Perceptions of Alignment and Personality in Generated Dialogue

Alastair J. Gill
University of Surrey
Guildford GU2 7XH, UK
A.Gill@surrey.ac.uk

Carsten Brockmann and **Jon Oberlander**
University of Edinburgh
Edinburgh EH8 9AB, UK
Carsten.Brockmann@gmx.net
J.Oberlander@ed.ac.uk

Abstract

Variation in language style can lead to different perceptions of the interaction, and different behaviour outcomes. Using the CRAG 2 language generation system we examine how accurately judges can perceive character personality from short, automatically generated dialogues, and how alignment (similarity between speakers) alters judge perceptions of the characters' relationship. Whilst personality perception of our dialogues is consistent with perceptions of human behaviour, we find that the introduction of alignment leads to negative perceptions of the dialogues and the interlocutors' relationship. A follow up evaluation study of the perceptions of different forms of alignment in the dialogues reveals that while similarity at polarity, topic and construction levels is viewed positively, similarity at the word level is regarded negatively. We discuss our findings in relation to the literature and in the context of dialogue systems.

1 Introduction

Personality describes characteristics which are central to human behaviour, and has implications for social interactions: It can affect performance on collaborative processes, and can increase engagement when incorporated within virtual agents (Hernault et al., 2008). In addition, personality has also been shown to influence linguistic style, both in written and spoken language (Pennebaker and King, 1999; Gill and Oberlander, 2002). Whilst individuals often possess individual styles of self-expression, such as those influenced by personality, in a conversation they may align or match the linguistic style of their partner: For example, by entraining, or converging, on a mutual vocabulary. Such alignment is associated with increased familiarity, trust, and task success (Shepard et al., 2001). People also adjust their linguistic styles when interacting with computers, and this affects their perceptions of the interaction (Porzel et al., 2006). However, when humans – or machines – are faced with a choice of matching the language of their conversational partner, this often raises a conflict: matching the language of an interlocutor may mean subduing one's own linguistic style. Better understanding these processes relating to language choice and interpersonal perception can inform our knowledge of human behaviour, but also have important implications for the design of dialogue systems and user interfaces.

In this paper, we present and evaluate novel automated natural language generation techniques, via the Critical Agent Dialogue system version 2 (CRAG 2), which enable us to generate dynamic, short-term alignment effects along with stable, long-term personality effects. We use it to investigate the following questions: Can personality be accurately judged from short, automatically generated dialogues? What are the effects of alignment between characters? How is the quality of the characters' relationship perceived? Additionally, in our evaluation study we examine perceptions of the different forms of alignment present in the dialogues, for example at the word, phrase or polarity levels. In the following we review relevant literature, before describing the CRAG 2 system and experimental method, and then presenting our results and discussion.

INLG 2012 Proceedings of the 7th International Natural Language Generation Conference, pages 40–48,
Utica, May 2012.

2 Background

Researchers from several traditions have studied aspects of similarity in dialogue, naming it: entrainment, alignment, priming, accommodation, coordination or convergence. For current purposes, we gloss over some important differences, and borrow the term 'alignment', because we will go on to adopt Pickering and Garrod's theoretical mechanisms in our system. Alignment usually means that if something has happened once in a dialogue (for instance, referring to an object as a vase), it is likely to happen again—and hence, alternatives become less likely (for instance, referring to the same object as a jug) (Pickering and Garrod, 2004). From this view, interlocutors align the representations they use in production and comprehension and the process is an automatic, labour-saving device, but there are of course limits to periods over which alignment processes operate; in corpus studies long-term adaptation predicts communicative success (Reitter, 2008). Alternative approaches view similarity as a process of negotiation leading to the establishment of common ground (Brennan and Clark, 1996), or a relatively conscious process resulting from attraction (Shepard et al., 2001). Although increased similarity (*convergence*) is generally regarded positively, it can sometimes arise during disagreement (Niederhoffer and Pennebaker, 2002), with cultural differences influencing both convergence and perceptions of others (Bortfeld and Brennan, 1997). Wizard-of-Oz studies have also shown convergence with a natural language interface (Brennan, 1996; Porzel et al., 2006).

