Re-Creating Dialogues from a Corpus

Amy Isard and **Carsten Brockmann** and **Jon Oberlander**
HCRC, University of Edinburgh
2 Buccleuch Place
Edinburgh EH8 9LW, UK
{Amy.Isard,Carsten.Brockmann,J.Oberlander}@ed.ac.uk

Abstract

We describe the collection and annotation of a corpus of dialogues about movies, and a system which uses utterances from this corpus in generating dialogues which vary according to the personalities assigned to two characters.

1 Introduction

This paper describes the design and implementation of the first version of the Critical Agent Dialogue (CrAg) system, which generates dialogues which vary according to the personalities assigned to two characters, based on recent research into different vocabulary, syntax and dialogue strategies exhibited according to personality type [Gill and Oberlander, 2002]. This research uses Eysenck's three factor model [Eysenck and Eysenck, 1991], in which personality is described in terms of the three dimensions psychoticism, extraversion, and neuroticism, each of which can separately influence language production.

This is not a dialogue system in the usual sense; both sides of the conversation are generated, in order to be able to manipulate the interactions between participants with different personalities. The output is currently being evaluated by human subjects, to ascertain whether they can recognise the personality characteristics which we have mimicked.

This system has some similarities to the NECA system [Piwek, 2003], but rather than focussing on the discourse structure, we have concentrated on the component of the system which assigns personality scores to utterances.

The dialogues are constructed using utterances from a corpus (see Section 2) which are ranked and combined into a coherent dialogue using the techniques described in Sections 3 and 4. Some example dialogues are presented in Section 4.4.

2 Corpus

2.1 Collection

Ten pairs of participants went to see a film of our choosing and were later recorded having a conversation about it. Three films were chosen which were showing at the same time, and were from three different genres: "League of Extraordinary Gentlemen" (action, sci-fi, fantasy), "Mystic River" (drama, crime) and "Intolerable Cruelty" (romantic comedy). The dialogues were recorded in a soundproof room, and participants were told that they could talk about any aspect of the film of their choosing, and asked to try to stay on the topic of the film they had just seen, but the conversation was not monitored. Dialogues ranged in length from 12 to 25 minutes, with an average of 19 minutes.

2.2 Transcription and Annotation

The dialogues were segmented into phrases and transcribed orthographically using the Transcriber tool [Barras *et al.*, 2001]. The dialogues were then annotated using the NITE XML Toolkit (NXT [Carletta *et al.*, 2003]). This toolkit provides utilities which allow users to create their own annotation interface. An interface was created which displayed the two participants' speech in separate windows, and allowed the annotator to listen to the speech and to combine phrases into utterances, while annotating the utterances as described below. A screenshot of the tool can be seen in Figure 1.

Topics

The annotator assigned one or more topics to each utterance from a pre-defined list, shown below. Topics without definitions are assumed to be self-explanatory.

- action sequences
- actors
- characters
- cinematography style: the look of the film
- dialogue
- directing: directing style, director(s)' intentions etc.
- humour
- music
- romance
- special effects
- whole movie
- other this film: a topic related to this film not included in the above list
- other film: a discussion about another film or films
- not film related: any discussion not related to films at all

