictionary. Once a base generator is available in the target domain, it can be used to collect ratings for any stylistic dimension. Our parameter

estimation method can then learn to control the generator's output from data, without any further handcrafting or parameter tuning. A second advantage of our parameter estimation method over handcrated approaches is that it can learn to target scalar combinations of the Big Five traits, whereas handcrafting all possible ways in which multiple stylistic dimensions can affect output utterances requires considerable effort. Thirdly, our parameter estimation method can easily model continuous stylistic dimensions, while producing such models by hand is not tractable. For example, one could interpolate various handcrafted parameter sets to handle continuous input scales — e.g., to generate an utterance that is 75% extravert — however the various weights can only be learned reliably from data.

Our results also suggest that SNLG systems may still require handcrafting to guarantee the naturalness of the output, but that data-driven models can be used to control various pragmatic effects. As naturalness depends on the combination of multiple generation decisions, modeling naturalness in the PE framework would require taking inter-parameter dependencies into account. At training time, this can be achieved by extending each PE model with independent variables for naturalness as well as all previous generation decisions made. During the generation process, the models would predict future generation decisions given all previous decisions and a high input naturalness value. While we believe this approach is tractable, it optimizes parameters sequentially in a greedy fashion. A global optimization method would require jointly optimizing all (past and future) parameters simultaneously. However, predictions in such a large output space would require a much larger training set. Nevertheless, we believe that future work should evaluate methods for jointly optimizing multiple NLG decisions.

Finally, this work focuses on controlling the perceptions of the output of a dialogue system, but an important next step is to use these techniques together with personality-based user modeling methods (Mairesse et al. 2007), to simultaneously model the personality of the user and the system in dialogue. Both models provide the tools for testing various hypotheses regarding personality-based alignment, such as the similarity-attraction effect (Isard, Brockmann, and Oberlander 2006). Furthermore, the optimal personality of the system is likely to be application-dependent, it would thus be useful to evaluate how the user's and the system's personality affect task performance in different applications.

The PERSONAGE language generator is available for download at `http://mi.eng.cam.ac.uk/~farm2/personage`, as well as the personality-annotated corpus collected for our experiments. An online demonstrator and a tutorial for customizing PERSONAGE for a new domain can be found at `http://nlds.soe.ucsc.edu/software/personage`.

Acknowledgments

This work was funded by a Royal Society Wolfson Research Merit Fellowship to Marilyn Walker and a Vice Chancellor's studentship to François Mairesse. Thanks to Chris Mellish, Donia Scott, Rob Gaizauskas and Roger Moore for detailed comments on earlier versions of this work.

References

André, Elisabeth, Thomas Rist, Susanne van Mulken, Martin Klesen, and Stephan Baldes. 2000. The automated design of believable dialogues for animated presentation teams. In J. Sullivan S. Prevost J. Cassell and E. Churchill, editors, *Embodied conversational agents*. MIT Press, Cambridge, MA, pages 220–255.

Bangalore, Srinivas and Owen Rambow. 2000. Exploiting a probabilistic hierarchical model for generation. In *Proceedings of the 18th International Conference on*

