Towards Personality-based User Adaptation: Psychologically-informed Stylistic Language Generation

François Mairesse
Cambridge University Engineering Department, Cambridge CB2 1PZ, United Kingdom
f.mairesse@eng.cam.ac.uk*

Marilyn A. Walker
Department of Computer Science, University of California Santa Cruz,
Santa Cruz, CA. 95060, U.S.A.
lynwalker@gmail.com

Abstract

Conversation is an essential component of social behavior, one of the primary means by which humans express intentions, beliefs, emotions, attitudes and personality. Thus the development of systems to support natural conversational interaction has been a long term research goal. In this regard, a major focus of spoken language processing research is to reliably interpret naturally occurring variations in user utterances. These variations are a function of individual differences, and of the conversational context, including attempts by the user to adapt to the system's conversational style. However, the system's output is much less varied, typically consisting of a relatively small number of highly handcrafted utterances, designed to portray a particular system personality or linguistic style. This approach produces high quality utterances, but makes it difficult to dynamically adapt the system's linguistic style to individual users and specific tasks. As part of a personality-based user adaptation framework, this article describes PERSONAGE, a highly parameterizable generator which provides a large number of parameters to support adaptation to user's linguistic style. We show how we can use results from psycholinguistic studies that document the linguistic reflexes of personality, in order to develop models to control PERSONAGE's parameters, and produce utterances matching particular personality profiles. When we evaluate these outputs with human judges, the results indicate that humans perceive the personality of system utterances in the way that the system intended.

1 Introduction

Conversation is an essential component of social behavior, one of the primary means by which humans express intentions, beliefs, emotions, attitudes and personality. Thus systems to support natural conversational interaction have been a long term research goal (Carberry, 1989; Cohen et al., 1982; Finin et al., 1986; Grosz, 1983; Kobsa and Wahlster, 1989; Litman and Allen, 1984; Power, 1974; Zukerman and Litman, 2001). In this regard, a major focus of spoken language processing research is to reliably interpret naturally occurring variations in user utterances (Johnson et al., 2005; Louchart et al., 2005; Mateas and Stern, 2003; Pieraccini and Levin, 1995). These variations are a function of individual differences, and of the conversational context, including attempts by the user to adapt to the system's conversational style (Brennan, 1991; Darves and Oviatt, 2002). However, the system's output is typically much less varied, consisting of a relatively small number of highly handcrafted utterances, designed to portray a particular system personality or linguistic style. This approach produces high quality utterances, but it means that it is difficult to personalize the

*This paper or a similar version is not currently under review by a journal or conference, nor will it be submitted to such within the next three months. This paper is void of plagiarism or self-plagiarism as defined in Section 1 of ACM's Policy and Procedures on Plagiarism. This research was carried out at at the University of Sheffield, where the authors were supported under a Vice Chancellor's studentship and Walker's Royal Society Wolfson Research Merit Award.

style of the dialogue interaction to individual users, or for the system to adapt during the dialogue to the user or the context.

In natural conversation, humans adapt to one another across many levels of utterance production via processes variously described as entrainment, alignment, audience design, and accommodation (Brennan, 1996; Brennan and Clark, 1996; Giles et al., 1991; Levelt and Kelter, 1982; Nenkova et al., 2008; Niederhoffer and Pennebaker, 2002; Pickering and Garrod, 2003). A number of recent studies strongly suggest that dialogue systems that adapted to the user in a similar way would be more effective (André et al., 2000; Brennan, 1991; Cassell and Bickmore, 2003; Forbes-Riley and Litman, 2007; Forbes-Riley et al., 2008; Hayes-Roth and Brownston, 1994; Hirschberg, 2008; Mott and Lester, 2006; Murray, 1997; Reeves and Nass, 1996; Reitter et al., 2006; Stenchikova and Stent, 2007; Tapus and Mataric, 2008).

Remarkably, Reeves and Nass show that alignment is also beneficial at the personality level (1996). They ran a series of experiments using utterances which were handcrafted for a particular task and discourse context and which were designed intuitively to express a particular personality. Their experiments showed that users perceptions of a system's intelligence or competence increase when the systems adapts to the user's personality. Although this *similarity-attraction* effect alone provides motivation for exploring methods for personality-based user adaptation—as it is suggested to hold when averaged over a large set of contexts—there is a case for a more general framework in which the system's personality is dependent on both the task and the user. While in general introvert users might be more attracted to introvert systems (Reeves and Nass, 1996), correlates between personality and teaching performance suggest that they would learn more from an extravert system in the context of a tutoring system (Rushton et al., 1987). Additionally, Wang et al. (2005) find that the use of politeness in tutoring dialogues produces higher learning outcomes, with a stronger effect for high extraverts learning difficult concepts. The contextual dependency of the similarity-attraction effect is even more striking for other traits, e.g. it is hard to imagine that a neurotic user would benefit from a system projecting neurotic cues (if not for the treatment of psycho-pathologies). Given the large number of task-specific requirements for adaptive behaviour such as those presented in Table 1, Fig. 1 illustrates how a generic adaptation capability for dialogue systems requires addressing three research problems: (a) acquiring relevant user traits (recognition), (b) deciding what traits should be conveyed by the system (adaptation), and (c) producing a consistent response matching those traits (generation). While the latter task is the focus of this article, let us briefly discuss the first two.

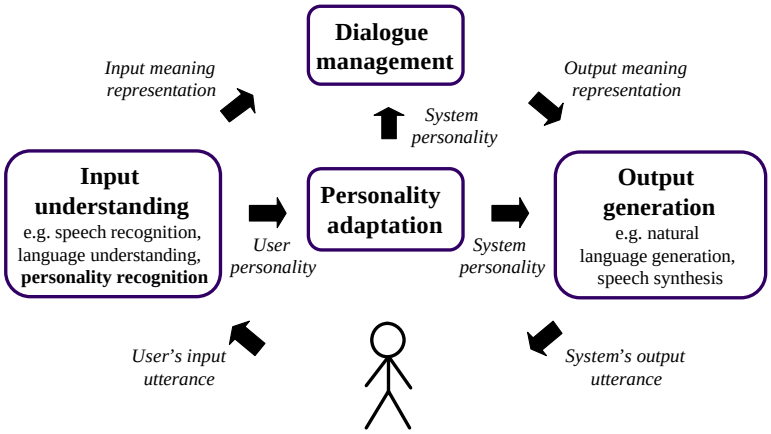


Fig. 1: High-level architecture of a dialogue system with personality-based user adaptation.

How can we acquire user personality information and why do we think it is an important component of user adaptation? The personality of the user can typically be assessed either by using questionnaires (Costa and McCrae, 1992; Gosling et al., 2003; John et al., 1991), or by identifying relevant behavioural cues, e.g. based on the user's interaction (Dunn et al., 2009) or the user's speech and language (Argamon et al., 2005; Mairesse et al., 2007; Oberlander and Nowson, 2006). While personality questionnaires have a high predictive value and only need to be filled once by the user, they lack the objectivity of observer-reports and require a significant effort from the user. For these reasons, we recently focused on automated methods for personality

recognition from user conversations (Mairesse et al., 2007). Results show that personality recognition models trained on content analysis and prosodic features predict the speaker's level of extraversion, neuroticism, agreeableness and conscientiousness better than chance. Similar methods have also been applied successfully to textual content (Argamon et al., 2005; Oberlander and Nowson, 2006). Automated recognition methods thus provide a promising alternative to questionnaires, however future work should evaluate whether they are accurate enough to improve task performance and user satisfaction within a task-oriented dialogue.

Table 1: Hypothesized personality adaptation policies for various applications. Specific traits are mapped to the Big Five or PEN framework (Eysenck et al., 1985; Norman, 1963).

Task	**User personality**	**System adaptation policy**
Information presentation system	novice	extravert, agreeable
	experienced	converge towards user
Tutoring system	any	extravert, agreeable, conscientious
Telesales system	any	potent (extravert), match the company's brand
Video games/entertainment	any	character-based
Crucial information retrieval (e.g. finance)	any	conscientious, not extravert, not agreeable
Psychotherapy	fearful (introvert, not open)	aggressive (extravert, psychotic)
	aggressive (extravert, psychotic)	fearful or aggressive
Psychotherapist training system	any	neurotic
	any	aggressive (extravert, psychotic)

Once the personality of the user has been assessed, findings from psychological studies can inform the *personality adaptation model* illustrated in Fig. 1. Table 1 presents several personality adaptation policies for different tasks, which remain to be evaluated. For example, we hypothesize that information presentation systems should produce extravert and agreeable language when facing novice users, and back-off to a similarity-attraction policy with more advanced users. Previous studies suggest that tutoring systems should be agreeable and extravert, based on findings associating the use of politeness forms with higher learning outcomes (Wang et al., 2005), as well as correlates between extraversion and human teachers performance (Rushton et al., 1987). Additionally, we hypothesize that systems providing crucial information—e.g. when requesting stock quotes or emergency advice—should produce outputs that are clear and concise. This suggests the need for a conscientious, introvert and non-agreeable operator, e.g. avoiding superfluous politeness forms. More generally, training systems should convey a large range of personalities. Examples include systems for training practitioners to interview anxious patients (Hubal et al., 2000), as well as systems training soldiers to gather information from uncooperative civilians through tactical questioning (Department of the Army, 2006). Personality modelling has also found applications in virtual reality for psychotherapy, e.g. to reduce the patient's anxiety when interacting with aggressive personalities or when speaking in public (Slater et al., 2004).

Additionally, task-dependent adaptation can also be beneficial to the system. Furnham et al. (1999) report that potency (extraversion) correlates positively with sales figures and superior ratings, and that impulsivity is a significant performance predictor of telesales employees selling insurance. Such findings can guide system designers to optimise the personality conveyed by an automated sales agent. Additionally, a large body of marketing research shows that consumers associate brands with personality types, and that they tend to select brands conveying traits that are desirable to them (Aaker, 1999; Fennis and Pruyn, 2007; Plummer, 1984). There is thus a strong incentive for companies to tailor the personality of their dialogue system to their target market.

Given the large number of applications benefiting from the projection of specific traits, what is needed is a computational method for generating dialogic utterances, which provides a wide selection of relevant parameters, whose values can be changed in real time in order to adapt to the user. If dialogue system utterances are handcrafted to portray a particular style, such parameters are not available, nor can they be modified in real time to adapt to a particular user.

In this article, we describe a highly parameterizable generator PERSONAGE, which provides a large

number of parameters to support adaptation to a user's linguistic style. These parameters operate across many levels of linguistic production and can support adaptation of content selection, lexical choice, and selection of syntactic and rhetorical structure. One way to control all these parameters is via a model of the user, representing either the total interaction history, or simply the current dialogue context (Kobsa and Wahlster, 1989). In this paper, we show how we can use personality models based on psycholinguistic studies to control PERSONAGE's parameters, and produce utterances adapted to particular user personality profiles. When we evaluate these outputs with human judges, the results indicate that humans perceive the personality of system utterances in the way that the system intended.

Our framework builds on the "Big Five" model of personality traits. The Big Five model is based on the observation that, when talking about a close friend, one can usually produce a large number of descriptive adjectives (Allport and Odbert, 1936). This observation is described as the *Lexical Hypothesis*, i.e. that any trait important for describing human behavior has a corresponding lexical token, which is typically an adjective, such as *trustworthy, modest, friendly, spontaneous, talkative, dutiful, anxious, impulsive, vulnerable.* The Lexical Hypothesis led to a great deal of subsequent work, and to a consensus that there are essential traits, known as the *Big Five* personality traits (Goldberg, 1990; Norman, 1963; Peabody and Goldberg, 1989). These traits (see Table 2) are *extraversion*, *emotional stability*, *agreeableness*, *conscientiousness* and *openness to experience*.[1]

Table 2: Example adjectives associated with the 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

The Big Five model has several advantages as the basis of a computational framework for generating variation in linguistic style. There are a large number of useful prior studies (Mehl et al., 2006; Oberlander and Gill, 2006; Pennebaker and King, 1999; Thorne, 1987), that carefully document correlations between Big Five traits and linguistic behavior (measured via lexical category, word or syntactic structure counts). These correlations suggest a large number of relevant parameters for generation. One important contribution of this paper is our survey of these studies, and our proposals for generation parameters that can affect these lexical category, word or syntactic structure counts. Another advantage is that prior work on the Big Five model uses validated personality surveys to assess personality traits in humans (e.g. Gosling et al., 2003; John and Srivastava, 1999; McCrae and Costa, 1987). Rather than inventing our own assessment methods for potentially ill-defined stylistic variations, we use these same surveys to evaluate our computational model of personality generation, and verify that the personality we intend to project is perceived correctly.

Table 3 shows some example outputs of PERSONAGE. In Table 3 the **set** column indicates whether the output utterance was based on a personality model for the low end of a trait (introversion) vs. the high end of a trait (extraversion). The examples in Table 3 manipulate parameters such as verbosity (verbal fluency, Study 7 in Table 25), polarity of the content selected (polarity, Study 14 in Table 25), and the occurrence of hedges or markers of tentativeness (content analysis category counts, Studies 3, 5, 6 in Table 25). The

[1]These traits explain the most variance of behaviour among people, and there is evidence for their biological basis (Eysenck et al., 1985; Revelle, 1991), e.g. studies comparing the personality of monozygotic twins with dizygotic twins raised apart show that the percentage of the variance accounted for by genetic factors—i.e. the heritability—is approximately 50% for each Big Five trait (Bouchard and McGue, 2003).

score column of Table 3 shows the average score of three judges when asked to assess the personality of the speaker of the utterance using the Ten Item Personality Inventory of Gosling et al. (2003). We explain in detail in this paper how we generate such utterances and how we collect human judgments to evaluate our framework.

Table 3: Example outputs of PERSONAGE for extraversion and emotional stability traits, with average judges ratings on the corresponding personality dimension (see Section 4). Personality ratings are on a scale from 1 to 7, with 1 = very low (e.g. introvert) and 7 = very high (e.g. extravert).

Trait	**Set**	**Example output utterance**	**Score**
Extraversion	low	Chimichurri Grill isn't as bad as the others.	1.00
	high	I am sure you would like Chimichurri Grill, you know. The food is kind of good, the food is tasty, it has nice servers, it's in Midtown West and it's a Latin American place. Its price is around 41 dollars, even if the atmosphere is poor.	6.33
Emotional stability	low	I am not sure! I mean, Ch-Chimichurri Grill is the only place I would recommend. It's a Latin American place. Err... its price is... it's damn ex-expensive, but it pr-pr-provides like, adequate food, though. It offers bad atmosphere, even if it features nice waiters.	4.00
	high	Let's see what we can find on Chimichurri Grill. Basically, it's the best.	6.00

Previous research on the generation of linguistic variation includes both rule-based and statistical approaches, as well as hybrid methods that combine rule-based linguistic knowledge with statistical methods (Langkilde and Knight, 1998; Langkilde-Geary, 2002). This includes work on variation using parameters based on pragmatic effects (Fleischman and Hovy, 2002; Hovy, 1988), stylistic factors such as formality, sentence length, and syntactic structure (Belz, 2005b; Bouayad-Agha et al., 2000; DiMarco and Hirst, 1993; Green and DiMarco, 1993; Paiva and Evans, 2005; Paris and Scott, 1994; Power et al., 2003; Walker et al., 2002), emotion (Cahn, 1990), lexical choice (Inkpen and Hirst, 2004), user expertise or confidence (DiMarco and Hirst, 1993; Forbes-Riley and Litman, 2007; Forbes-Riley et al., 2008; Porayska-Pomsta and Mellish, 2004; Wang et al., 2005), Brown and Levinson's theory of linguistic politeness (Brown and Levinson, 1987; Gupta et al., 2007, 2008; Porayska-Pomsta and Mellish, 2004; Walker et al., 1997a; Wang et al., 2005; Wilkie et al., 2005), theories of personality (André et al., 2000; Ball and Breese, 1998; Isard et al., 2006; Loyall and Bates, 1995), and individual differences and preferences for both style and content (Belz, 2005a; Lin, 2006; Reiter and Sripada, 2002; Stent et al., 2004; Walker et al., 2007). While there are strong relations between these different notions of style, and the types of linguistic variation associated with personality factors, here we limit our detailed discussion of prior work to personality generation. In Section 6, we will discuss how, in future work, PERSONAGE could be used to generate different types of stylistic variation.

Previous work on personality generation has primarily been associated with embodied conversational agents (ECAs). This research is very useful for identifying applications of personality generation, and showing how to integrate personality generation at the textual level with other modalities such as gesture and prosody (Rehm and Andre, 2008). While we do not know of any studies using ECAs that evaluate whether the personality of generated utterances is perceived by human users as the ECA intended, some of this work has shown an effect on task-related metrics, such as user satisfaction or perceptions of system competence (Isbister and Nass, 2000; Reeves and Nass, 1996). In contrast to our approach, the work on ECAs has typically modelled the generation task using templates, which have been labelled as expressing a particular personality, rather than by manipulating parameters within modules of the NLG pipeline.

Loyall and Bates (1995) is one of the first papers to suggest the use of personality models for language generation in ECAs. They present a model where personality factors are integrated with emotions, intentions and desires, and use template-based generation indexed by personality variables. Ball and Breese (1998) model the effect of the agent's personality (i.e. dominance and friendliness) and emotions (i.e. valence and arousal) on its behaviour. The personality values affect a layer of variables determining the paraphrase template to be selected by the system, such as the language strength, positivity and terseness. Scripted

dialogue is another venue for modelling the personality of multiple conversational agents. André et al. (2000) provide a system where the agents' utterances can be modified by selecting different values for extraversion, agreeableness and openness to experience (André et al., 2000; Rehm and Andre, 2008). Templates are annotated with intermediary variables (e.g. force) which in turn are associated with the personality traits (e.g. extravert agents use more forceful language, and they show more initiative in dialogue), and with gesture and facial expression. Lester et al. (1999a; 1999b) use handcrafted models of personality and emotion in pedagogical applications to teach children about science, and suggest that children become much more engaged in learning when the pedagogical agents exhibit colorful personalities and express emotions. Cassell and Bickmore (2003) extend their REA real estate agent with smalltalk generation capabilities, which is hypothesized to increase the user's trust in the system. Interestingly, they observe large perceptual variations between user groups with different personalities. Extravert users feel that they know REA better if she produces social language, resulting in a more satisfying interaction. On the other hand, introvert users are much less affected by REA's smalltalk, and rate that version of REA lower.

The most closely related work to this paper is the CRAG-2 system, which extends HALOGEN's methodology (Langkilde-Geary, 2002) to model personality and alignment in dialogue (Brockmann, 2009; Isard et al., 2006). CRAG-2 ranks a set of candidate utterances based on a linear combination of n-gram models, including a general-domain model trained on conversations from the Switchboard corpus, and models trained on a corpus of weblogs labeled with the author's personality. The system models linguistic alignment using a cache language model that primes particular syntactic forms on the basis of the conversational partner's previous utterance. This work is the first to combine personality control and alignment within the same framework. Brockmann (2009) shows that the alignment model affects personality perceptions of agreeableness and reduces the overall interaction quality, thus illustrating the trade-off between (a) benefits of the similarity attraction effect and (b) task-dependent personality requirements such as those presented in Table 1. The main difference between the current work and the CRAG-2 system lies in the generation methodology being used. While CRAG-2 produces variation at the realisation level based on heuristic rules, this article puts forward a systematic framework consisting of a large number of well-defined linguistic parameters that can be used to generate language manifesting personality at all levels of the generation process, thus ensuring re-usability and avoiding the need for an overgeneration phase.

In order to identify such parameters, we systematically organize and utilize findings from the psycholinguistic literature to develop personality models (see Table 25 in the Appendix). We then use these models to control the generation of system utterances, and show them to be perceived by the user as the system intended. The generation parameters that we propose are well-specified in terms of generation decisions that manipulate well-defined syntactic and semantic representations used in many standard NLG architectures. Thus they could be easily implemented in other generators and in other domains, and provide a basis for systematically testing, across domains and applications, which types of stylistic variation affect user perceptions and how.