Embodied conversational agents (Cassell et al., 2000) are implemented computer characters that exhibit multimodal behaviour; the technology can be exploited to give life to automatically generated scripted dialogues and to make them more engaging (van Deemter et al., 2008; Hernault et al., 2008). Aspects of the agents' personalities and their interests can be pre-configured and affect their dialogue strategies; the generation is template-based. A common way to describe personality is using the *Big Five* traits: Extraversion (preference for, and behavior in, social situations); Neuroticism (tendency to experience negative thoughts and feelings); Openness (reflects openness to new ideas); Agreeableness (how we tend to interact with others); and Conscientiousness (how organised and persistent we are in pursuing our goals). Relationships between personality dimensions and language use appear to be robust: For instance, in monological writing (essays and e-mails) high Extraverts use more social words, positive emotion words, and express more certainty; high Agreeableness scorers use more first person singular and positive emotion words, and fewer articles and negative emotion words (Pennebaker and King, 1999; Gill and Oberlander, 2002).

Personality can not only be projected through, but also perceived from, asynchronous textual communication. The extraversion dimension is generally perceived most accurately in a variety of contexts, while it was more difficult for raters to recognise neuroticism (Gill et al., 2006; Li and Chignell, 2010) Taking into account the difference between the language actually used by people with certain personality, and the language which others *expect* them to use, natural language generation (NLG) systems can exploit either to project personality. Perhaps the closest previous work to what we present here is the Personality Generator (PERSONAGE) (Mairesse and Walker, 2010) which mapped psychological findings relating to the personality to the components of the NLG system (e.g., content planning, sentence planning and realisation). Evaluation by human raters showed similar accuracy in perception of extraversion in the generated language compared with human-authored texts. There is evidence that computer users attribute personality to interfaces, and rate more highly those interfaces that exploit language associated with the user's own personality, and become more similar to the user over time (Isbister and Nass, 2000).

We now turn to describing our automated natural language generation techniques, implemented in CRAG 2, followed by a description of our experimental method and evaluation.

3 Generation Method

Dialogues are composed by CRAG 2, a Java program that provides a framework for generating dialogues between two computer characters discussing a movie. For more details of this system, see Brockmann (2009). Within CRAG 2, linguistic personality and alignment are modelled using the OPENNLP

CCG Library (OPENCCG) natural language realiser (White, 2006b). The realiser consults a grammar adapted to the movie review domain to allow the generation of utterances about the following topics: Action scenes, characters, dialogue, film, music, plot or special effects. The realiser also has access to a set of n-gram language models, used to compute probability scores of word sequences. The general conversational language model (LM) is based on data from the SWITCHBOARD corpus and a small corpus of movie reviews. The general LM is used for fallback probabilities, and is integrated with the personality and alignment language models (described below) using linear interpolation.

3.1 Personality Models

Language models were trained on a corpus of weblogs from authors of known personality (Nowson et al., 2005). For each personality dimension, the language data were divided up into high, medium and low bands so that the probability of a word sequence given a personality type could be derived; see Nowson et al. (2005) for further discussion of the positively skewed distribution of the openness dimension in bloggers. Each individual weblog was used 5 times, once for each dimension. The five models corresponding to the character's assigned personality are uniformly interpolated to give the final personality model, which is then combined with the general model (respective weights, 0.7 and 0.3).

3.2 Alignment via Cache Language Models

Meanwhile, alignment is modelled via cache language models (CLMs). For each utterance to be generated, a language model is computed based on the utterance that was generated immediately before it. This CLM is then combined with the personality LM. A character's propensity to align corresponds to the weight given to the CLM during this combination, and can be set to a value between 0 and 1.

3.3 Character Specification and Dialogue Generation

The characters are parameterised for their personality by specifying values (on a scale from 0 to 100) for the five dimensions: extraversion (E), neuroticism (N), agreeableness (A), conscientiousness (C) and openness (O). This parameterisation determines the extent to which utterances are weighted for their overlap with the personality generation model for each trait. Also, each character receives an agenda of topics they wish to discuss, along with polarities (POSITIVE/NEGATIVE) that indicate their opinion on each topic.

The character with the higher E score begins the dialogue, and their first topic is selected. Once an utterance has been generated, the other character is selected, and the system selects which topic should come next. This process continues until there are no topics left on the agenda of the current speaker. The system creates a simple XML representation of the character's utterance, using the specified topic and polarity. Following the method described in Foster and White (2004), the basic utterance specification is transformed, using stylesheets written in the Extensible Stylesheet Language Transformations (XSLT) language, into an OPENCCG logical form. We make use of the facility for defining optional and alternative inputs (White, 2006a) and underspecified semantics to mildly overgenerate candidate utterances.