Computational Linguistics (COLING), pages 42–48, Saarbrücken.
Barzilay, Regina and Lillian Lee. Bootstrapping lexical choice via multiple-sequence alignment. In *Proceedings of EMNLP*, pages 164–171, Philadelphia, PA.
Belz, Anja. 2005. Corpus-driven generation of weather forecasts. In *Proceedings of the 3rd Corpus Linguistics Conference*, Birmingham.
Belz, Anja. 2008. Automatic generation of weather forecast texts using comprehensive probabilistic generation-space models. *Natural Language Engineering*, 14(4):431–455.
Biber, Douglas. 1988. *Variation across Speech and Writing*. Cambridge University Press.
Bouayad-Agha, Nadjet, Donia R. Scott, and Richard Power. 2000. Integrating content and style in documents: A case study of patient information leaflets. *Information Design Journal*, 9(2-3):161–176.
Brennan, Susan E. and Herbert H. Clark. 1996. Conceptual pacts and lexical choice in conversation. *Journal of Experimental Psychology: Learning, Memory And Cognition*, 22:1482–1493.
Cassell, Justine and Timothy Bickmore. 2003. Negotiated collusion: Modeling social language and its relationship effects in intelligent agents. *User Modeling and User-Adapted Interaction*, 13:89–132.
Chambers, Nathanael and James Allen. 2004. Stochastic language generation in a dialogue system: Toward a domain independent generator. In *Proceedings 5th SIGdial Workshop on Discourse and Dialogue*, pages 9–18, Cambridge, MA.
Chklovski, Timothy and Patrick Pantel. 2004. VERBOCEAN: Mining the web for fine-grained semantic verb relations. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 33–40, Barcelona.
Clark, Herbert H. and Susan E. Brennan. 1991. Perspectives on socially shared cognition. In L. B. Resnick, J. Levine, and S. D. Bahrend, editors, *Grounding in communication*. APA.
Coltheart, Max. 1981. The MRC psycholinguistic database. *Quarterly Journal of Experimental Psychology*, 33A:497–505.
Costa, Paul T. and Robert R. McCrae, 1992. *NEO PI-R Professional Manual*. Psychological Assessment Resources, Odessa, FL.
Daumé III, Hal. 2007. Frustratingly easy domain adaptation. In *Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 256–263, Prague.
Dunbar, Robin. 1996. *Grooming, Gossip, and the Evolution of Language*. Harvard University Press.
Fellbaum, Christiane. 1998. *WordNet: An Electronic Lexical Database*. MIT Press, Cambridge, MA.
Funder, David C. 1997. *The Personality Puzzle*. W. W. Norton & Company, New York, 2nd edition.
Furnham, Adrian. 1990. Language and personality. In H. Giles and W. Robinson, editors, *Handbook of Language and Social Psychology*. Winley.
Goffman, Erving. 1970. *Strategic Interaction*. Blackwell, Oxford.
Goldberg, Lewis R. 1990. An alternative "description of personality": The Big-Five factor structure. *Journal of Personality and Social Psychology*, 59:1216–1229.
Gosling, Samuel D., Sam Gaddis, and Simine Vazire. 2007. Personality impressions based on facebook profiles. In *Proceedings of the International Conference on Weblogs and Social Media*, Boulder, CO.
Gosling, Samuel D., Peter J. Rentfrow, and William B. Swann. 2003. A very brief measure of the big five personality domains. *Journal of Research in Personality*, 37:504–528.
Green, Stephen J. and Chrysanne DiMarco. 1996. Stylistic decision-making in natural language generation. *Lecture Notes in Computer Science, Trends in Natural Language Generation An Artificial Intelligence Perspective*, 1036:125–143.
Higashinaka, Ryuichiro, Marilyn A. Walker, and Rashmi Prasad. 2007. An unsupervised method for learning generation lexicons for spoken dialogue systems by mining user reviews. *ACM Transactions on Speech and Language Processing*, 4(4):8.
Hovy, Eduard. 1988. *Generating Natural Language under Pragmatic Constraints*. Lawrence Erlbaum Associates.
Inkpen, Diana Zaiu and Graeme Hirst. 2004. Near-synonym choice in natural language generation. In Nicolas Nicolov, Kalina Bontcheva, Galia Angelova, and Ruslan Mitkov, editors, *Recent Advances in Natural Language Processing III*, pages 141–152. John Benjamins Publishing Company.
Isard, Amy, Carsten Brockmann, and Jon Oberlander. 2006. Individuality and alignment in generated dialogues. In

Proceedings of the 4th International Natural Language Generation Conference (INLG), pages 22–29, Sydney.

John, Oliver P., Eileen M. Donahue, and Robert L. Kentle. 1991. The "Big Five" Inventory: Versions 4a and 5b. Technical report, Berkeley: University of California, Institute of Personality and Social Research.

Jordan, Pamela W. 2000. *Intentional Influences on Object Redescriptions in Dialogue: Evidence from an Empirical Study*. Ph.D. thesis, Intelligent Systems Program, University of Pittsburgh.

Kittredge, Richard, Tanya Korelsky, and Owen Rambow. 1991. On the need for domain communication knowledge. *Computational Intelligence*, 7(4):305–314.