Section 2 describes the PERSONAGE base generator and all of its parameters. Section 3 presents a method for controlling these parameters, by developing personality models that target each end of the extraversion and emotional stability scales. Personality models for the remaining Big Five traits are detailed by Mairesse (2008). In order to test our framework, we instantiate it in a particular discourse situation and domain, namely producing recommendations in the restaurant domain. Sections 4 and 5 describe our evaluation experiment and present the results. Our evaluation metric is based on a standard personality measurement instrument (Gosling et al., 2003). Section 5.1 reports results showing that the judges agree significantly on their perceptions. Section 5.2 presents the correlations between PERSONAGE's linguistic parameters and extraversion and emotional stability ratings in order to test precisely which parameters are affecting user perceptions, and to test whether findings from previous work generalize to our domain and discourse situation. Results show that linguistic reflexes documented in naturally occurring genres can be manipulated in a language generator and that those reflexes in many cases have the same effect on perception of personality. Section 5.3 reports our evaluation of the naturalness of the generated utterances. We sum up and discuss future work in Section 6.

2 The Personage Base Generator

Personage builds on the SPaRKy sentence planner (Stent et al., 2004; Walker et al., 2007), which produces comparisons and recommendations of restaurants in New York City. We hypothesized that evaluative utterances such as recommendations, which are often more effective when personalized (Ardissono et al., 2003; Carenini and Moore, 2000), are well suited for expressing recognizable personality, because they allow for substantial variation of the utterance length, polarity and subjectivity of the opinion expressed, which correlate with various personality traits (See Section 3).

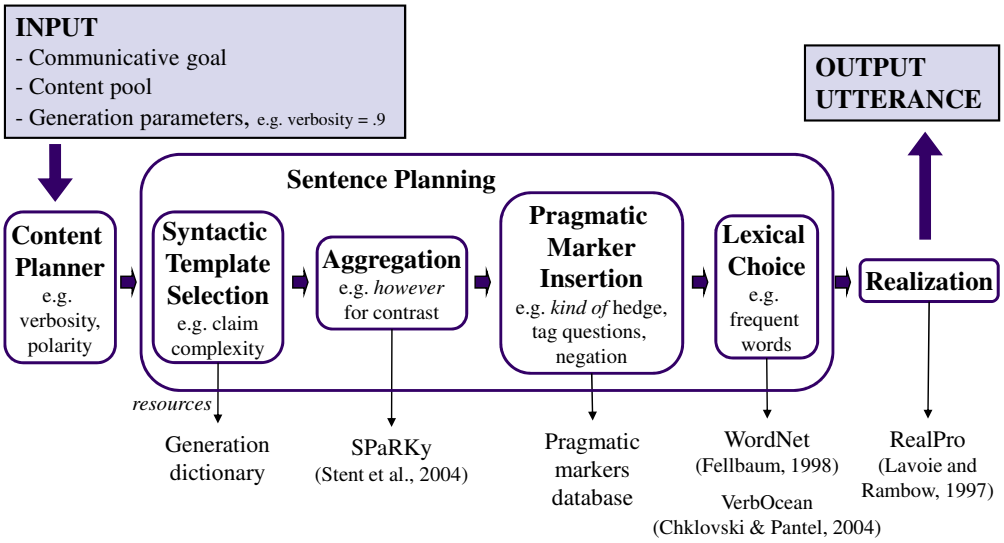


Fig. 2: The architecture of the Personage base generator.

Fig. 2 specifies Personage's architecture and gives examples of parameters introduced in each module in order to produce and control linguistic variation. The inputs are (1) a content plan representing a high-level communicative goal (speech act); (2) a content pool that can be used to achieve that goal, and (3) a set of parameter values for the generation parameters that we define below. In a dialogue system, the content plan is provided by the dialogue manager. Fig. 2 also shows how Personage uses multiple general, domain-independent, online lexical resources, such as WordNet and VerbOcean (Chklovski and Pantel, 2004; Fellbaum, 1998). It uses the standard NLG pipeline architecture (Kittredge et al., 1991; Reiter and Dale, 2000; Walker and Rambow, 2002; Walker et al., 2007). We know of no other work that exploits the modular nature of the standard NLG architecture to target personality-based variation.

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*.[2] The first component is the *content planner* which specifies the structure of the information to be conveyed (Section 2.1). The resulting content plan tree is then processed by the *sentence planner*, which selects syntactic structural templates for expressing individual propositions (Section 2.2), and aggregates them to produce the utterance's full syntactic structure (Section 2.3). The pragmatic marker insertion component then modifies the syntactic structure locally to produce various pragmatic effects, depending on the markers' insertion constraints (Section 2.4). The lexical choice component selects the most appropriate lexeme for each content word, given the lexical selection parameters (Section 2.5). Finally, the RealPro 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 (Section 2.6).[3] To make

[2]The attribute values used in the present work are derived from Zagat Survey's ratings, and mapped from a 30-point scale to the $[0, 1]$ interval.

[3]In a typical dialogue system, the output of the realizer is annotated for prosodic information by the prosody assigner, before being sent to the text-to-speech engine to be converted into an acoustic signal. Personage does not currently express personality through prosody, although studies of how personality is expressed in speech (Scherer, 1979) could be used to develop

PERSONAGE domain-independent, the input parameter values are normalized between 0 and 1 for continuous parameters, and to 0 or 1 for binary parameters, e.g. a VERBOSITY parameter of 1 maximizes the utterance's verbosity given the input, regardless of the actual number of propositions expressed.

2.1 Content planning

The input to the generation process is a *content plan*, a high level structure reflecting the communicative goal of the utterance. The content plan combines together propositions expressing information about individual attributes using *rhetorical relations* from Rhetorical Structure Theory, as in other generators (Mann and Thompson, 1988; Marcu, 1996; Moore and Paris, 1993; Scott and Souza, 1990). Two types of communicative goals are supported in PERSONAGE: *recommendation* and *comparison* of restaurants. Fig. 3 shows an example content plan for a recommendation.

Relations:	JUSTIFY (N:1, S:2); JUSTIFY (N:1, S:3); JUSTIFY (N:1, S:4); JUSTIFY (N:1, S:5); JUSTIFY (N:1, S:6); JUSTIFY (N:1, S:7)
Content:	1. assert(best (*Chanpen Thai*)) 2. assert(is (*Chanpen Thai*, cuisine (*Thai*))) 3. assert(has (*Chanpen Thai*, food-quality (.8))) 4. assert(has (*Chanpen Thai*, atmosphere (.6))) 5. assert(has (*Chanpen Thai*, service (.8))) 6. assert(is (*Chanpen Thai*, price (*24 dollars*))) 7. assert(is (*Chanpen Thai*, location (*Midtown West*)))

Fig. 3: An example content plan for a recommendation. N = nucleus, S = satellite.

The content plan is automatically converted into an equivalent tree structure, as illustrated in Fig. 4. This tree structure is referred to as the *content plan tree*. Each recommendation content plan contains a claim (nucleus) about the overall quality of the selected restaurant(s), supported by a set of satellite propositions describing their attributes. The propositions—the leaves in the content plan tree—are assertions labelled *assert-attribute(selection name)* in Fig. 4. Claims can be expressed in different ways, such as RESTAURANT NAME is the best, while the attribute satellites follow the pattern RESTAURANT NAME has MODIFIER ATTRIBUTE NAME, as in Le Marais has good food. Recommendations are characterized by a JUSTIFY rhetorical relation associating the claim with all the other propositions, which are linked together through an INFER relation.[4] In comparisons, the attributes of multiple restaurants are compared using the CONTRAST rhetorical relation. This relation combines propositions describing the same attributes for different restaurants, joined together through an INFER relation. An example content plan tree for a comparison between two restaurants is in Fig. 5.

Twelve content planning parameters are shown in Table 4 and discussed below. These parameters influence the size of the content plan tree, the content ordering, the rhetorical relations used, and the polarity of the propositions expressed. The correlational studies discussed in Section 3 suggests potential relationships between personality traits and a number of generation decisions at the content plan level.

Content size: Certain personality types tend to be more verbose, e.g. extraverts are more talkative than introverts (Furnham, 1990; Pennebaker and King, 1999). However because this finding is simply based on word count, it is not clear whether this involves the production of more content, or just being redundant and wordy. Thus we developed a number of parameters for our base generator that relate to the amount and type of content produced.

The VERBOSITY parameter controls the number of propositions selected from the content plan. The parameter value defines the ratio of propositions that are kept in the final content plan tree, while satisfying constraints dependent on the communicative goal: a recommendation must include a claim, and a comparison must include a pair of contrasted propositions. For example, the low extraversion utterance in Table 3 has a low VERBOSITY value, while the high extraversion utterance has high VERBOSITY, and expresses most of the items in the content plan.

such parameters for PERSONAGE.

[4]The INFER relation is similar to the JOINT relation in the RST literature.

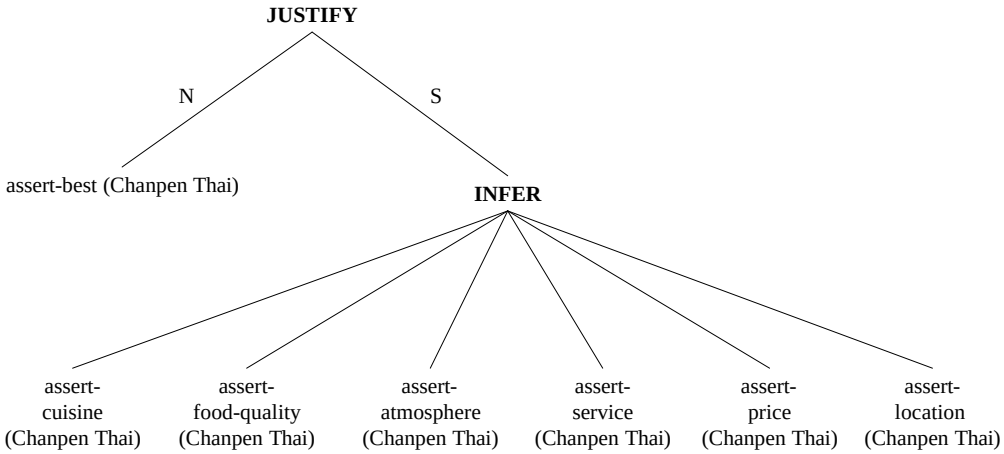


Fig. 4: An example content plan tree for a recommendation for Chanpen Thai, using all the restaurant attributes. N = nucleus, S = satellite.

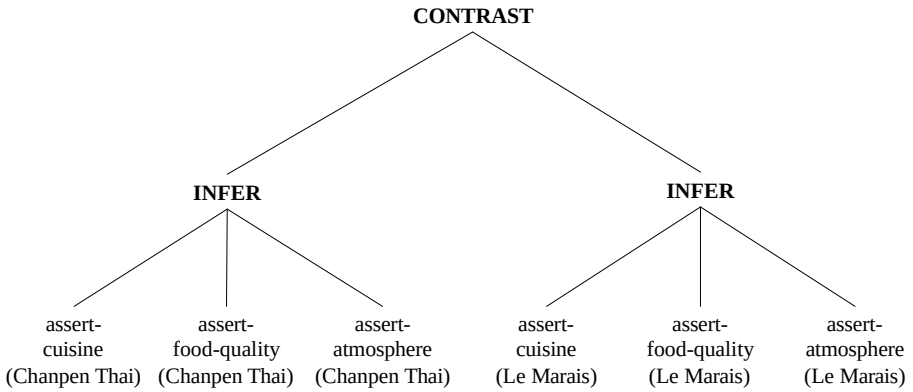


Fig. 5: An example content plan tree for a comparison between Chanpen Thai and Le Marais, using three attributes. All relations are multinuclear.

Table 4: Content Planning Parameters

Parameters	**Description**
VERBOSITY	Control the number of propositions in the utterance
RESTATEMENTS	Paraphrase an existing proposition, e.g. *'X has great Y, it has fantastic Z'*
REPETITIONS	Repeat an existing proposition
CONTENT POLARITY	Control the polarity of the propositions expressed, i.e. referring to negative or positive attributes
REPETITION POLARITY	Control the polarity of the restated propositions
CONCESSIONS	Emphasize one attribute over another, e.g. *'even if X has great Z, it has bad Y'*
CONCESSION POLARITY	Determine whether positive or negative attributes are emphasized
POLARIZATION	Control whether the expressed polarity is neutral or extreme
POSITIVE CONTENT FIRST	Determine whether positive propositions are uttered first
REQUEST CONFIRMATION	Begin the utterance with a confirmation of the request, e.g. *'did you say X?'*
INITIAL REJECTION	Begin the utterance with a rejection, e.g. *'I'm not sure'*
COMPETENCE MITIGATION	Express the speaker's negative appraisal of the hearer's request, e.g. *'everybody knows that . . . '*

The REPETITION parameter adds an exact repetition: the proposition node is duplicated and linked to the original content by a RESTATE rhetorical relation. The continuous parameter value (between 0 and

1) is mapped linearly to the number of repetitions in the content plan tree, i.e. between 0 and a domain-specific maximum (set to 2 in our domain). In Table 15, utterance 6 contains a repetition for the food quality attribute. The RESTATEMENT parameter adds a paraphrase to the content plan, obtained from the generation dictionary (see Section 2.2). If no paraphrase is found, one is created automatically by substituting content words with the most frequent WordNet synonym (see Section 2.5).

Polarity: We define parameters controlling polarity in order to bias the type of propositions that are selected to achieve the communicative goal, and to control whether positive or negative information is most salient in the utterance. This allows us to model findings that some personality types are more positive, e.g. they try to find something positive to say, while other traits tend to engage in more "problem talk" and expressions of dissatisfaction (Thorne, 1987). See Table 4 for definitions of the CONTENT POLARITY, REPETITION POLARITY, CONCESSIONS, CONCESSION POLARITY, and POLARIZATION parameters.

First, to support the CONTENT POLARITY parameter, propositions are defined as positive or negative. In our domain, propositions expressing attributes that received low ratings from users in Zagat surveys are defined as negative, although there are potentially many ways to assign positive and negative polarities to propositions (Fleischman and Hovy, 2002; Higashinaka et al., 2007; Wiebe, 1990). The claim in a recommendation is assigned a maximally positive polarity of 1, while the *cuisine* and *location* attributes are set at neutral polarity.[5] Then, the value of the CONTENT POLARITY parameter controls whether the content is mostly negative (e.g. *'Chanpen Thai has mediocre food'*), neutral (e.g. *'Le Marais is a French restaurant'*), or positive (e.g. *'Babbo has fantastic service'*). If there is enough polarized content given the required content plan tree size (VERBOSITY), the following propositions are selected depending on the value of the CONTENT POLARITY parameter:

Condition	Proposition set	
CONTENT POLARITY	$\in [0, .25[$	only negative propositions
	$\in [.25, .5[$	negative and neutral propositions
	$\in [.5, .75[$	neutral and positive propositions
	$\in [.75, 1]$	only positive propositions

If there are not enough propositions in the resulting set to satisfy the verbosity constraint, propositions with the closest polarity are added until the required content plan size is reached. Additionally, a constraint requiring that a comparison content plan tree contains at least one CONTRAST relation is enforced, thus the tree is likely to include propositions with different polarities.

From the filtered set of propositions, the POLARIZATION parameter determines whether the final content includes attributes with extreme scalar values (e.g. *'Chanpen Thai has fantastic staff'* vs. *'Chanpen Thai has decent staff'*). The REPETITIONS POLARITY parameters controls whether repetitions and paraphrases, if introduced, repeat and emphasize the positive content, or the negative content.

Rhetorical structure also affects the perceived polarity of an utterance, e.g. compare *'even if the food is good, it's expensive'* to *'even if the food is expensive, it's good'*. The CONCESSIONS parameter controls whether two propositions with different polarity are presented objectively, or if one is foregrounded and the other backgrounded. If two opposed propositions are selected for a concession, a CONCEDE relation is inserted between them.[6] The CONCESSION POLARITY parameter controls if the positive content is conceded (*'even if the food is good, it's expensive'*) or if the negative content is conceded *'even if the food is expensive, it's good'*).

Content ordering: Although extraverts use more positive language (Pennebaker and King, 1999; Thorne, 1987), an independent factor is where in the utterance the positive content is positioned. The position of the claim affects the persuasiveness of an argument (Carenini and Moore, 2000). The POSITIVE CONTENT FIRST parameter controls whether positive propositions—including the claim—appear first or last. [7]

The INITIAL REJECTION, REQUEST CONFIRMATION and COMPETENCE MITIGATION parameters are also defined in Table 4. From a theoretical perspective, these are content planning parameters, but since they

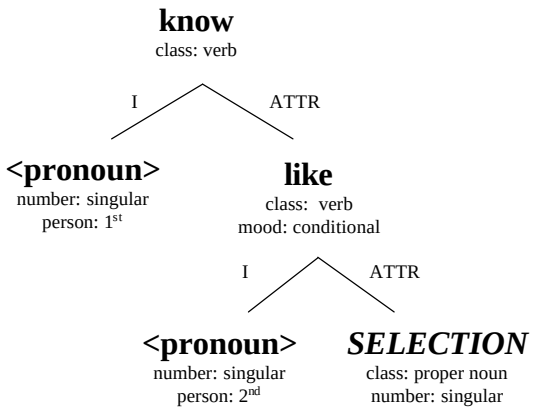
[5] An alternative would be to use individual user models to assign positive and negative polarities to categorical attributes as well (Ardissono et al., 2003; Carenini and Moore, 2006; Walker et al., 2004).

[6] We currently only can construct concessions between attributes of the same restaurant.

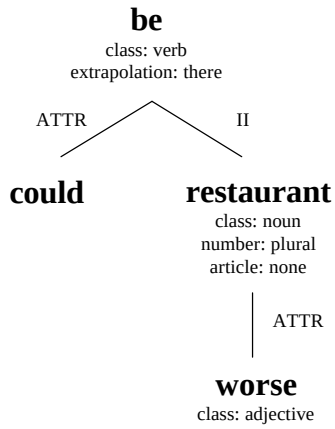
[7] Although this parameter determines the ordering of the nodes of content plan tree, some aggregation operations can still impose a specific ordering, e.g. BECAUSE CUE WORD to realize the JUSTIFY relation, see Section 2.3.

only affect the beginning of the utterance, we have implemented them in a later phase of utterance generation along with the insertion of pragmatic markers. We describe their implementation in Section 2.4.

2.2 Syntactic Structural Template Selection



(a) *I know you would like SELECTION*



(b) *There could be worse restaurants*

Fig. 6: Two example DSyntS for a recommendation claim. The lexemes are in bold, and the attributes below indicate non-default values in the RealPro realizer. Branch labels indicate dependency relations, i.e. I = subject, II = object and ATTR = modifier. Lexemes in italic are variables that are instantiated at generation time.

Once the content planner has determined *what* will be talked about, the remaining components control *how* the information is to be conveyed. The first phase of sentence planning looks in the generation dictionary for the set of syntactic elementary structures stored for each proposition in the content plan. PERSONAGE manipulates syntactic dependency tree representations inspired by Melčuk's Meaning-Text Theory (1988), and referred to as Deep Syntactic Structures (DSyntS), Fig. 6 shows two DSyntS expressing the recommendation claim. The DSyntS are stored in a small handcrafted generation dictionary, currently containing 18 DSyntS: 12 for the recommendation claim and one per attribute. Some attribute DSyntS contain variables that are instantiated based on the input restaurant (e.g. polarity adjectives in Section 2.5), see Fig. 7(a) for an example. These DSynts representations can be combined using domain-independent general-purpose linguistic operations to make more complex DSyntS (complex utterance structures) in order to produce a wide range of variation. The DSyntS can be converted to an output string using the RealPro surface realizer, which is also based on Melcuk's theory (Lavoie and Rambow, 1997). The DSyntS contain variables that are filled at generation time, such as the restaurant's name or cuisine. See Fig. 6.