Optional interjections (*I mean*, *you know*, *sort of*) and conversational markers (*right*, *but*, *and*, *well*) are added where appropriate given the discourse history. Using synonyms (e.g., *plot* = *story*, *comedy* = *humour*) and combining sentence types and optional expressions, up to 3000 possibilities are created per utterance, and the best candidate is chosen by the specific combination of n-gram models appropriate for dialogue history, personality and alignment.

4 Experimental Method

4.1 Participants

Data were collected from 80 participants with a variety of educational and occupational backgrounds using an online study (via the Language Experiments Portal; www.language-experiments.org). To ensure integrity of responses, submissions taking less than five minutes (five cases), or more than 45 minutes (one case) were examined in relation to the other responses before being included in the analysis. The demographics were as follows: 43 participants (54%) were native, and 37 (46%) non-native, speakers of English; 34 (42%) male, 46 (58%) fe-

Dialogue		Personality Parameter Setting					Propensity to
Type	Character	E	N	A	C	O	Align
1) High E	I	75	50	25	25	50	0
vs. Low E	II	25	50	75	75	50	0 or 0.7
2) Low E	I	25	50	25	25	50	0
vs. High E	II	75	50	75	75	50	0 or 0.7
3) High N	I	50	75	25	25	50	0
vs. Low N	II	50	25	75	75	50	0 or 0.7
4) Low N	I	50	25	25	25	50	0
vs. High N	II	50	75	75	75	50	0 or 0.7

Table 1: Dialogue type parameter settings.

male. Median age range was 25–29 (mode = 20–24). Other demographic information (right/left-handedness, area of upbringing, occupation) were collected, but are not considered here.

4.2 Materials

To be able to compare human judges' perceptions of characters demonstrating different personalities, and dialogues without and with alignment, dialogues were generated in four different dialogue types, as shown in Table 1. Each dialogue type sets the two computer characters to opposing extremes on either the E or the N dimension, while keeping the respective other dimension at a middle, or neutral, level (for example, in Dialogue Type 1, Character I is High E, Character II is Low E, and both characters are Mid N). Furthermore, Character I is always Low A and C, and Character II is always High A and C. All characters are set to Mid O.

Two dialogues were generated per type, giving a total of 8 dialogues, with aligning versions of each of these dialogues subsequently generated (giving 16 dialogues in total). The movie under discussion and the characters' respective agendas and their opinions about the topics were randomly assigned. Each dialogue was eight utterances long, with characters taking turns, each of them producing four utterances altogether. In each alignment dialogue, the High A/High C Character II aligned. The weight for the cache language model was set to 0.7. In both aligning and non-aligning versions of the dialogues, utterances for the non-aligning speaker were the same. The generation of utterances for the aligning speaker was seeded with the respective previous utterance functioning as the dialogue history. From the list of generated possible utterances, the top-ranked utterance was chosen.

4.2.1 Example Dialogue

To give an impression of the generated dialogues, Table 2 shows an example of Dialogue Type 1 (High E versus Low E) where the characters discuss the movie *Mystic River* (the first row of Table 1 gives the full parameter settings). The other generation parameters are (valence of opinions follows each topic): Character I, agenda (PLOT/−, CHARACTERS/−, MUSIC/−, FILM/−); further opinions (SPECIAL EFFECTS/−, ACTION SCENES/+, DIALOGUE/−); Character II, agenda (ACTION SCENES/+, SPECIAL EFFECTS/+, PLOT/−, DIALOGUE/−); further opinions (CHARACTERS/−, FILM/−, MUSIC/−).

Alignment is switched on for Character II, with examples of their utterances from the non-aligning dialogues included beneath. As can be seen, the aligned utterances closely match the surface form of their primes. The personality language models' influence on language choice can be seen when the High E Character I uses the phrase *I mean I didn't think*, while the (non-aligning) Low E Character II prefers *I (kind of) thought* throughout. Finally, the Low A/Low C Character I has a strategy of switching topics in order to follow his own agenda.

