Speaking without knowing what to say… or when to end

Anna Hjalmarsson
Centre for Speech Technology
KTH
SE-10044, Stockholm, Sweden
annah@speech.kth.se

Abstract

Humans produce speech incrementally and on-line as the dialogue progresses using information from several different sources in parallel. A dialogue system that generates output in a stepwise manner and not in preplanned syntactically correct sentences needs to signal how new dialogue contributions relate to previous discourse. This paper describes a data collection which is the foundation for an effort towards more human-like language generation in DEAL, a spoken dialogue system developed at KTH. Two annotators labelled cue phrases in the corpus with high inter-annotator agreement (kappa coefficient 0.82).

1 Introduction

This paper describes a data collection with the goal of modelling more human-like language generation in DEAL, a spoken dialogue system developed at KTH. The DEAL objectives are to build a system which is fun, human-like, and engaging to talk to, and which gives second language learners of Swedish conversation training (as described in Hjalmarsson et al., 2007). The scene of DEAL is set at a flea market where a talking animated agent is the owner of a shop selling used objects. The student is given a mission: to buy items from the shop-keeper at the best possible price by bargaining. From a language learning perspective and to keep the students motivated, the agent's language is crucial. The agent needs to behave human-like in a way which allows the users to suspend some of their disbeliefs and talk to DEAL as if talking to another human being. In an experimental study (Hjalmarsson & Edlund, in press), where a spoken dialogue system with human behaviour was simulated, two different systems were compared: a replica of human behaviour and a constrained version with less variability. The version based on human behaviour was rated as more human-like, polite and intelligent.

1.1 Human language production

Humans produce speech incrementally and on-line as the dialogue progresses using information from several different sources in parallel (Brennan, 2000; Aist et al., 2006). We anticipate what the other person is about to say in advance and start planning our next move while this person is still speaking. When starting to speak, we typically do not have a complete plan of how to say something or even what to say. Yet, we manage to rapidly integrate information from different sources in parallel and simultaneously plan and realize new dialogue contributions. Pauses, corrections and repetitions are used to stepwise refine, alter and revise our plans as we speak (Clark & Wasow, 1998). These human behaviours bring valuable information that contains more than the literal meanings of the words (Arnold et al., 2003).

In order to generate output incrementally in DEAL we need extended knowledge on how to signal relations between different segments of speech. In this paper we report on a data collection of human-human dialogue aiming at extending the knowledge of human interaction and in particular to distinguish different types of cue phrases used in the DEAL domain.

Proceedings of the 9th SIGdial Workshop on Discourse and Dialogue, pages 72–75,
Columbus, June 2008.

2 The DEAL corpus collection

The dialogue data recorded was informal, human-human, face-to-face conversation. The task and the recording environment were set up to mimic the DEAL domain and role play.

2.1 Data collection

The data collection was made with 6 subjects (4 male and 2 female), 2 posing as shop keepers and 4 as potential buyers. Each customer interacted with the same shop-keeper twice, in two different scenarios. The shop-keepers and customers were instructed separately. The customers were given a mission: to buy items at a flea market at the best possible price from the shop-keeper. The task was to buy 3 objects for a specific purpose (e.g. to buy tools to repair a house). The customers were given a certain amount of toy money, however not enough to buy what they were instructed to buy without bargaining. The shop-keeper sat behind a desk with images of different objects pinned to the wall behind him. Some of the object had obvious flaws, for example a puzzle with a missing piece, to open up for interesting negotiation. None of the shop-keepers had any professional experience of bargaining, which was appropriate since we were more interested in capturing naïve conceptual metaphors of bargaining rather than real life price negotiation. Each dialogue was about 15 minutes long, so about 2 hours of speech were collected altogether. The shop-keepers used an average of 13.4 words per speaker turn while the buyers' turns were generally shorter, 8.5 words per turn (in this paper *turn* always refers to speaker turns). In total 16357 words were collected.

3 Annotation

All dialogues were first transcribed orthographically including non-lexical entities such as laughter and hawks. Filled pauses, repetitions, corrections and restarts were also labelled manually.

3.1 Cue phrases

Linguistic devices used to signal relations between different segments of speech are often referred to as *cue phrases.* Other frequently used terms are discourse markers, pragmatic markers or discourse particles. Typical cue phrases in English are: *oh, well, now, then, however, you know, I mean, because, and, but* and *or*. Much research within discourse analysis, communicative analysis and psycholinguistics has been concerned with these connectives and what kind of relations they hold (for an overview see Schourup, 1999). Our definition of cue phrases is broad and all types of linguistic entities that the speakers use to hold the dialogue together at different communicative levels are included. A rule of thumb is that cue phrases are words or chunks of words that have little lexical impact at the local speech segment level but serve significant pragmatic function. To give an exact definition of what cue phrases are is difficult, as these entities often are ambiguous. According to the definition used here, cue phrases can be a single word or larger units, occupy various positions, belong to different syntactic classes, and be realized with different prosodic contours.