Table 5: Syntactic Structural Template Selection Parameters

Parameters	Description
SYNTACTIC COMPLEXITY	Control the syntactic complexity (e.g. syntactic embedding)
SELF-REFERENCES	Control the number of first person pronouns
TEMPLATE POLARITY	Control the syntactic structure's connotation (positive or negative)

Table 5 shows the PERSONAGE parameters that control the selection of DSyntS from the generation dictionary. The DSyntS selection process first assigns each candidate DSyntS to a point in a three-dimensional space, characterizing the DSyntS' syntactic complexity, number of self-references and polarity. Parameter values are normalized over all candidate DSyntS, so the DSyntS closest to the target values can be computed.

Syntactic complexity: Furnham (1990) suggests that introverts produce more complex constructions: the SYNTACTIC COMPLEXITY parameter controls the number of subordinate clauses of the DSyntS chosen to represent the claim, based on Beaman's definition of syntactic complexity (1984).[8] For example, the claim in Fig. 6(a) is rated as more complex than the one in Fig. 6(b), because the latter has no subordinate clause.

Self-references: Extraverts and neurotics make more self-references (Pennebaker and King, 1999). The SELF-REFERENCES parameter controls whether the claim is made in the first person (based on the speaker's own experience), or whether the claim is reported as objective or information obtained elsewhere. The SELF-REFERENCES value is computed from the DSyntS by counting the number of first person pronouns. For example, the DSyntS in Fig. 6(a) contains one self-reference, while that in Fig. 6(b) does not.

Polarity: While polarity can be expressed by content selection and structure, it can also be directly associated with the DSyntS. The TEMPLATE POLARITY parameter determines whether the claim has a positive or negative connotation (Fleischman and Hovy, 2002; Hovy, 1988; Wiebe, 1990). While automated methods for opinion extraction could be used in the future to annotate the generation dictionary (Higashinaka et al., 2007; Hu and Liu, 2004; Pang et al., 2002; Riloff et al., 2005; Wiebe, 1990; Wilson et al., 2004), at present DSyntS are manually annotated for polarity. An example claim with low polarity can be found in Fig. 6(b), i.e. *'There could be worse restaurants'*, while the claim in Fig. 6(a) is rated more positively.

2.3 Aggregation

Previous work in psychology suggests that personality affects the aggregation process, e.g. introverts prefer complex syntactic constructions, long pauses and rich vocabulary (Furnham, 1990). The role of the aggregation component is to combine syntactic structures for elementary DSyntS into a larger syntactic structure, by associating each pair of sibling propositions in the content plan tree with a *clause-combining operation* that determines how the parent rhetorical relation is to be expressed. Table 6 shows the PERSONAGE parameters that control the selection of clause-combining operators in the sentence planner. For example, poor food quality can be contrasted with good atmosphere using cue words such as *however*, or *but*. As in Rambow et al. (2001), the aggregation process randomly selects a clause-combining operation, for each rhetorical relation in the content plan tree, according to the probability distribution for that relation defined by the input aggregation parameters. See for example the distributions for the INFER relation for extraversion in Table 7. Aggregation settings for other traits are based on the findings detailed in Section 3.

The aggregation process then randomly selects pairs of propositions among the children propositions, until the two associated DSyntS satisfy the constraints of the clause-combining operation, e.g. the MERGE operation requires that both argument DSyntS have the same main verb. If none of the pairs satisfy the constraints, another clause-combining operation is chosen according to the input probability distribution. The aggregation process is guaranteed to terminate as each rhetorical relation implements at least one clause-combining operation with no constraint on the DSyntS, i.e. the PERIOD operation, which keeps both argument DSyntS in separate sentences. PERSONAGE augments the SPARKY clause-combining operations (Stent et al., 2004; Walker et al., 2007), with additional operations for the RESTATE and CONCEDE rhetorical relations. Table 8 shows the available operations for each rhetorical relation, the constraints on their application, and the result. Their effect on the final utterance at an abstract level is summarized in Table 6.

2.4 Pragmatic marker insertion

Psychological studies identify many pragmatic markers of personality that affect the utterance locally, and can be implemented as context-independent syntactic transformations. Table 9 describes all of the pragmatic marker insertion parameters and provides examples. For example, parameters in Table 9 include negations, tentative/softening hedges (e.g. *maybe*, *kind of*) and filled pauses (Pennebaker and King, 1999; Scherer, 1979; Siegman and Pope, 1965), expletives, emphasizer hedges (e.g. *really*) and exclamation marks (Mehl et al., 2006; Oberlander and Gill, 2004b). There are FILLED PAUSES and STUTTERING markers because

[8] The syntactic complexity is computed as the number of verb nodes in the DSyntS, which is equivalent to the number of subordinate clauses in the final utterance.

[9] The input selection probability does not entirely reflect the probability that an operation will appear in the output utterance, as the latter is also dependent on the constraints the operation imposes on its DSyntS arguments. For example, the MERGE operation requires both DSyntS to have the same verb, while the CONJUNCTION operation does not. Thus, individual probabilities are scaled to counterbalance these constraints.

Table 6: Aggregation parameters.

Parameters	Description
PERIOD	Leave two propositions in their own sentences, e.g. *'X has great Y. It has nice Z.'*
RELATIVE CLAUSE	Join propositions with a relative clause, e.g. *'X, which has great Y, has nice Z'*
WITH CUE WORD	Aggregate propositions using *with*, e.g. *'X has great Y, with nice Z'*
CONJUNCTION	Join propositions using a conjunction, or a comma if more than two propositions
MERGE	Merge the subject and verb of two propositions, e.g. *'X has great Y and nice Z'*
ALSO CUE WORD	Join two propositions using *also*, e.g. *'X has great Y, also it has nice Z'*
CONTRAST - CUE WORD	Contrast two propositions using *while, but, however, on the other hand*, e.g. *'While X has great Y, it has bad Z'*, *'X has great Y, but it has bad Z'*
WHILE CUE WORD	Contrast two propositions using *while*, e.g. *'While X has great Y, it has bad Z'*
HOWEVER CUE WORD	Contrast two propositions using *however*, e.g. *'X has great Y. However, it has bad Z'*
ON THE OTHER HAND CUE WORD	Contrast two propositions using *on the other hand*, e.g. *'X has great Y. On the other hand, it has bad Z'*
JUSTIFY - CUE WORD	Justify a proposition using *because, since, so*, e.g. *'X is the best, since it has great Y'*
BECAUSE CUE WORD	Justify a proposition using *because*, e.g. *'X is the best, because it has great Y'*
SINCE CUE WORD	Justify a proposition using *since*, e.g. *'X is the best, since it has great Y'*
SO CUE WORD	Justify a proposition using *so*, e.g. *'X has great Y, so it's the best'*
CONCEDE - CUE WORD	Concede a proposition using *although, even if, but/though*, e.g. *'Although X has great Y, it has bad Z'*, *'X has great Y, but it has bad Z though'*
ALTHOUGH CUE WORD	Concede a proposition using *although*, e.g. *'Although X has great Y, it has bad Z'*
EVEN IF CUE WORD	Concede a proposition using *even if*, e.g. *'Even if X has great Y, it has bad Z'*
BUT/THOUGH CUE WORD	Concede a proposition using *but/though*, e.g. *'X has great Y, but it has bad Z though'*
MERGE WITH COMMA	Restate a proposition by repeating only the object, e.g. *'X has great Y, nice Z'*
OBJECT ELLIPSIS	Replace part of a repeated proposition by an ellipsis, e.g. *'X has ... it has great Y'*

Table 7: Handcrafted probability distribution of aggregation operations expressing the INFER relation for the introvert and extravert parameter settings.[9] Parameter values must add up to 1.

Aggregation parameters for the INFER relation	Introvert distribution	Extravert distribution
INFER - MERGE	.20	.50
INFER - RELATIVE CLAUSE	.40	.00
INFER - WITH CUE WORD	.30	.10
INFER - ALSO CUE WORD	.00	.10
INFER - CONJUNCTION	.00	.29
INFER - PERIOD	.10	.01

neurotics produce more filled pauses and disfluencies (Scherer, 1979, 1981; Weaver, 1998). Neuroticism is associated with frustration and aquiescence, which are modeled with the EXPLETIVES and ACKNOWLEDGMENT parameters (Oberlander and Gill, 2004b; Weaver, 1998). Syntactic pattern matching controls the insertion of context-independent markers, while some parameters require more complex processing.

Syntactically embedded markers: Some pragmatic markers are inserted in the syntactic structure of an utterance (*like, you know, sort of*), and their insertion must respect particular syntactic constraints. Our approach is to add to the generation dictionary the syntactic representations that characterize each pragmatic marker. For each marker, the insertion process involves traversing the aggregated DSyntS to identify *insertion points* satisfying the syntactic constraints specified in the database.

Fig. 7 illustrates the matching and insertion process for the hedge *you know*. Each pragmatic marker dictionary entry consists of a syntactic pattern to be matched in the DSyntS, such as the root node in Fig. 7(b), and an insertion point element corresponding to the location in the DSyntS where the insertion should be made. Given the input DSyntS in Fig. 7(a) *'Chanpen Thai has good atmosphere'*, the verb *to have* is matched with the root node of the structure in Fig. 7(b), and thus the subtree below the insertion point is inserted under Fig. 7(a)'s root node. The resulting DSyntS is in Fig. 7(c). This DSyntS is realized as *'Chanpen Thai has good atmosphere, you know'*.

Table 8: Clause-combining operations from Rambow et al. (2001) and examples. [10]

Operation	Relations	Description	Sample 1st arg	Sample 2nd arg	Result
MERGE	INFER or CONTRAST	Two clauses can be combined if they have identical verbs and identical arguments and adjuncts except one. The non-identical arguments are coordinated.	Chanpen Thai has good service.	Chanpen Thai has good food quality.	Chanpen Thai has good service and good food quality.
WITH CUE WORD	JUSTIFY or INFER	Two clauses with identical subject arguments can be identified if one of the clauses contains the verb *to have*. The possession clause undergoes *with*-participial clause formation and is attached to the non-reduced clause.	Chanpen Thai is a Thai restaurant.	Chanpen Thai has good food quality.	Chanpen Thai is a Thai restaurant, with good food quality.
RELATIVE CLAUSE	JUSTIFY or INFER	Two clauses with an identical subject can be identified. One clause is attached to the subject of the other clause as a relative clause.	Chanpen Thai has good food quality.	Chanpen Thai is located in Midtown West.	Chanpen Thai, which is located in Midtown West, has good food quality.
CONJUNCTION	JUSTIFY, INFER or CONTRAST	Two clauses are conjoined with a coordinating conjunction. They are separated by a comma if the right clause already contains a conjunction.	Chanpen Thai has good food quality.	Chanpen Thai has good service.	Chanpen Thai has good food quality and it has good service.
ON THE OTHER HAND CUE WORD	CONTRAST	Combines clauses by inserting a cue word at the start of the second clause, resulting in two separate sentences.	Chanpen Thai has very good decor.	Baluchi's has mediocre decor.	Chanpen Thai has very good decor. On the other hand, Baluchi's has mediocre decor.
EVEN IF CUE WORD	CONCEDE	Combines clauses by inserting the *even if* adverbial at the start of the satellite clause. The order of the arguments is determined by the order of the nucleus (N) and the satellite (S), yielding two distinct operations, EVEN IF CUE WORD NS and EVEN IF CUE WORD SN.	Chanpen Thai has very good decor.	Chanpen Thai's has mediocre food quality.	Chanpen Thai has very good decor, even if it has mediocre food quality.
MERGE WITH COMMA	RESTATE	Merges repeated clauses in the same way as the MERGE operation, but ensures that the non-identical arguments are separated by a comma.	Chanpen Thai has very good service.	Chanpen Thai has fantastic waiters.	Chanpen Thai has very good service, fantastic waiters.
OBJECT ELLIPSIS	RESTATE	Coordinates clauses and replaces the object of the first clause by a three-dot ellipsis.	Chanpen Thai has fantastic waiters.	Chanpen Thai has fantastic waiters.	Chanpen Thai has... It has fantastic waiters.
PERIOD	Any	Two clauses are joined by a period.	Chanpen Thai is a Thai restaurant, with good food quality.	Chanpen Thai has good service.	Chanpen Thai is a Thai restaurant, with good food quality. It has good service.

This approach supports modifying utterances at the syntactic level rather than at the surface level, which allows the RealPro surface realizer to do what it was designed for, namely to enforce grammaticality. For example, pragmatic markers are added without controlling the final word order, while positional constraints can be enforced when required, e.g. the *position* attribute in Fig. 7(b) specifies that *you know* should be in sentence final position. Similarly, while the *punct* attribute specifies that the marker must appear between commas—irrespectively of its position in the utterance, the realizer ensures that the sentence is punctuated correctly by removing commas preceding the final period. We believe that the DSyntS representation is also what enables us to insert many different pragmatic markers into the same utterance while carrying out syntactic transformations such as negation insertion, while still ensuring that we produce grammatical outputs.

PERSONAGE implements a binary generation parameter for the pragmatic markers in Table 10 using the insertion mechanism described above. At generation time, syntactic patterns are randomly chosen (with a uniform distribution) among markers with parameter values set to 1, and matched against the aggregated DSyntS. The insertion process ends when there are no markers left in the database, or when the number of successful insertions is above a constant threshold (heuristically set to 5 for the current domain) to avoid producing unnatural utterances.

Table 9: Pragmatic marker insertion parameters.

Parameters	Description
SUBJECT IMPLICITNESS	Make the presented object implicit by moving its attribute to the subject, e.g. *'the Y is great'*
NEGATION	Negate a verb by replacing its modifier by its antonym, e.g. *'X doesn't have bad Y'*
SOFTENER HEDGES	Insert syntactic elements (*sort of*, *kind of*, *somewhat*, *quite*, *around*, *rather*, *I think that*, *it seems that*, *it seems to me that*) to mitigate the strength of a proposition, e.g. *'X has kind of great Y'* or *'It seems to me that X has rather great Y'*
EMPHASIZER HEDGES	Insert syntactic elements (*really*, *basically*, *actually*, *just*) to strengthen a proposition, e.g. *'X has really great Y'* or *'Basically, X just has great Y'*
ACKNOWLEDGMENTS	Insert an initial back-channel (*yeah*, *right*, *ok*, *I see*, *oh*, *well*), e.g. *'Ok, X has great Y'*
FILLED PAUSES	Insert syntactic elements expressing hesitancy (*I mean*, *err*, *mmhm*, *like*, *you know*), e.g. *'Err... X has, like, great Y'*
EXCLAMATION	Insert an exclamation mark, e.g. *'X has great Y!'*
EXPLETIVES	Insert a swear word, e.g. *'the Y is damn great'*
NEAR EXPLETIVES	Insert a near-swear word, e.g. *'the Y is darn great'*
TAG QUESTION	Insert a tag question, e.g. *'the Y is great, isn't it?'*
STUTTERING	Duplicate parts of a content word, e.g. *'X has gr-gr-great Y'*
IN-GROUP MARKER	Refer to the hearer as a member of the same social group, e.g. *pal*, *mate* and *buddy*
PRONOMINALIZATION	Replace references to the object by pronouns, as opposed to proper names or the reference *this restaurant*
REQUEST CONFIRMATION	Begin the utterance with a confirmation of the request, e.g. *'did you say X?'*
INITIAL REJECTION	Begin the utterance with a rejection, e.g. *'I'm not sure'*
COMPETENCE MITIGATION	Express the speaker's negative appraisal of the hearer's request, e.g. *'everybody knows that . . . '*

Other markers: The remaining pragmatic markers (marked with an asterisk in Table 10) require more complex syntactic processing and are implemented independently.

Proximal deictic expressions are a way to express involvement and empathy (Brown and Levinson, 1987), e.g. *'this restaurant has great service'*. Referring expression generation in PERSONAGE is based on a simple algorithm which identifies as potential anaphoric expressions any restaurant name that follows a previous reference to it, e.g. *'Chanpen Thai is the best, it has great service'*. Then, a PRONOMINALIZATION parameter controls whether referring expressions are realized as personal pronouns or proximal demonstrative phrases. by specifying the ratio of pronouns to other types of referring expressions. The RealPro surface realiser automatically selects the personal pronoun based on the selection's DSyntS node; inserting a demonstrative phrase requires replacing the selection's lexeme with a generic noun (e.g. *restaurant*) and setting the determiner to a demonstrative.

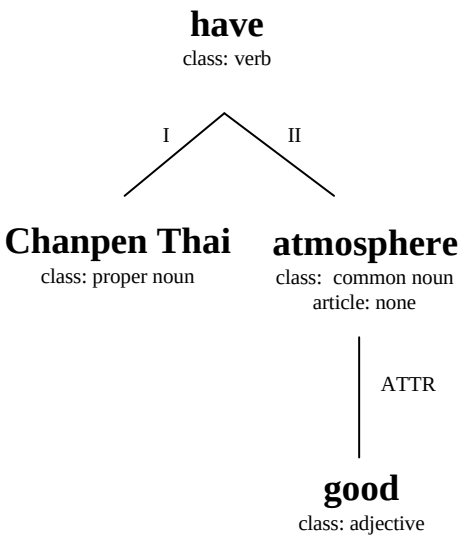
Negations indicate both introversion and a lack of conscientiousness (Mehl et al., 2006; Pennebaker and King, 1999), a NEGATION parameter allows PERSONAGE to insert a negation while preserving the initial communicative goal. An adjective modifying a verb or its object is randomly selected from the DSyntS, and its antonym is retrieved from WordNet (Fellbaum, 1998). If the query is successful, the adjective's lexeme is replaced by the antonym and the governing verb is negated,[11] e.g. *'Chanpen Thai has good atmosphere'* becomes *'Chanpen Thai doesn't have bad atmosphere'*. Adjectives in the domain are manually sense-tagged to ensure that they can be substituted by their antonym. Also, a maximum of one negation can be inserted to prevent the utterance from sounding unnatural.

Heylighen and Dewaele (2002) found that extraverts use more implicit language than introverts. A SUBJECT IMPLICITNESS parameter thus determines whether predicates describing restaurant attributes are expressed with the restaurant's name in the subject, or with the attribute itself by making the reference to the restaurant implicit (e.g. *'Chanpen Thai has good atmosphere'* vs. *'the atmosphere is good'*). The syntactic transformation involves shifting the object attribute to the subject, while promoting the adjective below the main verb, and changing the main verb's lexeme to *to be*. Hence, the transformation requires an

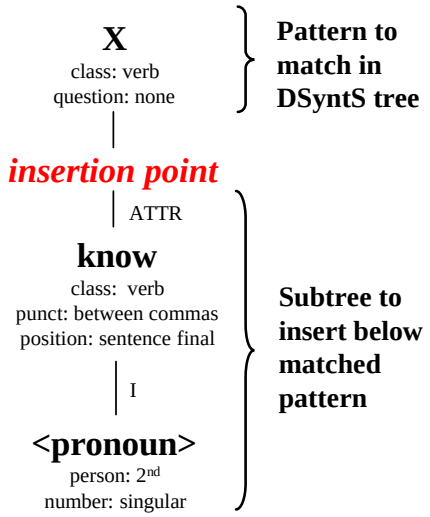
[11] At the DSyntS level the negation is represented as an attribute of the verb element, the actual inflection is done by RealPro in the realization phase.

Table 10: Pragmatic markers implemented in PERSONAGE, with insertion constraints and example realizations. An asterisk indicates that the pragmatic marker requires specific processing and was not implemented through pattern matching and insertion.