4.3 Procedure

An introductory web page explained the task and described the four personality dimensions of interest (E, N, A and C; adjectives relating to each personality type were included for reference after each dialogue). The first page also contained an example dialogue similar to those that would be displayed during the experiment. The participants were then presented with eight dialogues (two per type; one with, one without alignment). The dialogues were presented in random order and randomly chosen from the pool of available dialogues. One complete dialogue was shown at a time, and the participants were asked to judge which of the two characters scored higher on the E, N, A and C dimensions (using the previous personality descriptions and adjectives). They also answered questions probing their

	PLOT:
I	"I mean I didn't think that the story was great."
II	"I mean the story wasn't great."
	Without alignment:
	"Yes I mean I thought that it wasn't nicely tied up."
	CHARACTERS:
I	"And I mean I thought there was no characterization at all."
II	"I mean I thought there was no characterization at all."
	Without alignment:
	"Yeah there was no characterization at all I mean."
	MUSIC:
I	"I mean I didn't think that the music was great."
II	"I mean the music wasn't great."
	Without alignment:
	"Yeah I kind of thought that the music was also very bad you know."
	FILM:
I	"I mean I didn't think the film was good."
II	"I mean the film wasn't good."
	Without alignment:
	"I mean I kind of thought that it was a bit dull."

Table 2: Example Dialogue.

perceptions of the characters' relationship. They assessed on a seven-point Likert scale how well the characters 'got on' with each other (*very badly−very well*), interpreted as indicating positivity or rapport between characters, and how smoothly the conversation went (*not at all smoothly−very smoothly*), indicating how natural and coherent the interactions were. The participants were asked to rate each dialogue independently from the others.The experiment was open to both native and non-native speakers of English; upon supplying an email address, participants were entered into a draw for an Amazon gift token. All data were analysed anonymously. Note that this is a further evaluation of data previously presented in Brockmann (2009).

5 Experimental Results

5.1 Personality perception

To study the perception of personality in our dialogues, a nominal logistic regression was run on the perception ratings obtained from the judges. Here agreement between generated personality and rater judgements was coded as a binary value (agreement=1; disagreement=0), and entered into the regression model as the dependent variable (DV). The following independent variables (IVs) were entered into the model: Dialogue Alignment as a binary variable (alignment=1; no alignment=0); Personality Trait judged as a categorical variable ("Extraversion", "Neuroticism'", "Agreeableness", "Conscientiousness"). We also included an interaction variable, Generated Alignment × Personality Trait Rated. We ran this model in order to understand how each of the independent variables, such as Personality Trait judged, or combinations of variables (in the case of the interactions) best explain the accuracy of the personality perception judgements relative to our generated personality language (the DV). Throughout this section we report the parameter estimates and corresponding one degree of freedom for the more conservative Likelihood Ratio Chi Square effect tests for N=1920 (with the exception of the four-level variable, Personality Trait DF=3, and Participant ID DF=79).

The whole model is significant ($\chi^2 = 128.22$, $p < .0041$, R Square (U)= .05; although note that R Square (U) is not comparable to regular R Square, and therefore cannot be interpreted as a percentage of variance explained; model DF= 89). To investigate effects of native/non-native speaker effects on personality judgement accuracy, this variable was included in earlier models as a binary variable (Native Speaker: native=1; non-native=0), but no significant effect was found ($\chi^2 = 0.98, p = .3228$). Therefore data from all participants are included in the analyses here, and the native/non-native variable is not included in the model. For the interactions, there is a significant relationship between Dialogue Alignment and accuracy in judgement of Personality Trait ($\chi^2 = 13.67, p = .0034$). Further examination of this relationship shows that in the case of Agreeableness, accuracy decreases when alignment is present in the dialogue ($\chi^2 = 10.90, p = .0010$), whereas in the case of Conscientiousness, perception accuracy significantly increases with alignment ($\chi^2 = 4.38, p = .0364$). This is shown in Figure 1.

There is a significant main effect for Personality Trait judged ($\chi^2 = 17.04, p = .0007$): parameter estimates show that accuracy of judgement is significantly more accurate for Extraversion ($\chi^2 = 7.21, p = .0073$), but less accurate for Agreeableness ($\chi^2 = 5.54, p = .0186$) and Conscientiousness ($\chi^2 = 8.09, p = .0044$). No main effect was found for Dialogue Alignment relative to accuracy of personality judgement ($\chi^2 = 2.16, p = .1420$).