The first dialogue was analyzed and used to decide which classes to use in the annotation scheme. Nine of the classes were a subset of the functional classification scheme of discourse markers presented in Lindström (2008). A tenth class, *referring*, was added. There were 3 different classes for *connectives*, 3 classes for *responsives* and 4 remaining classes. The classes are presented in Table 1; the first row contains an example in its context from data, the word(s) in bold are the labelled cue phrase, and the second row presents frequently used instances of that class.

Additive Connectives (CAD)
och grönt är ju fint [**and** green is nice]
och, alltså, så [and, therefore, so]
Contrastive Connectives (CC)
men den är ganska antik [**but** it is pretty antique]
men, fast, alltså [but, although, thus]
Alternative Connectives (CAL)
som jag kan titta på **istället** [which I can look at **instead**]
eller, istället [or, instead]
Responsive (R)
ja jag tycker ju det [yeah I actually think so]
ja, mm, jaha, ok [yes, mm, yeah, ok]
Responsive New Information (RNI)
jaha har du några sådana [**right** do you have any of those]
jaha, ok, ja, mm [right, ok, yes, mm]

Responsive Disprefrence (RD)
ja men det är klart dom funkar [**yeah** but of course they work]
ja, mm, jo [yes, mm, sure]
Response Eliciting (RE)
vad ska du ha för den **då** [how much do you want for that one **then**]
då, eller hur [then, right]
Repair Correction (RC)
nej nu sa jag fel [**no** now I said wrong]
nej, jag menade [no, I meant]
Modifying (MOD)
ja jag tycker **ju** det [yeah I **actually** think so]
ju, liksom, jag tycker ju det [of course, so to speak, I like]
Referring (REF)
fyra hundra kronor **sa vi** [four hundred crowns **we said**]
sa vi, sa vi inte det [we said, wasn't that what we said]

Table 1: The DEAL annotation scheme

The labelling of cue phrases included a two-fold task, both to decide if a word was a cue phrase or not – a binary task – but also to classify which functional class it belongs to according to the annotation scheme. The annotators could both see the transcriptions and listen to the recordings while labelling. 81% of the speaker turns contained at least one cue phrase and 21% of all words were labelled as cue phrases. Table 2 presents the distribution of cue phrases over the different classes.

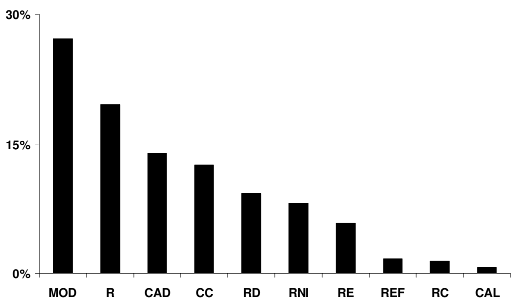

Table 2: Cue phrase distribution over the different classes

Two of the eight dialogues were annotated by two different annotators. A kappa coefficient was calculated on word level. The kappa coefficient for the binary task, to classify if a word was a cue phrase or not, was 0.87 (p=0.05). The kappa coefficient for the classification task was 0.82 (p=0.05). Three of the classes, referring, connective alternative and repair correction, had very few instances. The agreement in percentage distributed over the different classes is presented in Table 3.

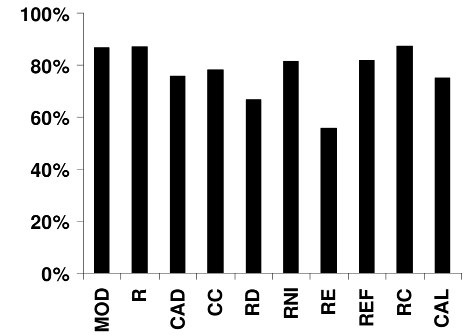

Table 3: % agreement for the different classes

4 Data analysis

To separate cue phrases from other lexical entities and to determine what they signal is a complex task. The DEAL corpus is rich in disfluencies and cue phrases; 86% of the speaker turns contained at least one cue phrase or disfluency. The annotators had access to the context and were allowed to listen to the recordings while labelling. The *responsives* were generally single words or non lexical units (e.g. "mm") and appeared in similar dialogue contexts (i.e. as responses to assertions). The classification is likely based on their prosodic realization. Acoustic analysis is needed in order to see if and how they differ in prosodic contour. In Hirschberg & Litman (1993) prosodic analysis is used to distinguish between discourse and sentential use of cue phrases. Table 4 presents how the different cue phrases were distributed over speaker turns, at initial, middle or end position.