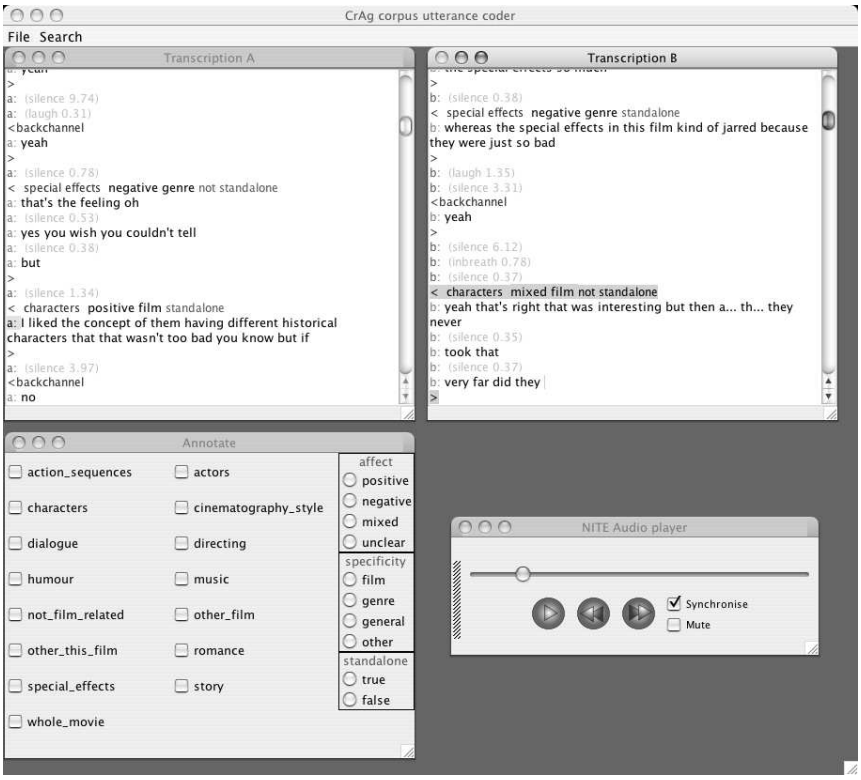

Figure 1: NXT topic annotation tool

- none: utterances where topic cannot be assigned, e.g. "um", "he it" "I ... I think'

Affect

The annotator also chose the affect of the utterance from the following list (one per utterance)

- positive
- negative
- mixed – both positive and negative e.g "all the cinematography was alright there was nothing interesting in it nothing daring"
- unclear
  - neutral e.g. "what did you think of Sean Connery", "well there's a clear implication that they had a relationship before"
  - not possible to assign affect – unclear is automatically assigned to all utterances whose topic is "none"

Generality

With re-generation in mind, utterances were labelled according to whether they make sense out of context. This means that most utterances with anaphoric references are rejected e.g. "there was no mention of that at all in the film" along with those which require knowledge of the previous utterance e.g. "and there wasn't even that much blood-sucking which is kind of disappointing for a vampire" and questions e.g. "what did you think?"

For the same reason, the utterances were also ranked for whether they could apply to just one film (e.g. "they'd start little storylines like when Sean Connery was teaching the American chap to shoot") or could be used to discuss any film (e.g. "I don't have anything positive to say about it actually").

2.3 Corpus Statistics

Topics/Films	LXG	IC	MR	all
action sequences	11	0	16	27
actors	30	66	95	171
characters	110	52	282	444
cinematography style	7	0	12	19
dialogue	37	8	8	53
directing	23	48	65	136
humour	5	76	2	83
music	0	0	25	25
romance	0	9	8	17
special effects	48	0	0	48
story	165	83	245	493
whole movie	74	36	44	154
other	106	173	124	403
Total	427	401	637	1465

Table 1: All Utterances

This resulted in a total of 1465 utterances averaging 73 per speaker. The topics are not distributed evenly throughout the dialogues since we used films from three different genres, and

some topics (e.g. special effects) do not apply to all types of film.

Table 1 includes all the utterances in the corpus (N.B. because there can be more than one topic per utterance, the totals at the bottom are less than the sum of their columns).

Table 2 shows all the utterances which were considered to be usable for re-generation. Utterances listed under each film are those which could only be used in a discussion of that particular film, and those in the column "general" could be used to talk about any film (see Section 2.2).