Labov, William. 2006. *The Social Stratification of English in New York City*. Cambridge University Press.

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

Langkilde-Geary, Irene. 2002. An empirical verification of coverage and correctness for a general-purpose sentence generator. In *Proceedings of the International Conference on Natural Language Generation*, pages 17–24, Harriman, NY.

Lavoie, Benoit and Owen Rambow. 1997. A fast and portable realizer for text generation systems. In *Proceedings of the 3rd Conference on Applied Natural Language Processing*, pages 265–268, Washington, D.C.

Lester, James C., Brian Stone, and Gary Stelling. 1999. Lifelike pedagogical agents for mixed-initiative problem solving in constructivist learning environments. *User Modeling and User-Adapted Interaction*, 9(1-2):1–44.

Lester, James C., Stuart G. Towns, and Patrick J. Fitzgerald. 1999. Achieving affective impact: Visual emotive communication in lifelike pedagogical agents. *The International Journal of Artificial Intelligence in Education*, 10(3-4):278–291.

Levelt, Willem J. and Stephanie Kelter. 1982. Surface form and memory in question answering. *Cognitive Psychology*, 14(1):78–106.

Litman, Diane J. and Kate Forbes-Riley. 2004. Predicting student emotions in computer-human tutoring dialogues. In *Proceedings of the 42nd Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 352–359.

Litman, Diane J. and Kate Forbes-Riley. 2006. Recognizing student emotions and attitudes on the basis of utterances in spoken tutoring dialogues with both human and computer tutors. *Speech Communication*, 48(5):559–590.

Mairesse, François. 2008. *Learning to Adapt in Dialogue Systems: Data-driven Models for Personality Recognition and Generation*. Ph.D. thesis, Department of Computer Science, University of Sheffield.

Mairesse, François, 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 48th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 1552–1561, Uppsala.

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

Mairesse, François and Marilyn A. Walker. 2010. Towards personality-based user adaptation: Psychologically-informed stylistic language generation. *User Modeling and User-Adapted Interaction*, 20(3):227–278.

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

Marcus, David K., Scott O. Lilienfeld, John F. Edens, and Norman G. Poythress. 2006. Is antisocial personality disorder continuous or categorical? A taxometric analysis. *Psychological Medicine*, 36(11):1571–1582.

McQuiggan, Scott W., Bradford W. Mott, and James C. Lester. 2008. Modeling self-efficacy in intelligent tutoring systems: An inductive approach. *User Modeling and User-Adapted Interaction*, 18(1):81–123.

Mehl, Matthias R., Samuel D. Gosling, and James W. 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.

Nakatsu, Crystal and Michael White. 2006. 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.

Norman, Warren T. 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.

Paiva, Daniel S. and Roger Evans. 2004. A framework for stylistically controlled generation. In *Proceedings of the International Conference on Natural Language Generation*, pages 120–129, Sydney.

Paiva, Daniel S. 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)*, pages 58–65, Ann Arbor, MI.

Paris, Cécile and Donia R. Scott. 1994. Stylistic variation in multilingual instructions. In *Proceedings of the 7th International Workshop on Natural Language Generation*, pages 45–52, Kennebunkport, MN.

Pennebaker, James W., Martha E. Francis, and Roger J. Booth. 2001. *Inquiry and Word Count: LIWC 2001*. Lawrence Erlbaum Associates, Mahwah, NJ.

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

Pickering, Martin J. and Simon Garrod. 2004. Towards a mechanistic theory of dialogue. *Behavioral and Brain Sciences*, 27:169–225.

Piwek, Paul. 2003. 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)*, pages 151–154, Budapest.

Pollack, Martha E. 1991. Overloading intentions for efficient practical reasoning. *Noûs*, 25:513–536.

Porayska-Pomsta, Kaska and Chris Mellish. 2004. Modelling politeness in natural language generation. In *Proceedings of the International Conference on Natural Language Generation*, pages 141–150, Sydney.

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

Rambow, Owen, Monica Rogati, and Marilyn A. Walker. 2001. 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)*, pages 434–441, Toulouse.