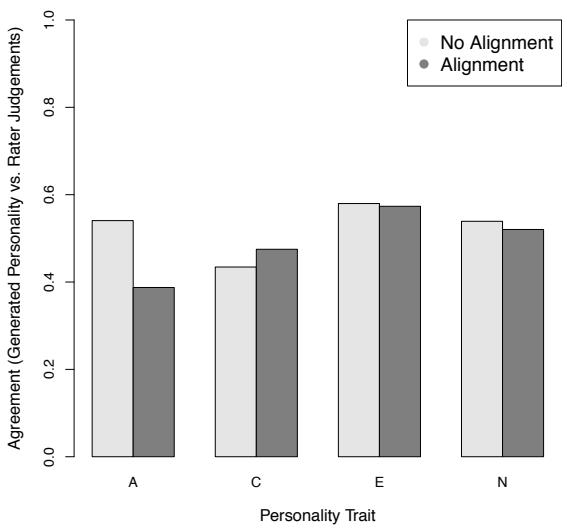

Figure 1: Accuracy of personality judgements.

5.2 Ratings of 'Getting on' and 'Smoothness'

In the following we are interested in examining what dialogue characteristics lead to the rater judgements of 'getting on'. Using an ordinal logistic regression (DV: how well the characters were judged to 'get on', seven point scale from 'very badly' to 'very well') the following independent variables, coded as described in the previous section, were entered into the model: Dialogue Alignment and Native Speaker (Personality Trait was also entered into the model, but did not reach significance). Participant ID was included in the model to account for the repeated measures design. Again, we use likelihood ratio effect tests and note parameter estimates for one degree of freedom (N=2560). The whole model is significant ($\chi^2 = 1396.75$, $p < .0001$, R Square (U)= .15; model DF=89): A main effect for Dialogue Alignment ($\chi^2 = 244.94$, $p < .0001$), shows alignment decreased perceptions of 'getting on'.

Similarly, ordinal logistic regressions were used to probe influencing factors in decisions of rating dialogue smoothness (DV: smoothness rated on a seven point scale from 'not at all smoothly' to 'very smoothly'). The following independent variables, coded as described in the previous section, were entered into the model: Dialogue Alignment and Native Speaker (again Personality Trait did not reach significance for inclusion). Again, Participant ID was included in the model to account for the repeated measures design (parameter estimates and likelihood ratio effect tests are for one degree of freedom, N=2560, Condition, DF=3; Participant ID, DF=78). The whole model is significant ($\chi^2 = 1291.28$, $p < .0001$), with an R Square (U) of 0.13 (model DF=89). There are strong main effects for Dialogue Alignment ($\chi^2 = 188.27$, $p < .0001$), and Native Speaker ($\chi^2 = 110.00$, $p < .0001$). Examination of the parameter estimates reveals negative relationships between ratings of smoothness and Native Speaker, and Dialogue Alignment, implying that native speakers significantly rated the dialogues as being less smooth than the non-native speakers, and also that dialogues with alignment were rated significantly less smooth than those without alignment.

6 Evaluation Method

To better understand the linguistic alignment processes which drive the participants' judgements in the previous experiment, we performed further analysis. In particular, we coded the forms of alignment present in each utterance of each dialogue, relative to the previous utterance. The forms of alignment were coded as follows: Polarity (matching a positive or negative opinion), Topic (whether the topic is the same or shifts), Word (instances of alignment of individual words of the previous utterance), Phrase (alignment of phrases), Construction (alignment at a grammatical construction level). Each instance of alignment for a given utterance was counted, with an overall score generated for the whole dialogue. This coding procedure was performed by one researcher and subsequently evaluated by a second, with disputes resolved by mutual agreement. In the following analysis we do not distinguish between dialogues intentionally generated with alignment and those without, but instead include all dialogues in the analysis to examine which objectively measured forms of alignment relate to the judges' perceptions for personality, 'getting on' and 'smoothness'.