Topics/Films	LXG	IC	MR	general	all
action sequences	6	0	1	0	7
actors	1	4	12	2	19
characters	7	1	12	3	23
cinematography style	5	0	2	2	9
dialogue	3	2	0	4	9
directing	1	3	2	1	7
humour	2	8	0	0	10
music	0	0	1	0	1
romance	0	0	0	0	0
special effects	14	0	0	0	14
story	14	2	12	7	35
whole movie	10	4	11	19	44
Total	44	15	41	32	132

Table 2: Re-usable Utterances

3 Ranking Utterances by Personality Features

3.1 Framework

The Critical Agent Dialogue (CrAg) 1.0 system is implemented as a collection of Open Agent Architecture (OAA, [Cheyer and Martin, 2001]) agents. Each agent is a program designed to fulfil a specific task; it informs the special OAA facilitator agent about its capabilities. Whenever an agent requires a task to be resolved, it sends a request to the facilitator, which then invokes the agent that can deal with the request, and returns the results to the requesting agent.

3.2 Augmenting the Annotation

In a first stage, the corpus utterances' annotation is augmented with information from a variety of linguistic and psycholinguistic resources. This knowledge is then used to compute neuroticism and extraversion scores (see Section 3.3).

Part of Speech Tagging and Lemmatisation

Each utterance is split into sentences, tokenised, and tagged with part of speech information using the *mxpost* part of speech tagger [Ratnaparkhi, 1996]. The *morph* tool [Minnen *et al.*, 2001] then determines each word's lemma form.

Based on the lemmata, we compute each utterance's type/token ratio, which measures the variety of words used; it equals 1 if every type is used only once, and decreases with each repetition.

MRC Psycholinguistic Database

The annotation is further augmented by information from the MRC Psycholinguistic Database (MRC PDb, [Wilson, 1988]), a machine readable dictionary of 150,837 words. For each word, it specifies up to 26 linguistic and psycholinguistic attributes, e.g.:

- written/spoken word frequencies
- familiarity, concreteness, imageability
- meaningfulness
- age of acquisition
- part of speech
- phonetic transcription, stress pattern

Linguistic Inquiry and Word Count

Linguistic Inquiry and Word Count (LIWC2001, [Pennebaker *et al.*, 2001]) is another machine readable dictionary. 2,300 words and word stems are annotated with one or more of 74 categories, e.g.:

- linguistic dimensions (pronouns, negations, articles, ...)
- psychological processes
  - positive/negative emotions
  - cognitive processes (insight, certainty, ...)
  - perceptual processes (seeing, hearing, feeling)
  - social processes (friends, family, ...)
- relativity (time, space, motion)
- personal concerns (occupation, leisure, physical states, ...)

The Formality Measure *F*

From each utterance's part of speech annotation we compute the formality measure F [Heylighen and Dewaele, 2002]; the authors propose the concept of formality as a "dimension of variation between linguistic expressions". The measure is based on frequency percentages of different word classes:

$$\begin{aligned} F = &(\text{noun freq.} + \text{adjective freq.} + \text{preposition freq.} \\ &+ \text{article freq.} - \text{pronoun freq.} - \text{verb freq.} \\ &- \text{adverb freq.} - \text{interjection freq.} + 100)/2 \end{aligned} \quad (1)$$

In Heylighen and Dewaele's study, oral female ($F = 38.7$) and oral male ($F = 41.6$) language was classified as informal; novels ($F = 52.5$) were average, while scientific text ($F = 65.7$) and newspapers ($F = 68.1$) ranked high on the formality scale.

3.3 Feature Combination

Previous research identified features characteristic for the language of extravert or neurotic speakers [Gill and Oberlander, 2002; Oberlander and Gill, 2004]. According to these results, we combine the utterance scores computed during the annotation phase using additive multiattribute value functions (AMVF). AMVF have been applied to represent user preferences [Carenini and Moore, 2000]; we use an implementation done for the user modelling component of the FLIGHTS system [Moore *et al.*, 2004].