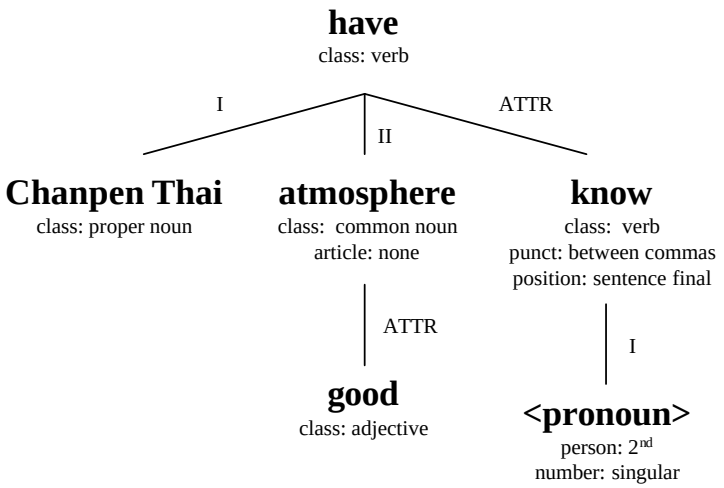
Marker	Constraints	Example
General:		
NEGATION*	adjective modifier + antonym	Chanpen Thai doesn't have bad atmosphere
EXCLAMATION	sentence-final punctuation	Chanpen Thai has good atmosphere!
IN-GROUP MARKER	clause-final adjunct, e.g. *pal*, *mate* and *buddy*	Chanpen Thai has good atmosphere pal
SUBJECT IMPLICITNESS*	requires a DSyntS of the form *NOUN has ADJ NOUN*	The atmosphere is good
TAG QUESTION*	none	Chanpen Thai has good atmosphere, doesn't it?
STUTTERING*	selection name	Ch-Chanpen Thai has good atmosphere
EXPLETIVES	adjective modifier (*damn*, *bloody*)	Chanpen Thai has damn good atmosphere
	clause-initial adjunct (*oh god*)	Oh god Chanpen Thai has good atmosphere
NEAR EXPLETIVES	adjective modifier (*darn*)	Chanpen Thai has darn good atmosphere
	clause-initial adjunct (*oh gosh*)	Oh gosh Chanpen Thai has good atmosphere
REQUEST	none	You want to know more about Chanpen Thai?
CONFIRMATION*		Let's see... Chanpen Thai Let's see what we can find on Chanpen Thai Did you say Chanpen Thai?
INITIAL REJECTION*	none	I don't know I'm not sure I might be wrong
COMPETENCE MITIGATION	main verb is subordinated to new clause (*everybody knows that* and *I thought everybody knew that*)	Everybody knows that Chanpen Thai has good atmosphere
	clause-initial adjunct (*come on*)	Come on, Chanpen Thai has good atmosphere
Softeners:		
KIND OF	adjective modifier	Chanpen Thai has kind of good atmosphere
SORT OF	adjective modifier	Chanpen Thai has sort of good atmosphere
SOMEWHAT	adjective modifier with verb *to be*	The atmosphere is somewhat good
QUITE	adjective modifier	Chanpen Thai has quite good atmosphere
RATHER	adjective modifier	Chanpen Thai has rather good atmosphere
AROUND	numeral modifier	Chanpen Thai's price is around $44
SUBORDINATE	main verb is subordinated to hedge clause, e.g. *I think that* and *it seems (to me) that*	It seems to me that Chanpen Thai has good atmosphere
Filled pauses:		
LIKE	verb modifier	Chanpen Thai has, like, good atmosphere
ERR	clause-initial adjunct	Err... Chanpen Thai has good atmosphere
MMHM	clause-initial adjunct	Mmhm... Chanpen Thai has good atmosphere
I MEAN	clause-initial adjunct	I mean, Chanpen Thai has good atmosphere
YOU KNOW	clause-final adjunct	Chanpen Thai has good atmosphere, you know
Emphasizers:		
REALLY	adjective modifier	Chanpen Thai has really good atmosphere
BASICALLY	clause-initial adjunct	Basically, Chanpen Thai has good atmosphere
ACTUALLY	clause-initial adjunct	Actually, Chanpen Thai has good atmosphere
JUST	pre-verbal modifier of *to have*	Chanpen Thai just has good atmosphere
	post-verbal modifier of *to be*	The atmosphere is just good
Acknowledgments:		
YEAH	clause-initial adjunct	Yeah, Chanpen Thai has good atmosphere
WELL	clause-initial adjunct	Well, Chanpen Thai has good atmosphere
OH	clause-initial adjunct	Oh, Chanpen Thai has good atmosphere
RIGHT	clause-initial adjunct	Right, Chanpen Thai has good atmosphere
OK	clause-initial adjunct	Ok, Chanpen Thai has good atmosphere
I SEE	clause-initial adjunct	I see, Chanpen Thai has good atmosphere



(a) Example input DSyntS realized as *'Chanpen Thai has good atmosphere'*.



(b) Syntactic representation of the insertion constraints for the pragmatic marker *you know*.



(c) Modified DSyntS after the insertion of the pragmatic marker below the main verb matching the pattern defined in Fig. 7(b)'s root node.

Fig. 7: Illustration of the pragmatic marker insertion process for the hedge *you know* in the DSyntS *'Chanpen Thai has good atmosphere'*.

input DSyntS matching the template *NOUN has ADJECTIVE NOUN*.

Speech disfluencies are associated with anxiety and neuroticism (Scherer, 1981), so we introduce a STUTTERING parameter that modifies the lexeme of a randomly selected proper noun by repeating the first two letters two or three times, e.g. *'Ch-Ch-Chanpen Thai'*. Only selection names are repeated as they are likely to be new to the speaker, the stuttering can therefore be interpreted as non-pathological. Allowing disfluencies to affect any word requires determining what words can be altered, which involves deep psycholinguistic modeling that is beyond the scope of this work.

PERSONAGE also implements politeness markers such as rhetorical questions. The TAG QUESTION parameter processes the DSyntS by (1) duplicating a randomly selected verb and its subject; (2) negating the

verb; (3) pronominalizing the subject; (4) setting the verb to the interrogative form and (5) appending the duplicated subtree as a sentence-final adjunct, e.g. *'Chanpen Thai has great food'* results in the insertion of *'doesn't it?'*. The duplicated verb is generally not realized,[12] i.e. only the negated auxiliary appears in the tag question. Additionally, whenever the subject is a first person pronoun, the verb is set to the conditional form and a second person pronoun is inserted, producing *'I would recommend Chanpen Thai, wouldn't you?'*. If the tag question insertion is unsuccessful, e.g. due to an extrapolated subject *'there is'*, a default tag question is appended, producing either *'you see?'*, *'alright?'* or *'okay?'*.

As mentioned above, the REQUEST CONFIRMATION, INITIAL REJECTION and COMPETENCE MITIGATION parameters are content level parameters that we implement as pragmatic markers, by inserting a full DSyntS at the beginning of the utterance, randomly chosen from the dictionary of such markers.[13] The INITIAL REJECTION parameter reduces the level of confidence of the speaker over the utterance's informational content, by beginning the utterance with either *'I don't know'*, *'I'm not sure'* or *'I might be wrong'*. The REQUEST CONFIRMATION parameter produces an implicit confirmation, which both redresses the hearer's positive face through grounding and emphasizes the system's uncertainty about the user's request, e.g. *'you want to know more about Chanpen Thai?'*. In order to convey disagreeableness, a COMPETENCE MITIGATION parameter also presents the user's request as trivial by embedding it as a subordinate clause, e.g. *'everybody knows that Chanpen Thai has good service'*. See Table 10 for additional examples of confirmation and competence mitigation DSyntS.

2.5 Lexical choice

Lexical features related to personality include word length, word frequency and verb strength. In addition, lexical choice is crucial to successful individual adaptation in dialogue systems (Brennan, 1996; Lin, 2006). Thus, PERSONAGE allows many different lexemes to be expressed for each content word, depending on input parameter values. See Table 11.

Table 11: Lexical choice Parameters

Parameters	**Description**
LEXICON FREQUENCY	Control the average frequency of use of each content word (e.g. according to frequency counts from a corpus)
LEXICON WORD LENGTH	Control the average number of letters of each content word
VERB STRENGTH	Control the strength of the verbs, e.g. *'I would suggest'* vs. *'I would recommend'*

The lexical selection component processes the DSyntS by sequentially modifying each content word. For each lexeme in the DSyntS, the corresponding WordNet synonyms are mapped to a multi-dimensional space defined by the lexeme's length, frequency of use and strength, using machine-readable dictionaries. The values along each dimension are normalized over the set of synonyms, and the synonym that is the closest to the target parameter values (in terms of Euclidean distance) is selected. Although word-sense disambiguation techniques could be used in the future, content words are manually sense-tagged to ensure that the synonyms are interchangeable in the dialogue domain. Fig. 8 illustrates the lexical choice process using the word length and word frequency dimensions, resulting in the selection of *cheap* over *inexpensive* because its length (5 letters) and its normalized frequency (1.0) are closer to the desired target values, i.e. a 6 letter word (normalized length of $\frac{6-5}{11-5} = .17$) with a normalized frequency of .7.

In order to enrich the pool of synonyms from Wordnet, adjectives extracted by Higashinaka et al. (2007) from a corpus of restaurant reviews and their synonyms are added to the synonym set of each attribute modifier. As the extraction process is not entirely accurate, the list of adjectives is filtered manually. As Higashinaka et al.'s method automatically extracts polarity values for each adjective on a scale from 1 to 5 based on the ratings of the associated reviews, the synonym set for a specific attribute is determined at generation time by mapping the attribute's scalar rating to the polarity scale, e.g. a DSyntS expressing a food quality rating of .42 is mapped to the adjective set with polarity 2 (as $\frac{2}{5} \sim .42$), consisting of the

[12]The verb *to be* is an exception.

[13]The constraint on the maximum number of pragmatic markers in the utterance also affects the insertion probability of the DSyntS.

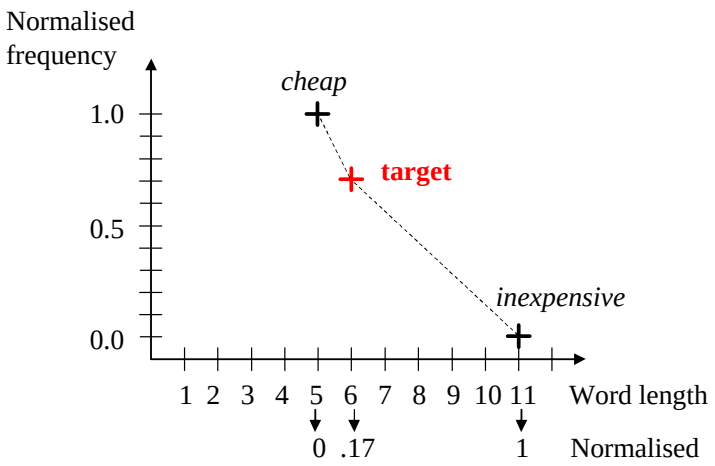


Fig. 8: Illustration of the lexical selection process between the synonyms *cheap* and *inexpensive* with two input dimensions.

modifiers *bland*, *mediocre* and *bad*. Table 12 lists the extracted adjective sets for the food quality attribute, ordered by polarity.

Table 12: Adjectives and polarity ratings (5=very positive) for the food quality attribute, extracted from a corpus of restaurant reviews by Higashinaka et al. (2007).

Polarity	**Adjectives**
1	awful, bad, terrible, horrible, horrendous
2	bland, mediocre, bad
3	decent, acceptable, adequate, satisfying
4	good, flavourful, tasty, nice
5	excellent, delicious, great, exquisite, wonderful, legendary, superb, terrific, fantastic, outstanding, incredible, delectable, fabulous, tremendous, awesome, delightful, marvellous

The synonym selection is implemented in PERSONAGE by jointly controlling the average normalized frequency of use, word length and verb strength in each DSyntS.

Frequency of use: Introvert and emotionally stable speakers use a richer vocabulary (Dewaele and Furnham, 1999; Gill and Oberlander, 2003). We model this with a LEXICON FREQUENCY parameter that selects lexical items associated with a particular part of speech using the frequency count in the British National Corpus, in order to support the selection of unusual low-frequency words.

Word length: Mehl et al. (2006) show that observers associate long words with agreeableness, conscientiousness and openness to experience. Thus we introduce a LEXICON WORD LENGTH parameter to control the number of letters of the selected synonym.

Verb strength: Verb synonyms, such as *appreciate*, *like* and *love*, differ in terms of their connotative strength (Inkpen and Hirst, 2004; Wilson et al., 2004). This variation is controlled in PERSONAGE through the VERB STRENGTH parameter, which orders each verb's synonym set according to the *stronger-than* semantic relation in the VERBOCEAN database (Chklovski and Pantel, 2004). The process is illustrated in Fig. 9 for synonyms of the verb *to know*. The ordered synonyms are mapped to equidistant points in the $[0, 1]$ interval to produce the final parameter value, i.e. the weakest verb is associated with 0.0 and the strongest with 1.0. This mapping is based on the assumption that the magnitude of the *stronger-than* relation is constant between contiguous synonyms, i.e. the verb strength is uniformly distributed over the synonym set.

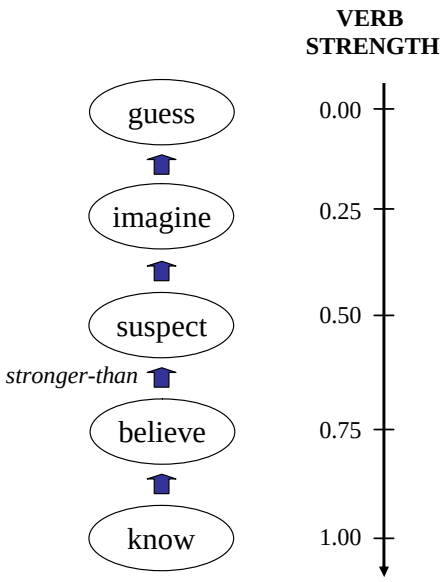


Fig. 9: Determination of the VERB STRENGTH parameter values for synonyms of the verb *to know*, based on the *stronger-than* semantic relation in VERBOCEAN.

The lexical choice parameters described above associate each candidate synonym with three values, and the one with the closest values to the target is selected. Since values are normalized over the members of the synonym set, all dimensions have the same weight in the selection process.[14] For example, consider the input DSyntS expressing *'I know you would like Chanpen Thai'*; a low VERB STRENGTH parameter value produces *'I guess you would like Chanpen Thai'*, while a high value yields *'I know you would love Chanpen Thai'*. Similarly, a proposition realized as *'this place has great ambiance'* is converted into *'this restaurant features fantastic atmosphere'* given high LEXICON WORD LENGTH and VERB STRENGTH parameter values, together with a low LEXICON FREQUENCY value.

2.6 Surface realization

Surface realization is the process of converting the DSyntS, i.e. dependency trees such as those as shown in Fig. 7, into a sentence string. Surface realization is a fairly well understood process, independent of other components, that involves applying rules of English grammar such as word order, word inflection, and function word and punctuation insertion. We use the RealPro surface realizer (Lavoie and Rambow, 1997) to convert the final sequence of DSyntS into a string, with each DSyntS corresponding to one sentence in the utterance.[15]

3 From Personality Markers to Generation Decisions

The PERSONAGE base generator described above can produce thousands of utterances for any input content plan, and this variation needs to be controlled to achieve particular communicative goals. Table 13 provides examples of utterances generated when the PERSONAGE parameters are randomly varied. Random variation can result in inconsistent personality cues that may sometimes, but not reliably, be recognized as manifesting a particular personality. See the personality ratings in Table 13. Figures 10 and 11, to be discussed in

[14] An exception is that verb selection is only affected by the VERB STRENGTH parameter, to ensure that its effect is perceptible in the output utterance.

[15] Some generation decisions in Sections 2.4 and 2.5 could be treated as part of the realization phase as well, such as the insertion of exclamation marks and lexical choice. This is only a matter of terminology.

more detail below, illustrate clearly the differences in the distributions of ratings for randomly generated utterances vs. those generated using personality models. Figures 10(a) and 11(a) are normal distributions (random utterances). Thus they do not convey any strong personality. In contrast, the bi-modal distributions (personality model utterances) of Figures 10(b) and 11(b) reliably project a particular personality trait.

Table 13: Example outputs of PERSONAGE with random parameter settings (random utterances), together with average user ratings (see evaluation in Section 4).

#	Content plan	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(Chimichurri 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(Chimichurri 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

In the appendix, Table 25 summarizes the studies that motivate our parameters. An important thing to realize about these studies is that they are typically based on lexical or syntactic tokens or types that can be counted, and the results reported consist of correlations between these counts and personality traits. In order to use such results, we explore various parameters that could have caused an increase or decrease in a particular count, e.g. a high word count can be associated with the expression of more content and/or more repetitions and redundancies. Thus, one of the important contributions of this work is that we test whether and when findings from other language genres (see the *Language source* column in Table 25) generalize to the production of a single utterance presenting information to the user in a controlled discourse situation, such as a recommender dialogue system.

Section 3.1 presents the findings from these studies relevant to extraversion, and Section 3.2 presents the findings for emotional stability (neuroticism). We then proceed to develop personality models from these findings. To do so, each finding (correlation) is first mapped to one or more parameters (generation decisions) of the PERSONAGE generator described in Section 2. Second, the personality model for each trait is expressed via parameter trends specified in terms of `high` or `low` settings for particular parameters relevant to that trait, based on the direction and the magnitude of the correlations. See Tables 14 and 16. Third, at generation time, these parameter trends are mapped to extreme parameter values to maximize their impact on the utterance, with $low = 0.0$ and $high = 1.0$ for most continuous and binary parameters,[16] and unspecified parameters set to default values.[17] However, if PERSONAGE always expressed a trait using identical parameter settings (e.g. neurotic), then identical generation decisions (e.g. hesitancy markers) could lead to excessive repetitions of particular linguistic forms. Therefore, parameter values are randomized before generation,

[16] Aggregation parameters are set to continuous values between 0 and 1 according to a predefined distribution reflecting the high/low values in the table, as for example the PERIOD operation probability is never 0 to ensure that the aggregation process will terminate.

[17] Default values are chosen to minimize the resulting pragmatic effect, e.g. VERBOSITY and CONTENT POLARITY are set to 0.5, while binary pragmatic markers are set to 0.

according to a normal distribution, with a 15% standard deviation around their predefined value,[18] in order to exploit PERSONAGE's variation capabilities. An exception is that the probability distributions of aggregation operations are handcrafted for each trait, to factor in the different probabilities of success of clause-combining operations, and to ensure that the parameter values add up to 1.

3.1 Extraversion

Extraverts tend to engage in social interaction, they are enthusiastic, risk-taking, talkative and assertive, while introverts are more reserved and solitary. Eysenck et al. (1985) suggest that this trait is associated with a lack of internal arousal: extraverts are thus seeking additional external stimulation, while introverts avoid it. The extraversion trait has received the most attention in linguistic studies. There are three reasons for this: (1) the extraversion dimension is often seen as the most important, since it explains the most variance among the adjective descriptors from which the Big Five factors are derived (Goldberg, 1990), (2) extraversion is present in most other personality frameworks—e.g. Eysenck et al.'s PEN model (Psychoticism, Extraversion and Neuroticism; 1985); and (3) extraversion may have the most influence on language, because it is strongly associated with talkativeness and enthusiasm (Furnham, 1990).

The findings about linguistic markers of extraversion are summarized in Table 14, together with the generation parameters that represent our hypotheses about how each finding can be mapped into the PERSONAGE framework. Most generation parameters are based on study results, however some are derived from hypotheses about how a specific trait affects language (indicated by a single asterisk). The right-most columns (e.g. *Intro* and *Extra*) contain the parameter values for expressing each end of the personality dimension, i.e. either introversion or extraversion. As mentioned above, parameter values are specified in terms of *low* and *high* settings, and then mapped to normalized scalar values between 0 and 1.

Table 15 shows examples generated by PERSONAGE using the extraversion personality model specified by the parameter settings in Table 14. Note how a given content plan (e.g. recommend Amy's Bread) can generate multiple examples for both extremes of the extraversion scale (introversion, extraversion). Below we present the derivation of each parameter value from the psychology findings.

Content planning: It has been repeatedly found that extraverts are more talkative than introverts (Cope, 1969; Dewaele and Furnham, 1999; Furnham, 1990; Mehl et al., 2006; Pennebaker and King, 1999). However, because findings are based on word count, it is not clear whether extraverts actually produce more content, or are just redundant and wordy. Therefore the extraversion personality model uses the VERBOSITY parameter to control the number of propositions expressed in the utterance, a REPETITIONS parameter to produce an exact repetition of a proposition, and a RESTATEMENTS parameter to produce a paraphrased repetition. Utterance 6 in Table 15 illustrates a strict repetition (*good food*), while utterance 13 restates the price information (*its price is 12 dollars, it's cheap*).