7 Evaluation Results

7.1 Alignment Forms and Personality

Accuracy of judgements of personality ratings and dialogue alignment was analysed for each of the four

personality traits (A, C, E, N) independently using nominal logistic regression (DV: rater vs. generated personality agreement coded 0 or 1; IVs: occurrence scores for Polarity, Topic, Word, Phrase, and Construction). For Agreeableness the whole model is significant ($\chi^2 = 85.74$, $p < .0001$, R Square (U)= .10; model DF=5, N=640), with Topic alignment ($\chi^2 = 16.68$, $p < .0001$), followed by Polarity ($\chi^2 = 10.13$, $p = .0015$) and Construction ($\chi^2 = 6.19$, $p = .0128$) alignment all positively related to perceptions of Agreeableness. For Conscientiousness (whole model $\chi^2 = 11.26$, $p = .0465$, R Square (U)= .01; DF=5, N=640), Polarity alignment is inversely related to perceptions of Conscientiousness ($\chi^2 = 5.12$, $p = .0236$). In the case of Neuroticism and Extraversion, the models are not significant ($\chi^2 = 5.37$, $p = .3719$, and $\chi^2 = 1.49$, $p = .2226$, respectively; both DF=5, N=320).

7.2 Alignment Forms and 'Getting On' and 'Smoothness'

The relationship between the different forms of alignment present in the dialogues and the judges' ratings of 'getting on' and 'smoothness' were evaluated in two separate ordinal logistic models, in which they were entered as the dependent variable. The five alignment types (Polarity, Topic, Word, Phrase, and Construction) were entered as independent variables. Participant ID was also entered into the model as an independent variable, since multiple responses were collected from each participant.

Ratings of 'getting on' (whole model $\chi^2 = 1595.10$, $p < .0001$, R Square (U)= .17; DF=84, N=2560) show that Polarity ($\chi^2 = 385.45$, $p < .0001$), Construction ($\chi^2 = 72.30$, $p < .0001$) and Topic ($\chi^2 = 16.68$, $p = .0014$) alignment all relate to greater scores of perceived getting on. Conversely, Word alignment leads to reduced scores of perceived getting on ($\chi^2 = 14.13$, $p = .0002$). For ratings of dialogue 'smoothness' ($\chi^2 = 1519.31$, $p = .0014$, R Square (U)= .16; DF=84, N=2560), again Polarity ($\chi^2 = 209.55$, $p < .0001$), Topic ($\chi^2 = 39.39$, $p < .0001$) and Construction ($\chi^2 = 28.01$, $p < .0001$) alignment all lead to increased ratings of 'smoothness'. Similarly, Word alignment has a negative impact upon perceptions of dialogue 'smoothness' ($\chi^2 = 29.24$, $p < .0001$).

8 Discussion

We now discuss the perception and evaluation results of the CRAG 2 system in greater detail. In terms of personality perception, extraversion is accurately perceived, with agreeableness and conscientiousness less so, which matches findings from personality perception studies in other contexts, including text based computer-mediated communication (Li and Chignell, 2010; Gill et al., 2006). It is interesting to note, however, that alignment helps perception of conscientiousness, but hurts ratings of agreeableness. Reduced accuracy in perception of agreeableness, which is important to relationships, may have a negative impact on the use of dialogues in collaborative settings (Rammstedt and Schupp, 2008). Further work could usefully examine ways in which these characteristics can be generated in more readily perceptible ways. Interestingly, personality perception is unaffected by whether the judges are native English speakers or not. This is a notable finding, and apparently implies that the social information relating to personality is available in the text only environment, or through the generation process, it is equally accessible to native and non-native English speakers. Native speaking judges were more critical in rating dialogue smoothness and characters getting on, perhaps indicating a finer-grained awareness of linguistic cues in interpersonal interaction, or else just greater confidence in making negative judgements of their native language.

Our finding that our generated alignment actually decreases the perceived positivity of the relationship is contrary to what is generally predicted by the literature (Brennan and Clark, 1996; Shepard et al., 2001; Pickering and Garrod, 2004); but cf. Niederhoffer and Pennebaker (2002). Likewise, we would also have expected the dialogues with alignment to have been perceived to have gone more smoothly. However, in our evaluation of the different types of alignment, we note that alignment per se is not necessarily a bad thing: Generally alignment of Polarity, Topic, and Construction are seen positively leading to higher ratings of 'getting on', 'smoothness', and increased accurate perception of Agreeableness; repetition of individual words is however viewed negatively, and leads to lower ratings of 'getting on' and 'smoothness'.