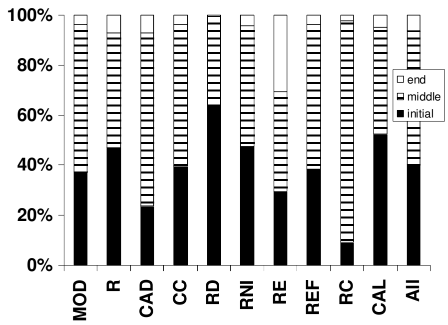

Table 4: Turn position distribution

5 Generation in DEAL

The collected and labelled data is a valuable resource of information for what cue phrases signal in the DEAL domain as well as how they are lexically and prosodically realized. To keep the response times constant and without unnaturally long delays, DEAL needs to be capable of grabbing the turn, hold it while the system is producing the rest of the message, and release it after completion. DEAL is implemented using components from the Higgins project (Skantze et al., 2006) an off-the-shelf ASR system and a GUI with an embodied conversational agent (ECA) (Beskow, 2003). A current research challenge is to redesign the modules and architecture for incremental processing, to allow generation of conversational speech. Deep generation in DEAL – the decision of what to say on an abstract semantic level – is distributed over three different modules; (1) the action manger, (2) the agent manager and the (3) communicative manager. The action manger is responsible for actions related to user input and previous discourse[1]. The agent manager represents the agents' personal motivations and personality. DEAL uses mixed initiative and the agent manager takes initiatives. It may for example try to promote certain objects or suggest prices of objects in focus. It also generates emotional facial gestures related to events in the dialogue. The communicative manager generates responses on a communicative level based on shallow analysis of input. For example, it initiates requests for confirmations if speech recognition confidence scores are low. This module initiates utterances when the user yields the floor, regardless of whether the system has a complete plan of what to say or not. Using similar strategies as the subjects recorded here, the dialogue system can grab the turn and start to say something before having completed processing input. Many cue phrases were used in combination, signalling function on different discourse levels; first a simple responsive, saying that the previous message was perceived, and then some type of connective to signal how the new contribution relates.

6 Final remarks

Since DEAL focuses on generation in role play, we are less interested in the ambiguous cue phrases and more concerned with the instances where the annotators agreed. The DEAL users are second language learners with poor knowledge in Swedish, and it may even be advisable that the agent's behaviour is exaggerated.

[1] For more details on the discourse modeller see Skantze et al, 2006.

Acknowledgments

This research was carried out at Centre for Speech Technology, KTH. The research is also supported by the Swedish research council project #2007-6431, GENDIAL and the Graduate School for Language Technology (GSLT). Many thanks to Jenny Klarenfjord for help on data collection and annotation and thanks to Rolf Carlson, Preben Wik and Jens Edlund for valuable comments.

References

G. Aist, J. Allen, E. Campana, L. Galescu, C. A. Gómez Gallo, S. Stoness, M. Swift, and M. Tanenhaus. 2006. Software Architectures for Incremental Understanding of Human Speech. In *Proc. of Interspeech*.

J. Arnold, M. Fagano, and M. Tanenhaus. 2003. Disfluencies signal theee, um, new information. *Journal of Psycholinguistic Research*, 32, 25-36.

J. Beskow. 2003. Talking heads - Models and applications for multimodal speech synthesis. *Doctoral dissertation, KTH*.

S. Brennan. 2000. Processes that shape conversation and their implications for computational. In *Proc. of the 38th Annual Meeting of the ACL*.

H. Clark, and T. Wasow. 1998. Repeating words in spontaneous speech. *Cognitive Psychology*, 37(3), 201-242.

J. Hirschberg, and D. Litman. 1993. Empirical studies on the disambiguation of cue phrases. *Computational Linguistics*, 19(3), 501-530.

A. Hjalmarsson, and J. Edlund. In press. Human-likeness in utterance generation: effects of variability. To be published in *Proc. of the 4th IEEE Workshop on Perception and Interactive Technologies for Speech-Based Systems*. Kloster Irsee, Germany.

A. Hjalmarsson, P. Wik, and J. Brusk. 2007. Dealing with DEAL: a dialogue system for conversation training. In *Proc. of SigDial*. Antwerp, Belgium.

J. Lindström. 2008. Diskursmarkörer. In *Tur och ordning; introduktion till svensk samtalsgrammatik* (pp. 56-104). Norstedts Akademiska Förlag. Stockholm, Sweden.

L. Schourup. 1999. Discourse markers. *Lingua*, 107(3-4), 227-265.

G. Skantze, J. Edlund, and R. Carlson. 2006. Talking with Higgins: Research challenges in a spoken dialogue system. In *Perception and Interactive Technologies* (pp. 193-196). Berlin/Heidelberg: Springer.