Extraverts are more positive; introverts are characterized as engaging in more 'problem talk' and expressions of dissatisfaction (Pennebaker and King, 1999; Thorne, 1987). This positivity can manifest itself through the choice of information presented to the user, which is controlled by a CONTENT POLARITY parameter. Utterance 10 in Table 15 provides an example negative claim, while utterance 14 contains a positive claim. In addition, polarity can also be implied by presenting information subjectively, thus a CONCESSION POLARITY parameter controls whether the positive or the negative content is emphasized, such as in *'even if the food is great, it's expensive'* vs. *'even if it is expensive, the food is great'* in utterance 16. Additional emphasis is conveyed using a REPETITION POLARITY parameter, controlling whether positive or negative information is more likely to be repeated in the utterance.

Carenini and Moore (2000) claim that starting with a positive claim facilitates the hearer's understanding, while finishing with it is more effective if the hearer disagrees. We hypothesize that extraverts begin their utterances with more positive content, as a consequence of their high enthusiasm, and control this by setting the POSITIVE CONTENT FIRST parameter to `high` for extraversion and `low` for introversion.

Weaver (1998) shows that extraverts are more sympathetic to other people—i.e. they show more concern—although this sympathy is not related to empathy, as they are not more inclined to feel other

[18] Binary parameter values are then rounded to 0 or 1.

Table 14: Summary of language cues for extraversion, with corresponding generation parameters. Asterisks indicate hypotheses, rather than results. An unreferenced asterisk indicates a new hypothesis. Referenced studies are detailed in Table 25.

Introvert findings	Extravert findings	Study	Parameters	Intro	Extra
Content planning:					
Single topic	Many topics, higher verbal output	1,3,4, 6,13	VERBOSITY	low	high
Strict selection	Think out loud	1*	RESTATEMENTS	low	high
			REPETITIONS	low	high
Problem talk, dissatisfaction, negative emotion words	Pleasure talk, agreement, compliment, positive emotion words	3,14	CONTENT POLARITY	low	high
			REPETITION POLARITY	low	high
			CONCESSION POLARITY	low	high
			POSITIVE CONTENT FIRST	low	high
Not sympathetic	Sympathetic, concerned about hearer (but not empathetic)	10	REQUEST CONFIRMATION	low	high
Syntactic template selection:					
Elaborated constructions	Simple constructions	1*	SYNTACTIC COMPLEXITY	high	low
Problem talk	Pleasure talk	3	TEMPLATE POLARITY	low	high
Aggregation:					
Few conjunctions	Many conjunctions	8	CONJUNCTION, BUT, ALSO CUE WORD	low	high
Many unfilled pauses	Few unfilled pauses	2,7	PERIOD	high	low
Many uses of *although*	Few uses of *although*	9	ALTHOUGH CUE WORD	high	low
Formal language	Informal language	1*,11	RELATIVE CLAUSE	high	low
Pragmatic marker insertion:					
Many nouns, adjectives, prepositions (explicit)	Many verbs, adverbs, pronouns (implicit)	11	SUBJECT IMPLICITNESS	low	high
Many negations	Few negations	3	NEGATION	high	low
Many tentative words (e.g. *maybe*, *guess*)	Few tentative words	3	SOFTENER HEDGES: ·SORT OF, SOMEWHAT, QUITE, RATHER, I THINK THAT, IT SEEMS THAT, IT SEEMS TO ME THAT	high	low
Formal language	Informal language	1*,11	·KIND OF, LIKE	low	high
			ACKNOWLEDGMENTS:		
			·YEAH	low	high
			·RIGHT, OK, I SEE, WELL	high	low
Few swear words	Many swear words	6	NEAR EXPLETIVES	low	high
Many unfilled pauses	Few unfilled pauses	2,7	FILLED PAUSES: · ERR, I MEAN, MMHM, YOU KNOW	high	low
Realism	Exaggeration (e.g. *really*)	9*	EMPHASIZER HEDGES: ·REALLY, BASICALLY, ACTUALLY, JUST	low	high
			EXCLAMATION	low	high
Not sympathetic	Sympathetic, concerned about hearer, minimise positive face threat	10	TAG QUESTION	low	high
Few words related to humans	Many words related to humans (e.g. *man*, *pal*)	12	IN-GROUP MARKER	low	high
Lexical choice:					
Rich vocabulary	Poor vocabulary	1*,4	LEXICON FREQUENCY	low	high
Longer words	Shorter words	6	LEXICON WORD LENGTH	high	low
Realism	Exaggeration	*	VERB STRENGTH	low	high

people's feelings. Concern for the user can be expressed in the information presentation domain by emphasizing the user's request through an explicit confirmation. This is controlled by setting the REQUEST CONFIRMATION parameter to `high` for high extraversion.

Syntactic template selection: The preference of extraverts for positive content can also be expressed through syntactic template selection, by controlling the connotation of the utterance's syntactic representation using a TEMPLATE POLARITY parameter, e.g. a high parameter value would favor the selection of the claim *'X is the best place'* (see utterance 16 in Table 15) rather than *'X is not as bad as the others'* (see utterance 9). In addition, Furnham (1990) hypothesizes that extraverts use simpler syntactic constructions,

Table 15: Example outputs of PERSONAGE for the **extraversion** personality model. Personality ratings are on a scale from 1 to 7, with 1 = very low and 7 = very high.

#	Content plan	End	PERSONAGE's output	Score
1	**compare(Lemongrass Grill, Monsoon)**	**low**	I think that Lemongrass Grill features mediocre ambience. Monsoon doesn't, on the other hand, have nasty ambiance.	3.50
2	**compare(Dishes, Joe's Shanghai)**	**low**	Right, I think that Dishes doesn't have good ambience. Joe's Shanghai, which features mediocre ambiance, is a chinese eating place. Its price is 23 dollars. There could be worse restaurants.	4.16
3	**compare(Aureole, Daniel)**	**low**	Right, Daniel's price is 82 dollars. Aureole, however, features marvelous waiters. I might recommend Daniel and Aureole.	3.00
4	**compare(Da Andrea, John's Pizzeria, Uguale)**	**low**	Da Andrea is costly while John's Pizzeria doesn't feature nasty waiters. John's Pizzeria, Da Andrea and Uguale are satisfactory restaurants.	2.33
5	**compare(Lemongrass Grill, Monsoon)**	**high**	Yeah, Lemongrass Grill's price is 22 dollars, even if the ambience is just really poor. Monsoon is low-cost and the atmosphere is nice.	5.67
6	**compare(Dishes, Joe's Shanghai)**	**high**	Yeah, I would recommend Joe's Shanghai and Dishes, wouldn't you? Basically, Dishes has good food. Actually, its price is 14 dollars, even if the atmosphere is poor. Joe's Shanghai is affordable and it's a kind of chinese place. Even if the atmosphere is bad, it has good food, good food.	6.17
7	**compare(Aureole, Daniel)**	**high**	I am sure you would like Daniel and Aureole, you know. Aureole has really great food, the servers are wonderful and the atmosphere is beautiful. Daniel has great service with lovely atmosphere. Even if it's expensive, the food is excellent.	7.00
8	**compare(Da Andrea, John's Pizzeria, Uguale)**	**high**	I would recommend John's Pizzeria, Da Andrea and Uguale, wouldn't you? Da Andrea has great servers and the atmosphere is acceptable. Uguale has nice atmosphere. Actually, the service is good, even if its price is 33 dollars. Even if John's Pizzeria just has poor atmosphere, its price is around 20 dollars. The servers are friendly.	5.33
9	**recommend(Amy's Bread)**	**low**	Amy's Bread's price is 12 dollars. It isn't as bad as the others.	2.50
10	**recommend(Amy's Bread)**	**low**	Amy's Bread, which is the only restaurant that is any good, isn't expensive.	3.83
11	**recommend(Bond Street)**	**low**	Bond Street is the only restaurant that is any good.	2.83
12	**recommend(Bond Street)**	**low**	It seems that Bond Street is the only restaurant with friendly staff that is any good.	3.67
13	**recommend(Amy's Bread)**	**high**	I am sure you would like Amy's Bread. Basically, its price is 12 dollars, it's cheap, you know, the food is good and the servers are friendly.	6.50
14	**recommend(Amy's Bread)**	**high**	I am sure you would like Amy's Bread, it's one of my favourite places, you know. Its price is around 12 dollars. The food is just good. It's in Midtown West and a cafe restaurant with nice servers.	6.33
15	**recommend(Bond Street)**	**high**	I am sure you would like Bond Street, you know. Basically, the food is great and the atmosphere is good with friendly service.	5.67
16	**recommend(Bond Street)**	**high**	Yeah, Bond Street is the best place. The atmosphere is good, it has nice service and it's a japanese and sushi place. Even if it's expensive, you know, the food is great.	6.67

so the template selection is influenced by a SYNTACTIC COMPLEXITY parameter controlling the template's level of subordination, e.g. the claim *'I am sure you would like X'* is more syntactically complex than *'X is the best'*.

Aggregation: Oberlander and Gill (2004b) show that extraverts use more conjunctions in their emails. Thus, an extravert system should combine pieces of information using conjunctions, such as *and*, *but* and *also*. Oberlander and Gill (2004b) also find that introversion is associated with the use of the adverbial clause *although*, which can be expressed by selecting the ALTHOUGH CUE WORD operation when conceding a piece of information over another, as opposed to EVEN IF CUE WORD for example. It has also been found that introverts produce more long unfilled pauses (Scherer, 1979; Siegman and Pope, 1965), which can be controlled at the aggregation level by enforcing that the utterance's propositions are expressed in separate sentences, using the PERIOD aggregation operation, such as in utterance 9 in Table 15.

The probabilities for the distributions for the INFER relation for extraversion were shown in Table 7. The probability of the operations biases the production of complex clauses, full stops and formal cue words for introverts, to express their preference for complex syntactic constructions, long pauses and rich vocabulary (Furnham, 1990). Thus, the introvert parameters favor operations such as RELATIVE CLAUSE and PERIOD for the INFER relation, HOWEVER CUE WORD for CONTRAST, and ALTHOUGH CUE WORD for CONCEDE, that we hypothesize to result in more formal language. Extravert aggregation produces longer sentences with

simpler constructions and informal cue words. Thus extravert utterances tend to use operations such as a CONJUNCTION to realize the INFER and RESTATE relations, and the EVEN IF CUE WORD for CONCEDE relations (see utterance 16 in Table 15).

Pragmatic marker insertion: Psychological studies identify many pragmatic markers of extraversion which only affect the utterance locally, and can thus be implemented as separate syntactic transformations. These studies show that introverts produce more negations, tentative words (e.g. *maybe*, *perhaps*) and filled pauses (Pennebaker and King, 1999; Scherer, 1979; Siegman and Pope, 1965). Negations can be controlled—while preserving the original meaning—by a NEGATION parameter that negates the logical inverse of a proposition, e.g. by producing *'X is not bad'* rather than *'X is good'*. See utterance 1 in Table 15. Tentativeness can be expressed through hedging expressions that mitigate the impact of the speaker's statement—referred to as SOFTENER HEDGES—including *sort of*, *somewhat*, *rather*, *it seems that*, etc. Utterance 12 provides an example of a subordinate hedge. Filled pauses can be expressed linguistically by inserting the adjuncts *err*, *mmhm*, *I mean*, *like* and *you know*, which are all placed under the FILLED PAUSES category.[19]

Extraverts use more informal, implicit language (Heylighen and Dewaele, 2002). We associate informal language with the use of adverbial hedges such as *kind of* and *like*, as well as acknowledgments such as *yeah* (as opposed to *well* or *right* for example, which are hypothesized to be more formal). Implicitness can be conveyed in the information presentation domain by referring to the object of interest implicitly through its attributes, such as in *'the food is good'* (X is implicit) vs. *'X has good food'* (X is explicit). This syntactic transformation is controlled by the SUBJECT IMPLICITNESS parameter. Oberlander and Gill (2004b) also find that extravert emails contain more occurrences of *'I really'* as well as more exclamation marks, suggesting the need for parameters controlling the insertion of adverbs such as *really*, *basically*, *actually* and *just*—referred to as EMPHASIZER HEDGES—as well as an EXCLAMATION parameter ending the utterance with an exclamation mark. See *the ambience is just really poor* in utterance 5 in Table 15.

Extraversion is also associated with more swearing and references to humans (Mehl et al., 2006; Nowson, 2006; Oberlander and Gill, 2004b). The use of (near-) swear words can be manipulated by inserting modifiers—e.g. *'the food is darn good'*—given a high NEAR EXPLETIVES parameter value. We use 'near' expletives to avoid conflicting with the positivity associated with extravert language. References to humans can be added locally as adjunct nouns—e.g. *'the food is good pal'*—using an IN-GROUP MARKER parameter. This linguistic marker can also be interpreted as the minimization of the positive face threat according to Brown and Levinson's politeness theory (1987; 2008; 1997a). Tag questions also fulfill the same politeness function, as well as contributing to the extravert's expression of sympathy (Weaver, 1998). They can therefore be inserted automatically using a TAG QUESTION parameter, such as in *'X has good food, doesn't it?'* or *I would recommend X, wouldn't you?*. See utterance 8 in Table 15.

Lexical choice: Introverts use richer and longer words (Dewaele and Furnham, 1999; Furnham, 1990). These aspects of the speaker's vocabulary can be controlled by a LEXICON FREQUENCY parameter and a LEXICON WORD LENGTH parameter, respectively biasing the selection of content words depending on their frequency of use and their length. Compare the use of *mediocre ambiance* in utterance 2 in Table 15 with *bad atmosphere* in utterance 6. Finally, we hypothesize that extraverts produce more exaggerations—as a consequence of their enthusiasm—which results in the use of stronger verbs, e.g. by favoring *love* over *like* in the utterance *'I think you would love X'*.

3.2 Emotional stability

Emotional stability—or neuroticism—is the second most studied personality trait; it is part of most existing frameworks of personality, such as the Big Five and the PEN model (Eysenck et al., 1985; Norman, 1963). Neurotics tend to be anxious, negative and oversensitive, while emotionally stable people are calm and even-tempered. Eysenck et al. (1985) suggest that this dimension is related to activation thresholds in the nervous system, i.e. neurotics turn more easily into a 'fight-or-flight' state when facing danger, resulting in

[19] As the hedge *you know* can have many other functions, we generally consider it as a filled pause, while we model it individually when needed, e.g. for projecting agreeableness in (Mairesse, 2008).

an increase of their heart beat, muscular tension, level of sweating, etc. In order to increase the number of relevant findings, we also include studies focusing on short-lived emotions that are symptomatic of the personality trait (Watson and Clark, 1992), e.g. markers of anxiety are considered as valid markers of neuroticism.[20]

Table 16 summarizes the linguistic cues for emotional stability and the hypothesized personality models, and Table 17 provides example utterances generated using the personality models.

Content planning: Even more than introversion, neuroticism is largely associated with negativity (Pennebaker and King, 1999), which can thus be controlled by the same polarity parameters—i.e. CONTENT POLARITY, CONCESSION POLARITY and REPETITION POLARITY. See Table 16 and note how utterances 1–2 and 5–6 in Table 17 include primarily negative and neutral content with negative content repeated and foregrounded. Neurotics also produce more lexical repetitions (Scherer, 1981), with a lower type-token ratio (Gill and Oberlander, 2003),[21] which we model with a high REPETITIONS parameter value. Additionally, we hypothesize that their overall lack of control makes neurotics more likely to present a positive claim first in their utterances—i.e. a high POSITIVE CONTENT FIRST parameter value, while more stable individuals would finish their utterances with more positive content to have a higher argumentative impact. Following the same assumption, we associate neuroticism with a high POLARIZATION parameter value, i.e. the production of more extreme content (regardless of whether it is positive or negative). Finally, we hypothesize that anxiety can be projected in the information presentation domain through explicit requests for confirmation as well as request rejections, producing utterances beginning with *'I'm not sure... did you say X?'* for example. The insertion of these markers is respectively controlled by the REQUEST CONFIRMATION and INITIAL REJECTION parameters.

Syntactic Structural Template Selection: Studies consistently show that neurotics produce more self-references (Mehl et al., 2006; Oberlander and Gill, 2004b; Pennebaker and King, 1999). Thus, neuroticism can be conveyed using a SELF-REFERENCES parameter that biases the template selection process by favoring templates with first-person pronouns, such as in the template *'I am sure you would like X'*. Furthermore, as with extraversion, polarity can also be expressed through template selection. Neuroticism can thus be projected by selecting negatively-connotated templates, with a low TEMPLATE POLARITY parameter value.

Aggregation: Emotionally stable people were shown to produce more *which* pronouns in their emails, while neurotics prefer the conjunction *and* (Oberlander and Gill, 2004a,b). The former preference can be modeled by the RELATIVE CLAUSE aggregation operation, such as in the utterance *'X, which has good food, has nice service'*. The production of conjunctions can be controlled by the MERGE operation, which combines propositions together by grouping their objects with a conjunction, e.g. *'X has good food and nice service'*. Siegman (1978) reports that emotionally stable speakers produce more short unfilled pauses, while anxious speakers produce longer pauses. Interestingly, these speech cues can be controlled at the aggregation level: short pauses can be realized textually by separating propositions with commas using the CONJUNCTION aggregation operation, while long pauses are conveyed by leaving propositions in separate sentences using the PERIOD operation. Oberlander and Gill (2004a) show that neurotics avoid using the *so* and *also* cue words, while they produce more inclusive words—e.g. *with*, *and*—as well as more occurrences of the adverb *though*. These specific cues can be used for combining, justifying and conceding information using the corresponding aggregation parameters (e.g. CONCEDE - BUT/THOUGH CUE WORD).[22] Finally, Scherer (1981) reports that neurotics are more likely to omit words in their speech. Such disfluencies can be reproduced in the information presentation domain by partially repeating a proposition with ellipsis dots, e.g. *'X has ... it has good food'*. This linguistic behavior is controlled at the aggregation level by repeating content using the OBJECT ELLIPSIS aggregation operation.[23]

[20]The term 'anxiety' is sometimes used to describe either an emotion or a permanent trait, the former is then referred to as *state anxiety* and the latter as *trait anxiety*.

[21]The type-token ratio is the number of unique words divided by the total number of words.

[22]Aggregation parameter names are prefixed with the rhetorical relation they realize.

[23]Ellipsis dots are only produced in restated content to make sure that the generator conveys all the information specified in the input.

Table 16: Summary of language cues for emotional stability, with corresponding generation parameters. One asterisk indicates an hypothesis, rather than a result. Two asterisks indicate a marker of a related emotion (e.g. anxiety). An unreferenced asterisk indicates a new hypothesis. Aggregation parameter names are prefixed with the rhetorical relation they realize.