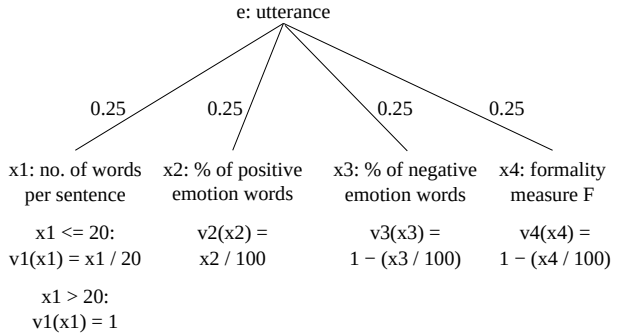

Figure 2: Partial additive multiattribute value function (AMVF) for extravert language.

In an AMVF, a value tree specifies the hierarchy of aspects of an entity e. Edges are weighted (w) according to the importance of their contribution to the parent node. For each leaf, a component value function v_i maps attribute value x_i to the $[0,1]$ interval (1 is most preferable). The weight w_i of a leaf node is computed as the product of the weights from the tree's root down to the leaf. Given this model, the value $v(e)$ of entity e can be computed:

$$v(e) = v(x_1, \ldots, x_n) = \sum w_i v_i(x_i) \quad (2)$$

A simplified example AMVF for extravert utterances is shown in Figure 2. Our complete set of features characteristic for high extravert language is listed below:

- high:
  - number of words per sentence
  - number of sentences per utterance
  - percentage of conjunctions (part of speech tag)
  - mean frequency count of spoken English (MRC PDb)
  - percentage of certainty words (*always*, *never*; LIWC)
  - percentage of positive emotion words (*happy*, *pretty*, *good*; LIWC)
  - percentage of social process words (*talk*, *us*, *friend*; LIWC)
- low:
  - percentage of determiners (part of speech tag)
  - mean concreteness (MRC PDb)
  - percentage of negation words (*no*, *never*, *not*; LIWC)
  - percentage of negative emotion words (*hate*, *worthless*, *enemy*; LIWC)
  - formality (F measure, cf. Section 3.2)
  - lemma-based type/token ratio

The following features characterise high neurotic language:

- high:
  - percentage of first person singular words (*I*, *my*, *me*; LIWC)
  - percentage of negative emotion words (LIWC)
- low:
  - percentage of determiners (part of speech tag)
  - percentage of positive emotion words (LIWC)
  - formality (F measure)

In the current version of the system, all features are given equal weight; we plan to fine-tune the weight adjustments in future.

4 Re-Generating Dialogue

4.1 Initialisation

Computer characters are defined by values for the personality dimensions extraversion (E), neuroticism (N), and psychoticism (P). These values are given in a range from 0 (low) to 1 (high). For psychoticism, in our current implementation, only the two settings low ($P < 0.5$) and high ($P \geq 0.5$) are distinguished, as explained below. The characters are also each assigned an agenda of topics about which they would like to talk; for each topic, their opinion about it (the *polarity*) is either positive or negative.

Dialogues between two computer characters are then regenerated by the OAA CrAg Steering Agent. Two character definitions and one of the three available films are selected, and the number of turns to generate is set.

4.2 The Affective Language Production Model

The generation process is informed by the Affective Language Production (ALP) model, developed by Oberlander and Gill. The simplest version of this model (ALP-1) starts from the idea that high extraverts have plenty of resource for linguistic interaction, but need to put less of it into detailed planning. High neurotics have less resource for linguistic interaction in the first place. It follows that extraversion finds its effects mostly at the stages of formulation (surface realisation). That is, the process and representations used in realisation differ between high and low extraverts. Secondly, neuroticism finds its effects at the stage of conceptualisation (content selection). That is, the process and representations used in content selection differ between high and low neurotics. Since content selection precedes surface realisation, variations in neuroticism will have consequences beyond the content selection stage, but this is their primary locus.

ALP-2 adds to ALP-1 a simple treatment of psychoticism. The idea is that some individuals pay more attention to their partner's utterances than others. Those who pay attention are more likely to produce utterances which overlap in both content (and form) with those of their partner.