Reiter, Ehud and Robert Dale. 2000. *Building Natural Language Generation Systems*. Cambridge University Press.

Scherer, Klaus R. 1979. Personality markers in speech. In K. R. Scherer and H. Giles, editors, *Social markers in speech*. Cambridge University Press, pages 147–209.

Scott, Donia R. and Clarisse Sieckenius de Souza. 1990. Getting the message across in RST-based text generation. In R. Dale, C. Mellish, and M. Zock, editors, *Current Research in Natural Language Generation*. Academic Press, London.

Singh, Satinder, Diane J. Litman, Michael Kearns, and Marilyn A. Walker. 2002. Optimizing dialogue managment with reinforcement learning: Experiments with the NJFun system. *Journal of Artificial Intelligence Research*, 16:105–133.

Snyder, Benjamin and Regina Barzilay. 2007. Database-text alignment via structured multilabel classification. In *Proceedings of the International Joint Conference on Artificial Intelligence (IJCAI)*, pages 1713–1718.

Srivastava, Sanjay, Steve Guglielmo, and Jennifer S. Beer. 2010. Perceiving others' personalities: Examining the dimensionality, assumed similarity to the self, and stability of perceiver effects. *Journal of Personality and Social Psychology*, 98(3):520–534.

Stent, Amanda and Hui Guo. 2005. A new data-driven approach for multimedia presentation generation. In *Proceedings of EuroIMSA*, Grindelwald.

Stone, Matthew and Bonnie Webber. 1998. Textual economy through close coupling of syntax and semantics. In *Proceedings of 1998 International Workshop on Natural Language Generation*, Niagara-on-the-Lake, Canada.

Tapus, Adriana and Maja Mataric. 2008. Socially assistive robots: The link between personality, empathy, physiological signals, and task performance. In *Proceedings of the AAAI Spring Symposium on Emotion, Personality and Social Behavior*, Palo Alto, CA.

Vogel, Klaus and Sigrid Vogel. 1986. L'interlangue et la personalité de l'apprenant. *International Journal of Applied Linguistics*, 24(1):48–68.

Walker, Marilyn A. 1993. *Informational Redundancy and Resource Bounds in Dialogue*. Ph.D. thesis, University of

Pennsylvania.
Walker, Marilyn A., Janet E. Cahn, and Stephen J. Whittaker. 1997. Improvising linguistic style: Social and affective bases for agent personality. In *Proceedings of the 1st Conference on Autonomous Agents*, pages 96–105, Marina Del Rey, CA.
Walker, Marilyn A., Jeanne C. Fromer, and Shrikanth Narayanan. 1998. Learning optimal dialogue strategies: A case study of a spoken dialogue agent for email. In *Proceedings of the 36th Annual Meeting of the Association for Computational Linguistics, COLING/ACL 98*, pages 1345–1352.
Walker, Marilyn A. and Owen Rambow. 2002. Spoken language generation. *Computer Speech and Language, Special Issue on Spoken Language Generation*, 16(3-4):273–281.
Walker, Marilyn A., Owen Rambow, and Monica Rogati. 2002. Training a sentence planner for spoken dialogue using boosting. *Computer Speech and Language*, 16(3-4).
Walker, Marilyn A., Amanda Stent, François Mairesse, and Rashmi Prasad. 2007. Individual and domain adaptation in sentence planning for dialogue. *Journal of Artificial Intelligence Research (JAIR)*, 30:413–456.
Wang, Ning, W. Lewis Johnson, Richard E. Mayer, Paola Rizzo, Erin Shaw, and Heather Collins. 2005. The politeness effect: Pedagogical agents and learning gains. *Frontiers in Artificial Intelligence and Applications*, 125:686–693.
White, Michael, Rajakrishnan Rajkumar, and Scott Martin. 2007. Towards broad coverage surface realization with CCG. In *Proceedings of the Workshop on Using Corpora for NLG: Language Generation and Machine Translation*, pages 22–30, Copenhagen.
Witten, Ian H. and Eibe Frank. 2005. *Data Mining: Practical machine learning tools and techniques*. Morgan Kaufmann, San Francisco, CA.