Neurotic findings	Stable findings	Ref	Parameters	Neuro	Emot
Content planning:					
Problem talk, dissatisfaction	Pleasure talk, agreement, compliment	3	CONTENT POLARITY	low	high
			REPETITION POLARITY	low	high
			CONCESSION POLARITY	low	high
Direct claim	Inferred claim	*	POSITIVE CONTENT FIRST	high	low
High verbal productivity	Low verbal productivity	15	VERBOSITY	high	low
Many lexical repetitions	Few lexical repetitions	9,16	REPETITIONS	high	low
Polarised content	Neutral content	*	POLARIZATION	high	low
Stressed	Calm	*	REQUEST CONFIRMATION	low	high
			INITIAL REJECTION	high	low
Syntactic Structural Template selection:					
Many self-references	Few self-references	3,6,9	SELF-REFERENCES	high	low
Problem talk	Pleasure talk	3	TEMPLATE POLARITY	low	high
Aggregation:					
Low use of 'punct *which*'	High use of 'punct *which*'	9	RELATIVE CLAUSE	low	high
Many conjunctions	Few conjunctions	8	MERGE	high	low
Few short silent pauses	Many short silent pauses	15	CONJUNCTION	low	high
Low use of 'punct *so*'	High use of 'punct *so*'	9	JUSTIFY - SO CUE WORD	low	high
Low use of clause final *also*	High use of clause final *also*	9	INFER - ALSO CUE WORD	low	high
Many inclusive words (e.g. *with*, *and*)	Few inclusive words	9,17	WITH CUE WORD	high	low
High use of final *though*	Low use of final *though*	8	CONCEDE - BUT/THOUGH CUE WORD	high	low
Many long silent pauses	Few long silent pauses	15	PERIOD	high	low
Many 'non-ah' disfluencies (omission)	Few 'non-ah' disfluencies	16**	RESTATE - OBJECT ELLIPSIS	high	low
Pragmatic marker insertion:					
Many pronouns, few articles	Few pronouns, many articles	3,8	SUBJECT IMPLICITNESS	low	high
			PRONOMINALIZATION	high	low
Few tentative words	Many tentative words	6	SOFTENER HEDGES: ·SORT OF, SOMEWHAT, QUITE, RATHER, IT SEEMS THAT,IT SEEMS TO ME THAT, KIND OF	low	high
Many self-reference	Few self-references	3,6,9	·I THINK THAT	high	low
Many filled pauses (apprehensive)	Few filled pauses	2,10	FILLED PAUSES: · ERR, I MEAN, MMHM, LIKE	high	low
			ACKNOWLEDGMENTS:		
More aquiescence	Few aquiescence	10	·YEAH, RIGHT, OK	high	low
Many self references	Few self references	3,6,9	·I SEE	high	low
High use of 'punct *well*'	Low use of 'punct *well*'	9	·WELL	high	low
Exaggeration	Realism	*	EMPHASIZER HEDGES:		
			·REALLY, ACTUALLY,	high	low
			·BASICALLY, JUST	low	high
Many rhetorical interrogatives	Few rhetorical interrogatives	*	TAG QUESTION	high	low
Frustration	Less frustration	10	EXPLETIVES	high	low
Many 'non-ah' disfluencies (stutter)	Few 'non-ah' disfluencies	16**	STUTTERING	high	low
Lexical choice:					
Many frequent words	Few frequent words	9,17	LEXICON FREQUENCY	high	low
High concreteness	Low concreteness	9,17			
Exaggeration	Realism	*	VERB STRENGTH	high	low

Pragmatic marker insertion: Emotionally stable people produce more articles and fewer pronouns (Oberlander and Gill, 2004a), so emotional stability is associated with a high SUBJECT IMPLICITNESS parameter value (which favours attributes preceded by an article over proper nouns, e.g. '*the food is good*'), and a low PRONOMINALIZATION value in Table 16. Interestingly, neuroticism is linked to the avoidance of tentative words (e.g. *maybe*, *rather*, *kind of*) (Mehl et al., 2006), which are controlled by the SOFTENER HEDGES parameters. This suggests that neurotics tend not to express their anxiety by verbalizing their uncertainties. However, their preference for self-references (Pennebaker and King, 1999) suggests a high use

Table 17: Example outputs of PERSONAGE for the **emotional stability** personality model. Personality ratings are on a scale from 1 to 7, with 1 = very low and 7 = very high.

#	Content plan	End	PERSONAGE's output	Score
1	**compare(Acacia, Marinella)**	**low**	I might be wrong. I might approve Acacia and Marinella. Err... Acacia provides like, acceptable food. It's in Midtown! It's a bistro and french place. Actually, I mean, Marinella is in Manhattan and an it-it-italian place.	3.50
2	**compare(Caffe Cielo, Trattoria Spaghetto)**	**low**	I am not sure! Actually, I might recommend Caffe Cielo and Trattoria Spaghetto. Caffe Cielo, which has poor atmosphere, bad atmosphere, even if it features like, nice food, is an italian re-restaurant. Trattoria Spaghetto is an it-italian place. I mean, it provides... it offers really bad atmosphere.	3.75
3	**compare(Acacia, Marinella)**	**high**	Let's see, Acacia and Marinella... I guess Acacia offers sort of decent food. Basically, Marinella, however, just has quite adequate food.	6.25
4	**compare(Caffe Cielo, Trattoria Spaghetto)**	**high**	Let's see, Caffe Cielo and Trattoria Spaghetto... Caffe Cielo offers kind of acceptable food and Trattoria Spaghetto just has sort of satisfying food. Basically, I guess they're outstanding restaurants.	5.75
5	**recommend(Cent anni)**	**low**	I am not really sure. Cent'anni is the only restaurant I would recommend. It's an italian place. It offers bad at-at-atmosphere, but it features like, nice waiters, though. It provides good food. I mean, it's bloody expensive. Err... its price is 45 dollars.	3.50
6	**recommend(Chimichurri Grill)**	**low**	I am not sure! I mean, Ch-Chimichurri Grill is the only place I would recommend. It's a latin american place. Err... its price is... it's damn ex-expensive, but it pr-pr-provides like, adequate food, though. It offers bad atmosphere, even if it features nice waiters.	4.00
7	**recommend(Cent anni)**	**high**	Did you say Cent'anni? Basically, it seems to me that it's the best because the staff is somewhat quite adequate, also this eating house offers kind of tasty food.	5.75
8	**recommend(Chimichurri Grill)**	**high**	Let's see what we can find on Chimichurri Grill. Basically, it's the best.	6.00

of subordination hedges, such as *I think that*. The literature also shows that neurotics produce more filled pauses (Scherer, 1979; Weaver, 1998), thus a neurotic generator requires high values for the FILLED PAUSES parameters (e.g. *err*, *mmhm*, *I mean*). Scherer (1981) reports that anxiety is also associated with 'non-ah' disfluencies, i.e. alterations of the intended lexical and phrasal output. While word omissions are modeled at the aggregation level, the repetition of syllables within words is modeled with a high STUTTERING parameter, as seen in utterances 1–4 and 9–12 in Table 17. Weaver (1998) shows that neuroticism is associated with frustration and acquiescence, which we model respectively with high EXPLETIVES and ACKNOWLEDGMENTS parameter values (e.g. *bloody,damn* in utterances 9 and 10 in Table 17). This last finding is confirmed by Oberlander and Gill (2004b), who found that neurotics use the sentence-initial *well* more often. Furthermore, we hypothesize that neurotics are more likely to exaggerate when presenting information, based on the impulsiveness facet of that trait. They are thus associated with high EMPHASIZER HEDGES parameter values (e.g. *really*, *actually*). Finally, we assume that neurotics use rhetorical questions to reduce their anxiety by seeking agreement, which we model with a high TAG QUESTION parameter value.

Lexical choice: Neurotics use more frequent and concrete words in their emails (Gill and Oberlander, 2003; Oberlander and Gill, 2004b), thus yielding a higher LEXICON FREQUENCY parameter value. As with extraversion, we do not know of any study focusing on the strength of the vocabulary used. We associate neuroticism with the use of stronger verbs—i.e. a higher VERB STRENGTH parameter value—based on the exaggeration hypothesis made in the previous paragraph.

4 Evaluation Experimental Design

As we discussed above, there has been considerable prior work on the linguistic expression of stylistic effects (Bouayad-Agha et al., 2000; DiMarco and Hirst, 1993; Hovy, 1988; Isard et al., 2006; Paiva and Evans, 2005; Power et al., 2003). However, there have been relatively fewer evaluations of whether humans perceive the variation as the system intended (Brockmann, 2009; Cahn, 1990; Cassell and Bickmore, 2003; Fleischman and Hovy, 2002; Porayska-Pomsta and Mellish, 2004; Rambow et al., 2001). Since the expressive effect of linguistic variation—e.g. style, emotion, mood and personality—can only be measured subjectively, an advantage of the Big Five framework is its standard questionnaires for testing the perception of personality (Costa and McCrae, 1992; Gosling et al., 2003; John et al., 1991).

Our evaluation of PERSONAGE exploits these questionnaires by asking human judges to rate the personality of a set of generated utterances by completing the Ten-Item Personality Inventory (TIPI) (Gosling et al., 2003). The TIPI instrument minimizes the number of judgments required to elicit personality ratings. To test whether personality can be recognized from a small sample of linguistic output, and to localize the effect of varying particular parameters, the judges evaluated the speaker's personality on the basis of a *single* utterance, i.e. ignoring personality perceptions that could emerge over the course of a dialogue. The judges rated the utterances as if they had been uttered by a friend responding in a dialogue to a request to recommend restaurants. In addition, the judges evaluated the naturalness of each utterance on the same scale. Naturalness was defined in the experimental instructions to mean how likely an utterance is to have been uttered by a real person.

The judges were researchers in psychology, history and anthropology who were familiarized with Big Five trait theory by being provided with associated lists of trait adjectives from Goldberg (1990), as exemplified by the adjectives shown with each trait in Table 2. Because of the high number of control parameters, a large number of utterances was needed to reveal any significance. We thus restricted the number of judges to three in order to ensure consistency over our dataset. The judges were not familiar with language generation engines, nor were they given any information about which linguistic reflexes are associated with different traits. The judges were alone in their own offices when they produced the ratings via the online TIPI, and they did not know each other or discuss their intuitions or judgments with one another. It took the judges approximately 10 to 14 hours, over several weeks of elapsed time to complete the TIPI for all utterances. The judgments from the three judges were averaged for each utterance to produce a rating for each trait ranging from 1 (e.g. highly neurotic) to 7 (e.g. very stable).

Our main hypothesis is that the personality models in Tables 14 and 16 can be used to control PERSONAGE's generation process, and the user's perception of the system's personality. There are two personality models for trait: (1) introversion and extraversion, and (2) neurotic and emotionally stable. Tables 15 and 17 provided example output utterances for each personality model.

However, if we only test utterances produced with the personality models, we cannot tell which parameters in each model are responsible for the judge's perceptions, because the same linguistic cues would be consistently used to convey a given personality. In other words, because the cues covary, it is not possible to identify which cue—or utterance feature—is responsible. Due to the high cost of collecting the personality judgments (each utterance has 11 associated questions), and the large number of parameters, it is not possible to systematically vary each parameter. Therefore, in our evaluation, we combine a sample of utterances generated with random parameter settings with utterances generated using the personality models, and then examine correlations between generation decisions and personality ratings on the random sample. This evaluation method was chosen because of its similarity with a large range of existing correlational studies between personalty and human language (Furnham, 1990; Pennebaker and King, 1999; Scherer, 1979).

Because extraversion is the most important of the Big Five traits (Goldberg, 1990), three judges evaluated PERSONAGE in a first experiment focusing strictly on that trait (Mairesse and Walker, 2007). After positive results were obtained for extraversion, two judges evaluated the four remaining traits in a second experiment. For clarity and brevity, results for both experiments are reported together, and we validate our approach by reporting only the results for extraversion and emotional stability. The interested reader is referred to Mairesse (2008) for similar results and analysis for the other three traits.

The judges rated a total of 240 utterances based on *personality models* (i.e., predefined parameter values based on Tables 14 and 16), and 320 *random* utterances generated with uniformly distributed parameter values.[24] The personality models parameter values were normally distributed with a 15% standard deviation to increase the range of variation for a given trait. There were 80 utterances generated using personality models for the extraversion experiment and 160 with personality models for the evaluation of the other four traits. Utterances were grouped into 20 sets of utterances generated from the same content plan. Each set contained two utterances per trait (four for extraversion), generated with parameter settings for both the low end and the high end of each dimension, and four random utterances. In each set, the personality model utterances were randomly ordered and mixed with utterances generated with random parameters. The judges rated one randomly ordered set at a time, but viewed all utterances in that set before rating them. All questionnaires were filled online. A total of 40 utterances were rated for each trait (80 for extraversion),

[24]The 320 random utterances were rated for extraversion, and half of them were also rated for the remaining traits.

with each half targeting one extreme of the dimension.

5 Experimental Results

Section 5.1 reports results showing that the judges agree significantly on their perceptions and that the personality models are perceived as intended. Section 5.2 presents the correlations between parameters used to generate the random utterances and judge's ratings, in order to test generalizations from other genres. Section 5.3 reports our evaluation of the naturalness of the generated utterances, and examines which generation parameters lead to unnatural utterances.

5.1 Personality perceptions

Table 18 compares the *inter-rater agreement* as the average correlations between the judges' ratings, showing that the judges agree significantly for both traits. Column **Personality models** provides correlations for the 40 utterances generated with personality models (80 for extraversion), while column **Random** shows the ratings of 160 random utterances (320 for extraversion). The agreement for the personality models for extraversion and emotional stability ($r = .73$ and $r = .67$) is very high. Although the use of a correlational analysis for evaluating inter-rater agreement is prone to biases, this allows us to compare our results with previous studies by Mehl et al. (2006), who report a correlation of $r = .84$ for personality perceptions based on human-human conversations. The similarity between both results is encouraging considering that the judgements presented here are based on a single utterance rather than audio conversation extracts collected over 48 hours.

Table 18: Average inter-rater correlation for the rule-based and random utterances. Correlations under the *All* column were computed over the full dataset.[24] All correlations are significant at the $p < .05$ level (two-tailed).

Parameter set	**Personality models**	**Random**	**All**
Extraversion	.73	.30	.48
Emotional stability	.67	.33	.39

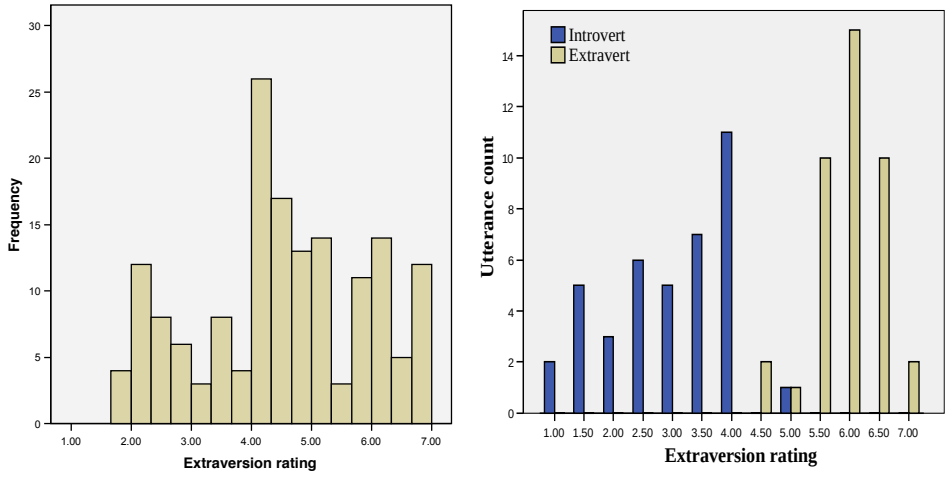
Table 18 shows that the judges also agree significantly on the personality of the random utterances ($p < .05$, two-tailed). Unsurprisingly, the agreement is lower than for the utterances produced by the personality models. Random generation decisions are more likely to produce utterances projecting inconsistent personality cues, which may be interpreted in different ways by the judges. An example of inconsistency can be found in the utterance *'Err... I am sure you would like Chanpen Thai!'*, as it expresses markers of both introversion (filled pause) and extraversion (exclamation mark).

Table 19: Average personality ratings for the utterances generated with the low and high personality models for each trait on a scale from 1 to 7. The ratings of the two extreme utterance sets differ significantly for all traits ($p < .001$, two-tailed).

Personality trait	**Low**	**High**
Extraversion	2.96	5.98
Emotional stability	3.29	5.96

Table 19 compares the average ratings of the 20 utterances expressing the low end of each trait and the 20 utterances expressing the high end (40 for extraversion). Paired t-tests show that the judges can discriminate between both extreme utterance sets for each trait ($p < .001$), e.g. utterances predicted to be perceived as introvert received an average rating of 2.96 out of 7, but utterances predicted to be perceived as extravert received an average rating of 5.98 (difference of 3.02). This difference can also be observed by comparing the distributions of the introvert and extravert utterances in Fig. 10(b). Comparing the distributions for the utterances generated with the personality models in Fig. 10(b) and Fig. 11(b) with typical ratings'

[23] Correlations over the full dataset (All) include cross-trait judgements such as extraversion ratings for utterances with neurotic parameters.



(a) Extraversion: Random (b) Extraversion: Personality Model

Fig. 10: Rating distributions for extraversion for random generation (160 utterances) vs. personality-model controlled generation (40 utterances). Ratings are averaged over all judges and rounded to the nearest half-integer.

distributions for random utterances in Fig. 10(a) and Fig. 11(a), clearly indicate that the personality models are having the desired effect on PERSONAGE's output. The ratings for emotional stability in Table 19 also shows that emotional stability is a highly recognisable trait, with a mean rating difference of 2.67 between neurotic and stable utterances.

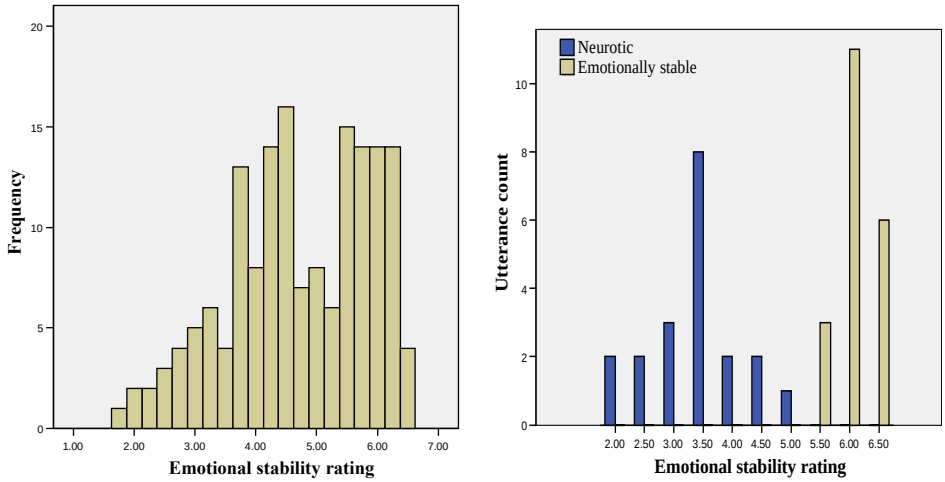
We can also compute *generation accuracy* by splitting ratings into two bins around the neutral rating (4 out of 7), and counting the percentage of utterances with an average rating falling in the bin predicted by its personality model. Since personality models aim to produce utterances manifesting extreme traits, neutral ratings count as errors. Table 20 summarizes generation accuracies for both traits, showing that PERSONAGE produces an average accuracy of 85%, i.e. a large majority of the utterances were recognized correctly.

Table 20: Generation accuracy (in %) for the utterance sets generated with the low and high parameter settings for each trait. An utterance is correctly recognized if its average rating falls in the half of the scale predicted by its parameter setting. Neutral ratings (4 out of 7) are counted as misrecognitions.

Personality trait	**Low**	**High**	**Overall**
Extraversion	82.5	100.0	91.3
Emotional stability	80.0	100.0	90.0
All utterances		85.0	

Fig. 10(b) shows that extravert utterances were all recognized as such, with approximately normally distributed ratings at the top end of the extraversion scale, while 17.5% of the introvert utterances were rated as neutral or extravert. Extraversion is the easiest trait to project in our domain, with ratings covering the full range of the scale and an overall accuracy of 91.3% over both utterance sets (See Table 20). Fig. 11(b) shows that PERSONAGE did not generate utterances perceived as extremely neurotic by all judges, as no utterance were rated below 2 out of 7 on that scale. Also, while all emotionally stable utterances were perceived correctly, 20% of the neurotic utterances were rated as neutral or moderately stable: the ratings' distribution of neurotic utterances is slightly biased towards the positive end of the scale.

Table 20 shows that the positive ends of both dimensions are modeled with high precision, while parameter settings for the low ends—typically associated with a low desirability—produce more misrecognized utterances. This overall trend can be explained by a bias of the judges towards the positive end, as suggested by the overall distributions of ratings. It could also be a consequence of a bias in PERSONAGE's predefined parameter settings, that could be attenuated by recalibrating the parameter values. Finally, it is likely that some aspects of personality cannot be conveyed through language only, or that more than a single utterance is required.



(a) Emotional stability: Random (b) Emotional stability: Personality Model

Fig. 11: Rating distributions for emotional stability for random generation (160 utterances) vs. personality-model controlled generation (40 utterances). Ratings are averaged over all judges and rounded to the nearest half-integer.