4.3 Re-Generation

Thus, the re-generation process operationalises ALP-2 in the following way.

First, the character with the higher extraversion score begins the dialogue, and the topic at the top of their agenda is used.

Subsequent topic choice is influenced by the characters' psychoticism score. A low psychotic character continues on

the same topic as the previous speaker, while a high psychotic changes topic to the one at the top of their own agenda.

Next the corpus is queried for all utterances with the appropriate film, topic, and polarity. By the previous annotation process (cf. Section 3.2) each utterance was assigned an N and an E score.

A character's N level relates to content selection. Thus, the N score is the first filter we apply to rank the utterance candidates. To implement this as simply as possible, we keep the half of the candidates which most closely match the desired N score, and discard the other half.

A character's E level relates to surface realisation behaviour. So the remaining candidates are sorted by E score, and we choose the one that is closest to the character's desired E score as their actual utterance.

After this, it is the other character's turn to say something. The routine loops back to the point of topic choice.

4.4 Example Dialogues

Here we present three example dialogues with different combinations of character personalities. In all these examples, character **a** begins the dialogue, and in the first two cases, since **a** is HiP and **b** is LoP, we end up with two utterances on each topic, as **b** always aligns to **a**'s topic, and **a** always starts a new subject. In the third dialogue, both characters are LoP, so they will continue speaking on the same topic until they run out of things to say.

Dialogue 1: Mystic River

Character a HiE/MidN/HiP

Character b LoE/MidN/LoP

a (characters) I mean I think maybe maybe one of the things is it had too many characters and you know you can't do a character building thing with you know three main characters two cops and you know all the supporting family and all that

b (characters) the other thing the eventual murderer the kid brother I don't think his character was very well developed

a (whole movie) you know what it was like for me it was like a b grade book that's been made into a movie

b (whole movie) for a mystery crime it's absolutely crap

a (story) I didn't really care for the story too much you know there are certain films that have good murder mysteries and you know the person dies at the beginning but you still don't know who did it or what happened but this didn't really engage I guess

b (story) it could have been just that there's a a big grand vision of like all these lives intertwined that was cut down

Dialogue 2: League of Extraordinary Gentlemen

Character a LoE/MidN/HiP

Character b LoE/MidN/LoP

a (characters) I think the particular disappointment for me was was not getting the characters interacting and kind of jostling enough

b (characters) I mean there is this mix of different well known characters but they never sort of explore their personalities in any way

a (special effects) maybe it was an incredibly low budget film and so they had to do everything on computers

b (special effects) the the set where the city scenes and buildings they were just so obviously fake as to be unreal I mean

a (whole movie) it was a bit dull

b (whole movie) I was on the whole disappointed with it

Dialogue 3: League of Extraordinary Gentlemen

Character a MidE/MidN/LoP

Character b MidE/MidN/LoP

a (whole movie) so basically we have nothing good to say about it

b (whole movie) I have to say this is very rare this is by far the the worst film I've seen in a in a very long time

a (whole movie) they didn't do anything particularly well at all

b (whole movie) I've seen a lot of kind of bad science fiction adventure fantasy type films before but this one is sort of bad in ways that none of the other bad films are bad you know

a (whole movie) I think probably I don't know I really thought it was going to be better than it was I think that the big names carried it and if you just took big names out of that movie it wouldn't have any I think it's a tv movie

b (whole movie) I'm still trying to grab something that I liked about it

5 Future Work

The first version of our generation system selects utterances from a corpus of human dialogues about films. The next version, currently under development, will use the OpenCCG realiser [White and Baldridge, 2003] to create more varied and flexible dialogues on the same subject, using the personality ranking algorithms presented here. We will also introduce the use of alignment as described in [Brockmann *et al.*, 2005].

There are many other possible applications for the ranking method presented in Section 3. It could for example be used in a real-time dialogue between a user and a system, to evaluate the probable personality of the user and provide appropriate responses.

