Utterance Generation in Spoken Dialogue Systems

Anna Hjalmarsson
Centre for Speech Technology
Department of Speech, Music and Hearing, CSC, KTH
annah@speech.kth.se
May 2006

Abstract

Generation of utterances in spoken dialogue systems often rely on simple, non-flexible methods such as templates or "canned text". However, spoken dialogue systems are becoming increasingly complex with new and more conversational domains. This new generation of spoken dialogue systems puts new demands on spoken dialogue systems to produce flexible and context aware output. The focus of this paper is research which tries to come up with strategies to generate more flexible spoken output in dialogue. A number of issues related to utterance generation and the specific characteristics of dialogue are discussed. The issues considered are: incrementality, form and content, fragmental utterances, referring expression generation, grounding, speech synthesis, lexical entrainment, adaptation and turn-taking. Different approaches to utterance generation and the relation between more traditional natural language generation and utterance generation are also discussed.

1. Introduction

A natural part of human conversation is to adapt what we say and how we say it depending on our conversational partners and the dialogue context. This includes syntactic, semantic and lexical variation. For machines to be perceived as natural conversational partners the system output needs to be coherent with the current state of the dialogue. However, the main effort within natural language processing has been put on the processes of understanding rather than on those responsible for generation. Moreover, the research area of Natural Language Generation (NLG) has mainly been concerned with the generation of coherent text and monologues rather than the generation of utterances in dialogue. Dialogue system developers have used basic non-flexible generation methods which work in simple systems designed for limited domains. NLG in dialogue has therefore been regarded as non-problematic and gained less attention as a research area. However, spoken dialogue systems are becoming increasingly complex with new and more conversational domains such as computer games. This new generation of spoken dialogue systems puts new demands on generation to produce flexible and context aware output. The focus

of this paper is research related to NLG in spoken dialogue systems which tries to meet these demands. This rest of this paper is structured as follows. Section 2 gives an overview of the chain of processes involved in more traditional natural language generation. Section 3 brings up issues related to spoken language generation in dialogue. Section 4 presents different approaches to utterance generation. Section 5 provides a few final remarks.

2. Natural Language Generation

Natural language generation is the process of deliberately constructing some kind of natural language output (speech or text) from a non-linguistic representation in order to meet some specified communicative goals (Jurafsky and Martin, 2000). In some sense NLG can be viewed as the inverse of natural language understanding (NLU) (Dale and Mellish, 1998). NLU transforms natural language input into abstract representations of meaning which can be processed by computers while NLG transforms representations of meaning into natural language (Figure 1).

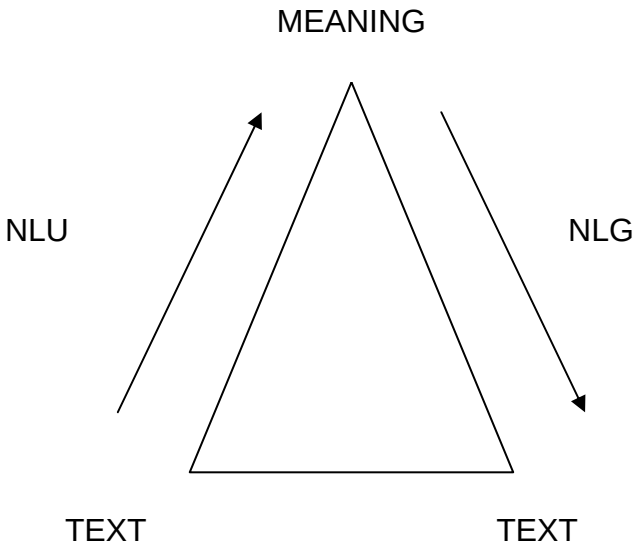

Figure 1 : The processes of NLU and NLG

In NLU the focus is on *hypothesis management*, ruling out possible interpretations of natural language input and determining which interpretation is the most appropriate one. When doing this the system has to deal with ambiguity, under-specification and ill-formed input. The focus of NLG is *choice*, i.e. choosing between different ways of realising a message given a specific context. Applications of NLG include document production such as weather forecasts and letters, dialogue systems, text summarization, machine translation and question answering systems. NLG spans research in several disciplines including linguistics, psychology, engineering and computer science.

The broad challenges of natural language generation are to decide what to say and how to say it. Most research within NLG has been concerned with producing monologues as text or monologues to be synthesised as speech without user interruption. Compared to text documents, dialogue is produced on line in collaboration with a human interlocutor and timing is an important factor. Due to the specific characteristics of dialogue, methods for monologue generation cannot be directly applied for spoken utterance generation in dialogue systems. In the rest of this paper this task will be referred to as utterance generation (UG) to separate from the task of more traditional monologue generation. First, the basic chain of processes involved in monologue generation and their relation to UG will be discussed.

2.1 Process stages in natural language generation

Computers were able to produce natural language output years before they were able to process natural language input (Jurafsky and Martin, 2000). These early NLG systems were based on canned text or templates. Canned text is the simplest approach where predefined texts or pre-recorded utterance are used. Template based systems allow a bit more flexibility with slots to be filled. Templates are often used for generation of personalized letters or utterances in database access systems such as time table information applications. Canned text and template filling are straightforward methods to implement but difficult to reuse and tedious to maintain (Theune, 2003b). More important, these methods are inflexible and have little to do with natural language produced by humans.