There are a number of possible explanations for these negative responses to our generated dialogue alignment. They hinge on understanding what is involved in generating alignment, or similar behaviour, in dialogue participants. First, it could be that our dialogues encode the 'wrong' type of similarity. For example, the alignment and entrainment approaches to similarity usually study task-based dialogues, which often focus on establishing a shared vocabulary for referencing objects (i.e., at the word level). In such cases, the similarity arises either through priming mechanisms, or the establishment of common ground. Given that we used an alignment model to generate similarity in our dialogues, this kind of repetition or similarity may seem incongruent or out of place in dialogues that are not task-based (cf. negative impact of word-level alignment).

A second explanation might be that similarity relates to positive outcomes when it occurs over a longer, rather than shorter, period of time (Reitter, 2008). In the current study the dialogues consisted of eight turns, thus similarity was not generated over a long period. Indeed, linguistic similarity over a longer period of time may be more consistent with perceptions of social similarity, such as ingroup, rather than outgroup, membership (Shepard et al., 2001). Indeed, in such contexts word choice *is* an important feature in dialogue and would be useful to incorporate into a dialogue model to simulate ingroup membership.

Third, in communication accommodation theory it is 'convergence' – the process of *increasing* similarity between interlocutors – which is important, rather than similarity alone. In the current study, convergence was not examined since the dialogues were generated with static levels of alignment.

So how do these findings relate back to the area of dialogue generation for applied contexts? Similarly to findings for the PERSONAGE system (Mairesse and Walker, 2010), personality in our generated dialogues is perceived with similar accuracy to the way humans perceive personality of other humans. This suggests that our CRAG 2 system can create believable characters to whom the user can potentially relate while auditing the dialogues, or using a dialogue-based interface. That alignment can have negative effects on dialogue perception we propose is due to the form of alignment depicted in these generated dialogues (i.e., task-based nature emphasising similarity at the word level), rather than alignment in general. We do not take this result to necessarily indicate that alignment in generated dialogues should be avoided. Rather, its implementation should be carefully considered, especially to ensure that the form of similarity achieved makes sense in the communicative context. Indeed, as we show in the evaluation of the generated dialogues, alignment at the Polarity, Topic, and Construction levels is generally viewed positively, however in contrast alignment at the Word level tends to be viewed more negatively. One of the key suggestions arising from this study is that the different forms of dialogue similarity cannot simply be used interchangeably, with alignment found in task-based dialogues which may include many instances of word-level repetition and alignment not necessarily appropriate in non-task dialogues, and thus not automatically resulting in perceptions of positivity. We note that non-native speakers were more forgiving in their ratings of the dialogues containing alignment. Given that they were equally able to perceive the personality of the characters, this may be due to non-native speakers having fewer expectations of alignment behaviour in dialogue. Indeed in some contexts, greater alignment, and thus repetition, may be beneficial for non-native speakers auditing dialogues.

To conclude, personality in our generated dialogues was perceived with comparable accuracy to human texts, but alignment or similarity between speakers – especially at the word level – regarded negatively. We would like to see future work examine further the responses to different forms of alignment, including convergence, in generated dialogue.

9 Acknowledgements

We acknowledge Edinburgh-Stanford Link funding, and the partial support of the Future and Emerging Technologies programme FP7-COSI-ICT of the European Commission (project QLectives, grant no.: 231200). We thank Amy Isard, Scott Nowson and Michael White for their assistance in this work. A version of the paper was presented at the Twentieth Society for Text and Discourse conference; thanks to Herb Clark, Max Louwerse and Michael Schober for their insights regarding linguistic similarity.

References

[Bortfeld and Brennan1997] H. Bortfeld and S. E. Brennan. 1997. Use and acquisition of idiomatic expressions in referring by native and non-native speakers. *Discourse Processes*, 23:119–147.

[Brennan and Clark1996] Susan E. Brennan and Herbert H. Clark. 1996. Conceptual pacts and lexical choice in conversation. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 22(6):1482–1493, November.

[Brennan1996] Susan E. Brennan. 1996. Lexical entrainment in spontaneous dialog. In *International Symposium on Spoken Dialog*, pages 41–44.

[Brockmann2009] Carsten Brockmann. 2009. *Personality and Alignment Processes in Dialogue: Towards a Lexically-Based Unified Model*. Ph.D. thesis, University of Edinburgh, UK.