References

[Barras *et al.*, 2001] Claude Barras, Edouard Geoffrois, Zhibiao Wu, and Mark Liberman. Transcriber: development and use of a tool for assisting speech corpora production. *Speech Communication*, 33(1–2):5–22, 2001. See also `www.etca.fr/CTA/gip/Projets/Transcriber`.

[Brockmann *et al.*, 2005] Carsten Brockmann, Amy Isard, Jon Oberlander, and Michael White. Modelling alignment for affective dialogue. In *Proceedings of the Workshop on Adapting the Interaction Style to Affective Factors*, Edinburgh, July 2005. To appear.

[Carenini and Moore, 2000] Giuseppe Carenini and Johanna D. Moore. A strategy for generating evaluative arguments. In *Proceedings of the 1st International Natural Language Generation Conference (INLG-00)*, pages 47–54, Mitzpe Ramon, Israel, 2000.

[Carletta *et al.*, 2003] Jean Carletta, S. Evert, U. Heid, J. Kilgour, J. Robertson, and H. Voorman. The NITE XML toolkit: flexible annotation for multi-modal language data. *Behavior Research Methods, Instruments, and Computers*, 35(3):353–363, 2003.

[Cheyer and Martin, 2001] Adam Cheyer and David Martin. The Open Agent Architecture. *Journal of Autonomous Agents and Multi-Agent Systems*, 4(1):143–148, 2001.

[Eysenck and Eysenck, 1991] H. J. Eysenck and S. B. G. Eysenck. *The Eysenck Personality Questionnaire-Revised*. Hodder & Stoughton, Sevenoaks, 1991.

[Gill and Oberlander, 2002] Alastair J. Gill and Jon Oberlander. 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, 2002.

[Heylighen and Dewaele, 2002] Francis Heylighen and Jean-Marc Dewaele. Variation in the contextuality of language: An empirical measure. *Foundations of Science*, 7(3):293–340, 2002.

[Minnen *et al.*, 2001] Guido Minnen, John Carroll, and Darren Pearce. Applied morphological processing of English. *Natural Language Engineering*, 7(3):207–223, 2001.

[Moore *et al.*, 2004] Johanna Moore, Mary Ellen Foster, Oliver Lemon, and Michael White. Generating tailored, comparative descriptions in spoken dialogue. In *Proceedings of the 17th International FLAIRS Conference*, Miami Beach, FL, USA, 2004.

[Oberlander and Gill, 2004] Jon Oberlander and Alastair J. Gill. Individual differences and implicit language: personality, parts-of-speech and pervasiveness. In *Proceedings of the 26th Annual Conference of the Cognitive Science Society (CogSci2004)*, pages 1035–1040, Chicago, IL, USA, 2004.

[Pennebaker *et al.*, 2001] James W. Pennebaker, Martha E. Francis, and Roger J. Booth. *Linguistic Inquiry and Word Count (LIWC): LIWC2001 Manual*. Erlbaum Publishers, Mahwah, NJ, USA, 2001.

[Piwek, 2003] Paul Piwek. A flexible pragmatics-driven language generator for animated agents. In *Proceedings of EACL-03, Research Notes*, Budapest, Hungary, 2003.

[Ratnaparkhi, 1996] Adwait Ratnaparkhi. A Maximum Entropy model for Part-Of-Speech tagging. In *Proceedings of the Empirical Methods in Natural Language Processing Conference*, University of Pennsylvania, 1996.

[White and Baldridge, 2003] Michael White and Jason Baldridge. Adapting chart realization to CCG. In *Proceedings of the 9th European Workshop on Natural Language Generation*, pages 119–126, Budapest, Hungary, 2003.

[Wilson, 1988] Michael D. Wilson. The MRC Psycholinguistic Database: Machine readable dictionary, version 2. *Behavior Research Methods, Instruments, & Computers*, 20(1):6–11, 1988.