5.2 Correlational analysis of random utterances

As discussed above, the utterances generated from personality models do not allow us to understand which generation decisions are responsible for the judge's ratings. In other words, because the same linguistic cues are consistently used to convey a given personality, it is not possible to identify which cue—or utterance feature—is responsible for observed discrepancies between the target personality and the judges' ratings. Thus we apply the same method as used in psycholinguistic studies: we test correlations between linguistic markers that can be counted and the Big Five Traits. Tables 21 and 22 present the correlations for the random utterances between the average judges' ratings and generation decisions for the extraversion and emotional stability traits.[24] Generation decision features are labeled with the parameter's name prefixed with its component in the NLG architecture. The correlations indicate that some generation decisions have higher impact (magnitude of r). The values of `yes` and `no` in the Prediction **Pred** column indicate which hypotheses are confirmed, i.e. which predictions from other language genre carry over to our domain. Interestingly, some results also contradict our hypotheses, as indicated by `opp` in the **Pred** columns in Tables 21 and 22. It is important to note that the reported level of significance is potentially overestimated due to the large number of significance tests performed. While there exist methods for adjusting the significance level accordingly (e.g., Bonferroni correction), we choose to report the unadjusted significance level in order to provide a fair comparison with the psychology studies in Table 25 in the appendix.

While most correlation would appear to be quite low, they are in the same range as personality studies on human language production (Pennebaker and King, 1999), which reflects the fact that perception of personality through language is a non-trivial task, especially in the absence of acoustic cues. Nevertheless, we find that many parameters correlate significantly with personality. Table 21 shows that exclamation marks are the strongest indicators of extraversion, with a correlation of .34 with the average ratings. As suggested by the literature, verbosity is also associated with extraversion, however the use of the INFER rhetorical relation—joining propositions together without emphasis—produces a higher association, suggesting that extraverts do not put pieces of information into perspective.[25] Explicit confirmations are also associated with extraversion, as well as the use of conjunctions, frequent lexical items, near-expletives (e.g. *darn*), the adverb *really*, restatements, and more positive claims. Negative correlations at the bottom of Table 21

[24]The values used for the generation decision correlations are the actual decisions that were taken in each utterance rather than input parameter values.

[25]To improve readability throughout this paper, the perception of the judges regarding a personality type is referred to as a characteristic of individuals that possess that personality trait, e.g. *extravert utterances*, *extravert speakers* and *extraverts* are used interchangeably.

Table 21: Correlations between generation decision features and average extraversion ratings at the $p < .1$ level (* = $p < .05$, ** = $p < .01$). The *Pred* column indicates whether the relation was predicted by the psychology findings reviewed in Section 3 (*opp* = predicted opposite relation). Because extraversion ratings were collected over two experiments, with different generation parameters, generation decisions that were implemented in both experiments produce higher significance levels.

Generation decision features	r_{extra}	**Pred**
Pragmatic marker - exclamation	0.34**	yes
Aggregation - infer	0.21**	no
Content planning - verbosity	0.19**	yes
Content planning - request confirmation: you want to know	0.16**	yes
Content planning - request confirmation: did you say	0.16*	yes
Aggregation - justify - since ns	0.16	no
Aggregation - conjunction	0.16**	yes
Lexical choice - lexicon frequency	0.15*	yes
Pragmatic marker - near expletives	0.15	yes
Content planning - syntactic complexity	0.15**	opp
Pragmatic marker - emphasizer: really	0.14*	yes
Content planning - request confirmation	0.14*	yes
Aggregation - restate - conjunction with comma	0.13*	yes
Aggregation - infer - period	0.13*	opp
Content planning - restatements	0.13*	yes
Aggregation - restate	0.12*	yes
Aggregation - infer - conjunction	0.12*	yes
Content planning - repeated positive content	0.12*	yes
Aggregation - period	0.12*	yes
Content planning - template polarity	0.12*	yes
Content planning - repetition polarity	0.12*	yes
Aggregation - infer - merge	0.11	no
Pragmatic marker - softener: kind of	-0.10	yes
Pragmatic marker - softener: rather	-0.11*	yes
Pragmatic marker - softener: like	-0.11*	yes
Content planning - initial rejection	-0.18*	no
Pragmatic marker - filled pause: err	-0.23**	yes

indicate markers of introversion. The filled pause *err* is the strongest indicator of introversion, with a correlation of $-.23$. Introverts are also perceived as producing initial rejections, as well as hedges such as *like* and *rather*.

Table 22 provides correlations between generation decisions and emotional stability. The correlations indicate that neuroticism is associated with the use of short, frequent words ($r = .25$ and $r = -.28$). Neurotics use the discourse connective *so* to express justifications, while *since* is associated with stable speakers. Interestingly, in-group markers indicate stability as well (especially *pal*), while filled pauses (i.e. *err*) and repetitions indicate neuroticism. As in other genres, negative content and swear words are also associated with a lack of stability, with a stronger association for the expletive *damn* ($r = -.21$).

This correlational analysis provides insight into which generation parameters help the judges to discriminate between various traits. The knowledge of strong markers of personality is useful for controlling the generation process. More importantly, these correlations show clearly that the findings from other genres of language that we summarized in Table 25 in the Appendix generalize to our domain. Interestingly, we also find that many new markers emerge, while some results contradict our hypotheses (i.e. indicated by *opp* in the *Pred* columns). Future work could thus enhance Personage's rule-based approach based on the correlations presented here, by taking domain-specific information into account to refine the predefined parameter settings derived from psychological studies.

Table 22: Correlations between generation decision features and average emotional stability ratings at the $p < .1$ level (* = $p < .05$, ** = $p < .01$). The *Pred* column indicates whether the relation was predicted by the psychology findings reviewed in Section 3 (*opp* = predicted opposite relation).

Generation decision features	r_{emot}	**Pred**
Lexical choice - lexicon word length	0.25**	no
Pragmatic marker - in-group marker: pal	0.22**	no
Pragmatic marker - in-group marker	0.20**	no
Aggregation - justify - since ns	0.16*	no
Pragmatic marker - acknowledgment: yeah	0.15	opp
Content planning - content polarity	0.14	yes
Aggregation - with	-0.13	yes
Content planning - restatements	-0.15	no
Aggregation - infer - with ns	-0.15	yes
Aggregation - infer - also	-0.16	opp
Pragmatic marker - acknowledgment: ok	-0.16*	yes
Aggregation - merge	-0.16*	yes
Aggregation - concede - although ns	-0.17*	no
Content planning - repetitions	-0.18*	yes
Pragmatic marker - expletives	-0.18*	yes
Aggregation - restate - merge with comma	-0.19*	yes
Pragmatic marker - tag question	-0.19*	yes
Content planning - syntactic complexity	-0.19*	no
Content planning - negative content	-0.20*	yes
Pragmatic marker - tag question: alright	-0.21**	yes
Pragmatic marker - expletives: damn	-0.21**	yes
Pragmatic marker - filled pause: err	-0.22**	yes
Aggregation - restate	-0.23**	yes
Aggregation - justify - so sn	-0.25*	opp
Content planning - repeated negative content	-0.26**	yes
Lexical choice - lexicon frequency	-0.28**	yes

5.3 Naturalness

Judges also evaluated the naturalness of each utterance, i.e. to what extent it could have been uttered by a human. Results in Table 23 show that the utterances were seen as moderately natural on average, with a mean rating of 4.59 out of 7 for the personality model utterances. Although naturalness could possibly be improved by adding generation constraints to avoid inconsistent outputs, these results are promising.

Table 23: Average naturalness ratings for the utterance sets generated with the personality models and the random utterances.

Personality trait	**Low**	**High**	**Random**
Extraversion	4.93	5.78	4.75
Emotional stability	3.43	4.63	4.72

Table 23 also shows that extravert utterances are rated as the most natural, with an average rating above 5.5 out of 7. Introvert utterances are also perceived as natural, with ratings close to 5. On the other hand, utterances expressing neuroticism are rated as moderately unnatural, with average scores below 3.5. A comparison between Tables 20 and 23 suggests a correlation between naturalness and generation accuracy, however it is not clear whether (1) poor personality recognition is a consequence of unnatural utterances, or whether (2) the projection of inconsistent personality cues causes the low naturalness scores, or whether (3) extreme traits are likely to be perceived as unnatural because they are not commonly observed.

Naturalness ratings can also help identify the generation parameters responsible for unnatural utterances. Table 24 shows the correlations between generation decisions and average naturalness ratings of the random utterances. Results show that negative content is perceived as unnatural ($r = -.32$). Negations ($r = -.27$),

Table 24: Correlations between generation decisions and average naturalness ratings of the random utterances, at the $p < .1$ level (* = $p < .05$, ** = $p < .01$). Generation parameter names are prefixed with their component in the NLG architecture.

Generation decisions	r_{nat}
Content planning - content polarity	0.32**
Content planning - positive content	0.26**
Content planning - polarization	0.26**
Pragmatic marker - in-group marker	0.24**
Lexical choice - lexicon frequency	0.21**
Aggregation - justify - since sn	0.17*
Pragmatic marker - in-group marker: pal	0.16*
Pragmatic marker - stuttering	0.14
Aggregation - infer - merge	0.12*
Aggregation - justify - since ns	0.12*
Pragmatic marker - tag question: okay	0.11*
Content planning - syntactic complexity	0.10
Content planning - concessions	-0.10
Pragmatic marker - softener: rather	-0.10
Aggregation - justify - with ns	-0.11
Pragmatic marker - filled pause: err	-0.11
Aggregation - concede - even if ns	-0.11*
Pragmatic marker - subject implicitness	-0.13*
Pragmatic marker - tag question: you see	-0.13*
Aggregation - restate - merge with comma	-0.13*
Aggregation - concede - but/though ns	-0.13*
Aggregation - infer - period	-0.14*
Pragmatic marker - softener: somewhat	-0.14*
Pragmatic marker - pronominalization: demonstrative	-0.14*
Aggregation - justify - with ns	-0.14
Aggregation - restate - conjunction with comma	-0.14*
Pragmatic marker - softener: quite	-0.15**
Pragmatic marker - softener: sort of	-0.17**
Aggregation - concede - even if ns	-0.18
Content planning - repeated neutral content	-0.18**
Pragmatic marker - pronominalization	-0.18**
Content planning - repetitions	-0.22**
Pragmatic marker - negation	-0.27**
Content planning - negative content	-0.32**

strict repetitions ($r = -.22$) and pronouns ($r = -.18$) also affect naturalness. On the other hand, positive content ($r = .32$) and in-group markers ($r = .24$) are perceived as natural, as well as more frequent words ($r = .21$). Interestingly, stuttering also increases naturalness ($r = .14$).

6 Discussion and Conclusion

We present and evaluate PERSONAGE, a highly parameterizable generator that produces outputs that are reliably perceived by human judges as expressing Big Five personality traits. We believe that such a generation capability is a necessary step towards personality-based user adaptation. This paper makes four contributions:

1. We present a systematic review of psycholinguistic findings, organized by the NLG reference architecture;
2. We propose a mapping from these findings to generation parameters for each NLG module and a real-time implementation of a generator using these parameters.[26] Our parameters are defined in terms of well-defined operations on standard semantic and syntactic representations, which should therefore be replicable in other systems;
3. We present an evaluation experiment showing that we can use personality models based on psycholinguistic findings to control the parameters, in order to produce recognizable linguistic variation for both extraversion and emotional stability. See Mairesse (2008) for similar results for other Big Five personality dimensions;
4. We analyze the correlations between judges' ratings of personality and PERSONAGE generation decisions, showing which linguistic reflexes of personality generalize from naturally-occurring genres to our application domain.

Our evaluation shows that human judges reliably interpret PERSONAGE's personality cues. To our knowledge, the only other generation system to be evaluated in such a way is CRAG-2, a system generating movie review dialogues (Brockmann, 2009). In a first experiment, Brockmann presents human judges with dialogues combining utterances selected from an annotated corpus. Results show that the judges perceive variations between extravert and introvert utterances correctly ($p < .001$), however results for emotional stability are not significant. Interestingly, the introduction of n-gram language models for re-ranking paraphrases generated from logical forms produces non-significant results for both traits.[27] Brockmann hypothesizes that this is a consequence of the bias of n-gram language models towards shorter utterances. Although future work should investigate other data-driven methods for stylistic generation, these results suggest that controlling the target personality from within the generation process is beneficial both in terms of perceptual accuracy and efficiency.

As we discussed above, there has been considerable prior work on the linguistic expression of stylistic effects (Bouayad-Agha et al., 2000; DiMarco and Hirst, 1993; Hovy, 1988; Isard et al., 2006; Paiva and Evans, 2005; Power et al., 2003). However, to our knowledge, many of the parameters that we have systematically and replicably implemented in PERSONAGE, such as hedges, negation insertion, tag questions and polarity, have never been implemented within the standard NLG architecture. Many of our parameters are not only useful for generating language expressing personality, but could also be used for other types of affective generation, such as politeness (Gupta et al., 2008; Porayska-Pomsta and Mellish, 2004; Walker et al., 1997a; Wang et al., 2005), or formality (DiMarco and Hirst, 1993), if the appropriate models for controlling these parameters were developed. For example, hedges and tag questions can convey politeness or status (Brennan and Ohaeri, 1994, 1999; Brown and Levinson, 1987; Lakoff, 1973a,b; Prince et al., 1980).

Another novel aspect of PERSONAGE is the idea of indexing and selecting content by polarity. In every personality model we tested, content selection according to polarity has a significant effect on human perceptions of personality. This type of content selection mechanism suggests that there might be many other potential ways to index and discriminate content, in order to make different versions of a story, a play, or a tutorial, or indeed any dialogue. For example, other work implicitly distinguishes content according to how "personal" particular questions or statements might be in a conversational context (Cassell and Bickmore, 2003; Mateas and Stern, 2003), or how "threatening" a teacher's criticism might be, using ideas from politeness theory (Porayska-Pomsta and Mellish, 2004; Wang et al., 2005). Thus the idea of content

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

[27] Agreeableness is the only trait that is perceived correctly above chance level, however that trait is not evaluated in the first experiment.

that is interchangeable and selectable according to particular social or pragmatic criteria is potentially very powerful.

There are a number of issues that deserve further research. We examined only the effect of manipulations of linguistic form, and tested these manipulations by asking judges to read the generated utterances. However many findings in the literature suggest that personality affects dialogue strategy, prosody, and gesture (Scherer, 1979; Vogel and Vogel, 1986). Our approach could be extended to the parametrization of these other modules.

In addition, while we have shown that evaluative utterances in the restaurant domain can manifest personality, more research is needed to identify whether only some types of speech acts can express personality. Although PERSONAGE's parameters were implemented with domain-independence in mind, future work should assess the extent to which our results are dependent on these speech acts or on the application domain. We believe our work can be trivially extended to any tourist domain (e.g. hotels), and more generally to any domain producing evaluative utterances (e.g. film reviews). An extension to these domains would simply be a matter of keyword substitution in PERSONAGE's output. We have started to explore PERSONAGE's generalization capabilities, for interactive drama systems and personal assistants (Mairesse and Walker, 2008a; Walker et al., 1997b). Personalization is often an important technical requirement for such applications (Hayes-Roth and Brownston, 1994; Mott and Lester, 2006; Murray, 1997).

Another limitation of this work is that we treat personality as a discrete phenomenon, with personality models controlling generated utterances expressing either the low or the high end of each personality trait, and only one trait at a time. This capability can be used for dialogue system adaptation in systems supporting a limited range of user models, or other applications that do not require fine-grained variation of the generation output, e.g. artificial characters with static behavior. However, the wide range of individual differences reflected by the literature on the Big Five (Allport and Odbert, 1936; Goldberg, 1990; Norman, 1963) as well as recent work in medical research (Marcus et al., 2006) suggest that personality varies continuously. This continuity is also reflected by the continuous scales used in personality psychology instruments (Costa and McCrae, 1992; Gosling et al., 2003; John et al., 1991). In other work, we investigate methods for producing language targeting any arbitrary value on the Big Five dimensions (Mairesse and Walker, 2008b).

Additionally, our approach currently does not offer fine-grained control of the use of various pragmatic markers, but this might be needed to increase the naturalness of generated utterances. For example, previous work on the placement of cue words suggests constraints that we do not capture, such as avoiding the repetition of the same cue within a single turn (Di Eugenio et al., 1997; Moser and Moore, 1995).

Our long term goal is to be able to adapt to the user's personality and linguistic style during dialogic interaction. In other work, we have developed models and techniques for recognizing the user's personality from conversational data (Mairesse et al., 2007); these models could be used to produce a system that models similarity-attraction (Byrne and Nelson, 1965; Nass and Lee, 2001) and task-specific personality adaptation, based on the adaptation policies outlined in Section 1. In future work, we plan to use PERSONAGE to evaluate the effect of lexical, syntactic, and personality-based adaptation on various dialogue system tasks.

7 Appendix

Table 25: Psychological studies on the identification of personality markers in language. Each study is labeled with a reference number that will be used in Section 3, when the findings are mapped to generation parameters for each personality trait. An asterisk indicates a review, rather than a specific study.

Study ref	Authors	Language source	Cues	Assessment method	Personality dimensions
1	Furnham (1990)*	spoken	speech, linguistic markers	self-report	extraversion, type A behavior, self-monitoring
2	Scherer (1979)*	spoken	speech markers	self-report, emotion induction	extraversion, emotional stability, anxiety *inter alia*
3	Pennebaker and King (1999)	essays	content-analysis category counts	self-report	Big Five traits
4	Dewaele and Furnham (1999)*	spoken	various	self-report	extraversion
5	Oberlander and Gill (2006)	emails	content-analysis category and n-gram counts	self-report	extraversion, neuroticism, psychoticism
6	Mehl et al. (2006)	daily-life conversations	content-analysis category counts	observer, self-report	Big Five traits
7	Siegman and Pope (1965)	spoken	verbal fluency	self-report	extraversion
8	Oberlander and Gill (2004a)	emails	part-of-speech n-grams	self-report	extraversion, neuroticism, psychoticism
9	Oberlander and Gill (2004b)	emails	content-analysis category and n-gram counts	self-report	extraversion, neuroticism
10	Weaver (1998)	questionnaires	communicative behavior	self-report	extraversion, neuroticism, psychoticism
11	Heylighen and Dewaele (2002)	essays, oral examinations	measure of formality	self-report	extraversion
12	Nowson (2006)	blogs	content-analysis category and n-gram counts	self-report	Big Five traits
13	Cope (1969)	spoken	output size, type-token ratio	self-report	extraversion
14	Thorne (1987)	spoken	polarity, focus	self-report	extraversion
15	Siegman (1978)*	spoken	speech markers	various	socio-economic background, extraversion, anxiety, anger, *inter alia*
16	Scherer (1981)*	spoken	speech markers	various	stress, anxiety
17	Gill and Oberlander (2003)	emails	part-of-speech n-gram counts	self-report	extraversion, neuroticism
18	Infante (1995)*	spoken	communicative behavior	emotion induction	verbal aggressiveness

References

Aaker, J. L. (1999). The malleable self: the role of self-expression in persuasion. *Journal of Marketing Research*, 36(1):45–57.

Allport, G. W. and Odbert, H. S. (1936). Trait names: a psycho-lexical study. *Psychological Monographs*, 47(1, Whole No. 211):171–220.

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

Ardissono, L., Goy, A., Petrone, G., Segnan, M., and Torasso, P. (2003). Intrigue: personalized recommendation of tourist attractions for desktop and hand held devices. *Applied Artificial Intelligence*, 17(8):687–714.

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

Ball, G. and Breese, J. (1998). Emotion and personality in a conversational character. In *Proceedings of the Workshop on Embodied Conversational Characters*, pages 83–86.

Beaman, K. (1984). Coordination and subordination revisited: Syntactic complexity in spoken and written narrative discourse. In Tannen, D. and Freedle, R., editors, *Coherence in Spoken and Written Discourse*, pages 45–80. Ablex.

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

Belz, A. (2005b). Statistical generation: Three methods compared and evaluated. In *Proceedings of the 10th European Workshop on Natural Language Generation*.