[Cassell et al.2000] Justine Cassell, Joseph Sullivan, Scott Prevost, and Elizabeth Churchill, editors. 2000. *Embodied Conversational Agents*. MIT Press, Cambridge, MA, USA.

[Foster and White2004] Mary Ellen Foster and Michael White. 2004. Techniques for text planning with XSLT. In *Proceedings of the 4th Workshop on NLP and XML (NLPXML-04) at the 42nd Annual Meeting of the Association for Computational Linguistics (ACL-04)*, pages 1–8, Barcelona, Spain.

[Gill and Oberlander2002] Alastair J. Gill and Jon Oberlander. 2002. Taking care of the linguistic features of extraversion. In *Proceedings of the 24th Annual Conference of the Cognitive Science Society (CogSci2002)*, pages 363–368, Fairfax, VA, USA.

[Gill et al.2006] Alastair J. Gill, Jon Oberlander, and Elizabeth Austin. 2006. Rating e-mail personality at zero acquaintance. *Personality and Individual Differences*, 40(3):497–507.

[Hernault et al.2008] Hugo Hernault, Paul Piwek, Helmut Prendinger, and Mitsuru Ishizuka. 2008. Generating dialogues for virtual agents using nested textual coherence relations. In *Proceedings of Intelligent Virtual Agents*, pages 139–145.

[Isbister and Nass2000] Katherine Isbister and Clifford Nass. 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.

[Li and Chignell2010] J. Li and M. Chignell. 2010. Birds of a feather: How personality influences blog writing and reading. *Int. J. Human-Computer Studies*, 68:589–602.

[Mairesse and Walker2010] François Mairesse and Marilyn Walker. 2010. Towards personality-based user adaptation: Psychologically informed stylistic language generation. *User Modeling and User-Adapted Interaction*, 20(3):227–278.

[Niederhoffer and Pennebaker2002] Kate G. Niederhoffer and James W. Pennebaker. 2002. Linguistic style matching in social interaction. *Journal of Language and Social Psychology*, 21(4):337–360.

[Nowson et al.2005] S. Nowson, J. Oberlander, and A.J. Gill. 2005. Weblogs, genres and individual differences. In *Proceedings of the 27th Annual Conference of the Cognitive Science Society*, pages 1666–1671.

[Pennebaker and King1999] James W. Pennebaker and Laura A. King. 1999. Linguistic styles: Language use as an individual difference. *Journal of Personality and Social Psychology*, 77(6):1296–1312.

[Pickering and Garrod2004] Martin J. Pickering and Simon Garrod. 2004. Toward a mechanistic psychology of dialogue. *Behavioral and Brain Sciences*, 27(2):169–225.

[Porzel et al.2006] Robert Porzel, Annika Scheffler, and Rainer Malaka. 2006. How entrainment increases dialogical efficiency. In *Proceedings of Workshop on on Effective Multimodal Dialogue Interfaces*.

[Rammstedt and Schupp2008] Beatrice Rammstedt and Jürgen Schupp. 2008. Only the congruent survive – personality similarities in couples. *Personality and Individual Differences*, 45(6):533–535.

[Reitter2008] David Reitter. 2008. *Context Effects in Language Production: Models of Syntactic Priming in Dialogue Corpora*. Ph.D. thesis, University of Edinburgh, UK.

[Shepard et al.2001] Carolyn A. Shepard, Howard Giles, and Beth A. Le Poire. 2001. Communication accommodation theory. In W. Peter Robinson and Howard Giles, editors, *The New Handbook of Language and Social Psychology*, chapter 1.2, pages 33–56. John Wiley & Sons, Chichester, UK.

[van Deemter et al.2008] Kees van Deemter, Brigitte Krenn, Paul Piwek, Martin Klesen, Marc Schröder, and Stefan Baumann. 2008. Fully generated scripted dialogue for embodied agents. *Artificial Intelligence*, 172(10):1219–1244.

[White2006a] Michael White. 2006a. CCG chart realization from disjunctive inputs. In *Proceedings of the 4th International Natural Language Generation Conference (INLG-06)*, pages 9–16, Sydney, Australia.

[White2006b] Michael White. 2006b. Efficient realization of coordinate structures in Combinatory Categorial Grammar. *Research on Language and Computation*, 4(1):39–75.