Bouayad-Agha, N., Scott, D., and Power, R. (2000). Integrating content and style in documents: a case study of patient information leaflets. *Information Design Journal*, 9(2-3):161–176.

Bouchard, T. J. and McGue, M. (2003). Genetic and environmental influences on human psychological differences. *Journal of Neurobiology*, 54:4–45.

Brennan, S. and Ohaeri, J. (1994). Effects of message style on users' attributions toward agents. *Conference on Human Factors in Computing Systems*, pages 281–282.

Brennan, S. and Ohaeri, J. (1999). Why do electronic conversations seem less polite? the costs and benefits of hedging. *Proceedings of the international joint conference on Work activities coordination and collaboration*, pages 227–235.

Brennan, S. E. (1991). Conversations with and through computers. *User Modeling and User-Adapted Interaction*, 1:67–86.

Brennan, S. E. (1996). Lexical entrainment in spontaneous dialog. In *Proceedings of the International Symposium on Spoken Dialogue*, pages 41–44.

Brennan, S. E. and Clark, H. H. (1996). Lexical choice and conceptual pacts in conversation. *Journal of Experimental Psychology: Learning, Memory And Cognition*.

Brockmann, C. (2009). *Personality and Alignment Processes in Dialogue: Towards a Lexically-Based Unified Model*. PhD thesis, University of Edinburgh, School of Informatics.

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

Byrne, D. and Nelson, D. (1965). Attraction as a linear function of proportion of positive reinforcements. *Journal of Personality and Social Psychology*, 1:659–663.

Cahn, J. E. (1990). The Generation of Affect in Synthesized Speech. *Journal of the American Voice I/O Society*, 8:1–19.

Carberry, S. (1989). Plan recognition and its use in understanding dialogue. In Kobsa, A. and Wahlster, W., editors, *User Models in Dialogue Systems*, pages 133–162. Springer Verlag, Berlin.

Carenini, G. and Moore, J. D. (2000). A strategy for generating evaluative arguments. In *Proceedings of International Conference on Natural Language Generation*, pages 47–54, Mitzpe Ramon, Israel.

Carenini, G. and Moore, J. D. (2006). Generating and Evaluating Evaluative Arguments. *Artificial Intelligence Journal*.

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

Chklovski, T. and Pantel, P. (2004). VerbOcean: Mining the web for fine-grained semantic verb relations. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, Barcelona.

Cohen, P. R., Perrault, C. R., and 1982, J. F. A. (1982). Beyond question answering. In Lehnert, W. and Ringle, M., editors, *Strategies for Natural Language Processing*, pages 245–274. Lawrence Erlbaum Ass. Inc, Hillsdale, N.J.

Cope, C. (1969). Linguistic structure and personality development. *Journal of Counselling Psychology*, 16:1–19.

Costa, P. T. and McCrae, R. R. (1992). *NEO PI-R Professional Manual*. Psychological Assessment Resources, Odessa, FL.

Darves, C. and Oviatt, S. (2002). Adaptation of users' spoken dialogue patterns in a conversational interface. In *Proceedings of the International Conference on Spoken Language Processing (ICSLP'2002)*.

Department of the Army (2006). *Police intelligence operations. Field Manual FM 3-19.50. Appendix D: Tactical Questioning*.

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

Di Eugenio, B., Moore, J. D., and Paolucci, M. (1997). Learning features that predict cue usage. In *Proceedings of the 35th Annual Meeting of the ACL, ACL/EACL 97*, pages 80–87.

DiMarco, C. and Hirst, G. (1993). A computational theory of goal-directed style in syntax. *Computational Liguistics*, 19(3):451–499.

Dunn, G., Wiersema, J., Ham, J., and Aroyo, L. (2009). Evaluating interface variants on personality acquisition for recommender systems. In *Proceedings of the International Conference on User Modeling, Adaptation, and Personalization*.

Eysenck, S. B. G., Eysenck, H. J., and Barrett, P. (1985). A revised version of the psychoticism scale. *Personality and Individual Differences*, 6(1):21–29.

Fellbaum, C. (1998). *WordNet: An Electronic Lexical Database*. MIT Press, Cambridge, MA.

Fennis, B. M. and Pruyn, A. T. H. (2007). You are what you wear: Brand personality influences on consumer impression formation. *Journal of Business Research*, 60(6):634–639.

Finin, T. W., Joshi, A. K., and Webber, B. L. (1986). Natural language interactions with artificial experts. *Proceedings of the IEEE*, 74(7):921–938.

Fleischman, M. and Hovy, E. (2002). Towards emotional variation in speech-based natural language generation. In *Proceedings of the 1st International Conference on Natural Language Generation*, pages 57–64.

Forbes-Riley, K. and Litman, D. (2007). Investigating Human Tutor Responses to Student Uncertainty for Adaptive System Development. *Lecture Notes in Computer Science*, 4738:678.

Forbes-Riley, K., Litman, D., and Rotaru, M. (2008). Responding to Student Uncertainty During Computer Tutoring: An Experimental Evaluation. *Lecture Notes in Computer Science*, 5091:60–69.

Furnham, A. (1990). Language and personality. In Giles, H. and Robinson, W., editors, *Handbook of Language and Social Psychology*. Winley.

Furnham, A., Jackson, C. J., and Miller, T. (1999). Personality, learning style and work performance. *Personality and Individual Differences*, 27:1113–1122.

Giles, H., Coupland, N., and Coupland, J. (1991). 1. Accommodation theory: Communication, context, and consequence. *Contexts of accommodation: Developments in applied sociolinguistics*, page 1.

Gill, A. and Oberlander, J. (2003). Perception of e-mail personality at zero-acquaintance: Extraversion takes care of itself; neuroticism is a worry. In *Proceedings of the 25th Annual Conference of the Cognitive Science Society*, pages 456–461.

Goldberg, L. R. (1990). An alternative "description of personality": The Big-Five factor structure. *Journal of Personality and Social Psychology*, 59:1216–1229.

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

Green, S. J. and DiMarco, C. (1993). Stylistic decision-making in natural language generation. In *Proceedings of the 4th European Workshop on Natural Language Generation*.

Grosz, B. J. (1983). Team: A transportable natural language interface system. In *Proc. 1st Applied ACL, Association for Computational Linguistics, Santa Monica, Ca.*

Gupta, S., Walker, M. A., and Romano, D. M. (2007). How rude are you?: Evaluating politeness and affect in interaction. In *Proceedings of ACII*, pages 203–217.

Gupta, S., Walker, M. A., and Romano, D. M. (2008). Polly: A conversational system that uses a shared, representation to generate action and social language. In *IJCNLP 2008, The Third International Joint Conference on Natural Language Processing*, pages 203–217.

Hayes-Roth, B. and Brownston, L. (1994). Multiagent collaboration in directed improvisation. Technical Report KSL 94-69, Knowledge Systems Laboratory, Stanford University.

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

Higashinaka, R., Walker, M. A., and Prasad, R. (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).

Hirschberg, J. (2008). Speaking more like you: Lexical, acoustic/prosodic, and discourse entrainment in spoken dialogue systems. In *Proceedings of the 8th SIGdial Workshop on Discourse and Dialogue*, page 128.

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

Hu, M. and Liu, B. (2004). Mining and summarizing customer reviews. In *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*.

Hubal, R. C., Kizakevich, P. N., Guinn, C. I., Merino, K. D., and West, S. L. (2000). The virtual standardized patient: Simulated patient-practitioner dialogue for patient interview training. In Westwood, J. D., Hoffman, H. M., Mogel, G. T., Robb, R. A., and Stredney, D., editors, *Envisioning Healing: Interactive Technology and the Patient-Practitioner Dialogue*. IOS Press, Amsterdam.

Infante, D. A. (1995). Teaching students to understand and control verbal aggression. *Communication Education*, 44(1):51–63.

Inkpen, D. Z. and Hirst, G. (2004). Near-synonym choice in natural language generation. In Nicolas Nicolov, Kalina Bontcheva, G. A. and Mitkov, R., editors, *Recent Advances in Natural Language Processing III*. John Benjamins Publishing Company.

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

Isbister, K. and Nass, C. (2000). Consistency of personality in interactive characters: verbal cues, non-verbal cues, and user characteristics. *International Journal of Human-Computer Studies*, 53(2):251 – 267.

John, O. P., Donahue, E. M., and Kentle, R. L. (1991). The "Big Five" Inventory: Versions 4a and 5b. Technical report, Berkeley: University of California, Institute of Personality and Social Research.

John, O. P. and Srivastava, S. (1999). The Big Five trait taxonomy: History, measurement, and theoretical perspectives. In Pervin, L. A. and John, O. P., editors, *Handbook of personality theory and research*. New York: Guilford Press.

Johnson, L., Mayer, R., André, E., and Rehm, M. (2005). Cross-cultural evaluation of politeness in tactics for pedagogical agents. In *Proceedings of the 12th International Conference on Artificial Intelligence in Education (AIED)*.

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

Kobsa, A. and Wahlster, W., editors (1989). *User Models in Dialog Systems*. Springer Verlag, Berlin.

Lakoff, R. (1973a). Language and woman's place. *Language in society*, pages 45–80.

Lakoff, R. (1973b). The logic of politeness; or, minding your p's and q's. In *ninth regional meeting of the Chicago Linguistic Society*, volume 292, page 305.

Langkilde, I. and Knight, K. (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.

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

Lavoie, B. and Rambow, O. (1997). A fast and portable realizer for text generation systems. In *Proceedings of the 3rd Conference on Applied Natural Language Processing*, pages 265–268.

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

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

Levelt, W. J. M. and Kelter, S. (1982). Surface form and memory in question answering. *Cognitive Psychology*, 14:78–106.

Lin, J. (2006). Using distributional similarity to identify individual verb choice. In *Proceedings of the 4th International Natural Language Generation Conference*, pages 33–40, Sydney, Australia.

Litman, D. and Allen, J. (1984). A plan recognition model for subdialogues in conversation. Technical Report 141, University of Rochester.

Louchart, S., Aylett, R., Dias, J., and Paiva, A. (2005). Unscripted narrative for affectively driven characters. In *Proceedings of the First International conference on Artificial Intelligence and Interactive Digital Media (AIIDE)*.

Loyall, A. B. and Bates, J. (1995). Behavior-based language generation for believable agents. Technical Report CMU-CS-95-139, School of Computer Science, Carnegie Mellon University.

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

Mairesse, F. and Walker, M. (2008a). A personality-based framework for utterance generation in dialogue applications. In *Proceedings of the AAAI Spring Symposium on Emotion, Personality, and Social Behavior*.

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

Mairesse, F. and Walker, M. A. (2008b). Trainable generation of Big-Five personality styles through data-driven parameter estimation. In *Proceedings of the 46th Annual Meeting of the Association for Computational Linguistics (ACL)*.

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

Mann, W. C. and Thompson, S. A. (1988). Rhetorical structure theory. Toward a functional theory of text organization. *Text*, 8(3):243–281.

Marcu, D. (1996). Building up rhetorical structure trees. In *Proceedings of AAAI/IAAI 1996*, volume 2, pages 1069–1074.

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

Mateas, M. and Stern, A. (2003). Façade: An experiment in building a fully-realized interactive drama. In *Proceedings of the Game Developers Conference, Game Design track*.

McCrae, R. R. and Costa, P. T. (1987). Validation of the five-factor model of personality across instruments and observers. *Journal of Personality and Social Psychology*, 52:81–90.

Mehl, M. R., Gosling, S. D., and Pennebaker, J. W. (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.

Melčuk, I. A. (1988). *Dependency Syntax: Theory and Practice*. SUNY, Albany, New York.

Moore, J. D. and Paris, C. L. (1993). Planning text for advisory dialogues: Capturing intentional and rhetorical information. *Computational Linguistics*, 19(4).

Moser, M. G. and Moore, J. (1995). Investigating cue selection and placement in tutorial discourse. In *ACL 95*, pages 130–137.

Mott, B. and Lester, J. (2006). U-director: a decision-theoretic narrative planning architecture for storytelling environments. In *Proceedings of the fifth international joint conference on Autonomous agents and multiagent systems*.

Murray, J. (1997). *Hamlet on the Holodeck: The Future of Narrative in Cyberspace*. The Free Press, NY, USA.

Nass, C. and Lee, K. (2001). Does computer-synthesized speech manifest personality? experimental tests of recognition, similarity-attraction, and consistency-attraction. *Journal of Experimental Psychology: Applied*, 7(3):171–181.

Nenkova, A., Gravano, A., and Hirschberg, J. (2008). High frequency word entrainment in spoken dialogue. In *Proceedings of ACL-08: HLT*. Association for Computational Linguistics.

Niederhoffer, K. and Pennebaker, J. (2002). Linguistic style matching in social interaction. *Journal of Language and Social Psychology*, 21:337–360.

Norman, W. 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.

Nowson, S. (2006). *The Language of Weblogs: A study of genre and individual differences*. PhD thesis, University of Edinburgh.

Oberlander, J. and Gill, A. (2004a). Individual differences and implicit language: personality, parts-of-speech, and pervasiveness. In *Proceedings of the 26th Annual Conference of the Cognitive Science Society*, pages 1035–1040.

Oberlander, J. and Gill, A. (2004b). Language generation and personality: two dimensions, two stages, two hemispheres? In *Proceedings from the AAAI Spring Symposium on Architectures for Modeling Emotion: Cross-Disciplinary Foundations*, pages 104–111.

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

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

Paiva, D. S. and Evans, R. (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.

Pang, B., Lee, L., and Vaithyanathan, S. (2002). Thumbs up? Sentiment classification using machine learning techniques. In *Proceedings of the 2002 Conference on Empirical Methods in Natural Language Processing*, pages 79–86.

Paris, C. and Scott, D. (1994). Stylistic variation in multilingual instructions. In *The 7th International Conference on Natural Language Generation*.

Peabody, D. and Goldberg, L. R. (1989). Some determinants of factor structures from personality-trait descriptor. *Journal of Personality and Social Psychology*, 57(3):552–567.

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

Pickering, M. and Garrod, S. (2003). Towards a mechanistic theory of dialogue. *Behavioural and Brain Science*.

Pieraccini, R. and Levin, E. (1995). A learning approach to natural language understanding. In *Speech Recognition and Coding, New Advances and Trends, NATO ASI Series*, pages 139–155. Springer Verlag.

Plummer, J. T. (1984). How personality makes a difference. *Journal of Advertising Research*, 24:27–31.

Porayska-Pomsta, K. and Mellish, C. (2004). Modelling politeness in natural language generation. In *Proceedings of the 3rd International Conference on Natural Language Generation*, pages 141–150.

Power, R. (1974). *A Computer Model of Conversation*. PhD thesis, University of Edinburgh.

Power, R., Scott, D., and Bouayad-Agha, N. (2003). Generating texts with style. *Lecture notes in computer science*, pages 444–452.

Prince, E., Frader, J., and Bosk, C. (1980). On hedging in physician-physician discourse. In *AAAL Symposium on Applied Linguistics in Medicine*.

Rambow, O., Rogati, M., and Walker, M. A. (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)*.

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

Rehm, M. and Andre, E. (2008). From annotated multimodal corpora to simulated human-like behaviors. *Lecture Notes in Computer Science*, 4930:1.

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

Reiter, E. and Sripada, S. (2002). Human variation and lexical choice. *Computational Linguistics*, 28:545–553.

Reitter, D., Keller, F., and Moore, J. (2006). Computational modelling of structural priming in dialogue. *Proc. Human Language Technology conference-North American chapter of the Association for Computational Linguistics annual mtg*.

Revelle, W. (1991). Personality processes. *Annual Review of Psychology*, 46:295–328.

Riloff, E., Wiebe, J., and Phillips, W. (2005). Exploiting subjectivity classification to improve information extraction. In *Proceedings of the National Conference On Artificial Intelligence*, page 1106.

Rushton, J. P., Murray, H. G., and Erdle, S. (1987). Combining trait consistency and learning specificity approaches to personality, with illustrative data on faculty teaching performance. *Personality and Individual Differences*, 8:59–66.

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

Scherer, K. R. (1981). Vocal indicators of stress. In Darby, J., editor, *Speech evaluation in psychiatry*, pages 171–187. Grune & Stratton, New York.

Scott, D. R. and Souza, C. S. d. (1990). Getting the message across in RST-based text generation. In Dale, R., Mellish, C., and Zock, M., editors, *Current Research in Natural Language Generation*, pages 47–73. Academic Press.

Siegman, A. and Pope, B. (1965). Personality variables associated with productivity and verbal fluency in the initial interview. In *Proceedings of the 73rd Annual Conference of the American Psychological Association*.

Siegman, A. W. (1978). The telltale voice: Nonverbal messages of verbal communication. In Feldstein, S. and Siegman, A. W., editors, *Nonverbal Behavior and Communication*, chapter 7, pages 183–243. Lawrence Erlbaum Associates, Hillsdale, NJ.

Slater, M., Pertaub, D.-P., Barker, C., and Clark, D. (2004). An experimental study on fear of public speaking using a virtual environment. In *3rd International Workshop on Virtual Rehabilitation*.

Stenchikova, S. and Stent, A. (2007). Measuring adaptation between dialogs. *Proc. of the 8th SIGdial Workshop on Discourse and Dialogue*.

Stent, A., Prasad, R., and Walker, M. A. (2004). Trainable sentence planning for complex information presentation in spoken dialog systems. In *Proceedings of the 42nd Annual Meeting of the Association for Computational Linguistics (ACL)*.

Tapus, A. and Mataric, M. (2008). Socially assistive robots: The link between personality, empathy, physiological signals, and task performance. In *AAAI Spring Symposium*.

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

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

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

Walker, M. A., Cahn, J. E., and Whittaker, S. J. (1997b). Improvising linguistic style: Social and affective bases for agent personality. In *Proceedings of the 1st Conference on Automous Agents, AGENTS 97*, pages 96–105.

Walker, M. A. and Rambow, O. (2002). Spoken language generation. *Computer Speech and Language, Special Issue on Spoken Language Generation*, 16(3-4):273–281.

Walker, M. A., Rambow, O., and Rogati, M. (2002). Training a sentence planner for spoken dialogue using boosting. *Computer Speech and Language*, 16(3-4).

Walker, M. A., Stent, A., Mairesse, F., and Prasad, R. (2007). Individual and domain adaptation in sentence planning for dialogue. *Journal of Artificial Intelligence Research (JAIR)*, 30:413–456.

Walker, M. A., Whittaker, S., Stent, A., Maloor, P., Moore, J., Johnston, M., and Vasireddy, G. (2004). Generation and evaluation of user tailored responses in multimodal dialogue. *Cognitive Science*, 28(5):811–840.

Wang, N., Johnson, W. L., Mayer, R. E., Rizzo, P., Shaw, E., and Collins, H. (2005). The politeness effect: Pedagogical agents and learning gains. *Frontiers in Artificial Intelligence and Applications*, 125:686–693.

Watson, D. and Clark, L. A. (1992). On traits and temperament: General and specific factors of emotional experience and their relation to the five factor model. *Journal of Personality*, 60(2):441–76.

Weaver, J. B. (1998). Personality and self-perceptions about communication. In McCroksey, J. C., Daly, J. A., Martin, M. M., and Beatty, M. J., editors, *Communication and Personality: Trait perspectives*, chapter 4, pages 95–118. Hampton Press.

Wiebe, J. (1990). *Recognizing subjective sentences: a computational investigation of narrative text*. PhD thesis, State University of New York at Buffalo Buffalo, NY, USA.

Wilkie, J., Jacka, M. A., and Littlewood, P. J. (2005). System-initiated digressive proposals in automated human-computer telephone dialogues: the use of contrasting politeness strategies. *International Journal of Human-Computer Studies*, 62:41–71.

Wilson, T., Wiebe, J., and Hwa, R. (2004). Just how mad are you? finding strong and weak opinion clauses. In *Proc. the Nineteenth National Conference on Artificial Intelligence (AAAI-2004)*.

Zukerman, I. and Litman, D. (2001). Natural language processing and user modeling: Synergies and limitations. *User Modeling and User-Adapted Interaction*, 11(1-2):129–158.