It is not obvious where the processes of NLG start and which knowledge sources they rely on. Within the research area of NLG there is no simple consensus on what the input of a generation component should be and its form and content vary between systems. Neither is there a universally accepted architecture. However, in general terms the processes are often split up into three stages: document planning, microplanning and surface realization. Figure 2 demonstrates a reference architecture as defined by Reiter and Dale (2000).

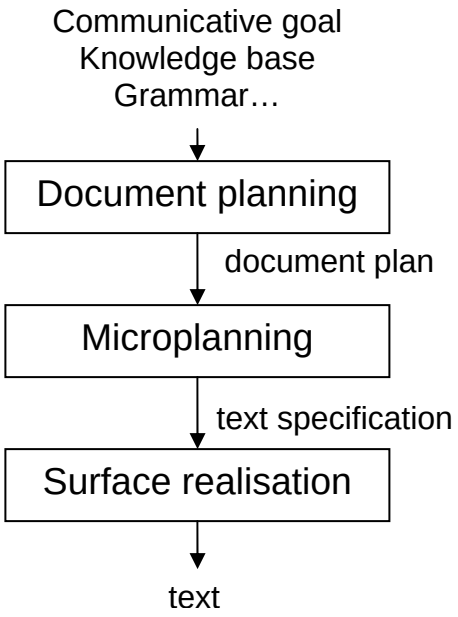

Figure 2 : A pipeline NLG architecture

2.1.1 Document planning

The phase of document planning is divided into processes for *content determination* and *document structuring*. The task of content determination is to produce data objects (or messages) which can be manipulated by the components responsible for micro planning and surface realisation. These messages are normally non-linguistic semantic representations based on domain and system specific knowledge. In dialogue these units are often represented in terms of the utterance's communicative goal and some aspect of its semantic content (Theune, 2003b). The task of document structuring is to decide the structure of a document, i.e. ordering and chunking the text. Document planning is generally language independent but domain specific. In spoken dialogue systems the dialogue manager is generally responsible for both these tasks, i.e. determining what to say in the next utterance and when to say it.

2.1.2 Micro planning

Microplanning include processes for *lexicalisation*, *aggregation* and *referring expression generation*. Lexicalisation is the task of putting words to the concepts in the abstract message. A concept can often be expressed in many different ways and the task of lexicalisation is to choose the word which is most appropriate in the given context. In utterance generation this can include considering which words have been used previously in the dialogue. Studies have shown that humans adjust their vocabulary and expressions to each other (Garrod and Andersson, 1987). It is likely that a dialogue system would benefit from behaving in a similar way and adopt the lexical choices of the user. This will be discussed further under Lexical entrainment. Aggregation is the

processes responsible for structuring the document linguistically into paragraphs and sentences. For example the sentences “Johan has a book” and “Mary has a book” can be written as “John and Mary have a book”. Aggregation can also include determining how the information should be ordered. Utterance aggregation differs from monologue aggregation in that sense that utterances are typically short and consist of only one sentence or even shorter fragments. So rather than structuring the messages into sentences and paragraphs utterance aggregation is about making the utterances more concise (Appelt, 1985). Referring expression generation is determining which expressions to use when referring to entities; definite descriptions or pronouns. A dialogue system which only uses definite descriptions will likely be experienced as repetitive and a waste of time. Unjustified pronoun use on the other hand can cause misunderstandings which will be experienced as frustrating and which can be difficult to recover from.

2.1.3 Surface realisation

The surface realisation component can be divided into processes for linguistic *realisation* and *structure realisation*. Linguistic realisation refers to the processes that convert abstract representations into real text. This includes applying syntactic and morphological rules to form “grammatically correct” texts. Structure realisation is an additional task which is related to converting the text to some external format such as XML, HTML or Latex-codes (Theune, 2003b). The reason for doing structure realization can be to meet the format requirements of the text-to-speech-synthesis. For natural and context aware spoken output some kind of prosodic markers are needed. The acoustic realisation of speech will be discussed more later on.

3. Utterance generation in spoken dialogue systems

Spoken dialogue systems allow users to interact with a computer-based application by natural spoken language. A spoken dialogue system takes natural human speech as input and produces speech, synthesized or pre-recorded human speech, as output. The basic chain of processes includes automatic speech recognition, natural language understanding, dialogue management, utterance generation and text to speech synthesis. The task of the dialogue manager is to control the flow of the dialogue. This includes determining if the system has elicited adequate information from the user, contextual understanding, information retrieval and response generation. The response generated by the dialogue manager is normally on a non-linguistic semantic level. The utterance generation component takes this abstract representation and produces a surface generation (often in some textual form) which can be passed on to the text to speech synthesis component and transformed it into speech. The process includes deciding which information should be included, how the information should be semantically structured and its syntactic structure. The response generation component can use simple pre-defined templates or complex natural

language generation. The generation process is sometimes divided into deep and surface generation, where deep generation refers to the decisions related to content selection ("what to say") and surface generation refers to the realization of the utterance into actual text ("how to say it").

3.2 Challenges in Utterance Generation

Pre-recorded speech and simple templates are straightforward methods which are easy and fast to implement. However, utterance generation based on these simple methods might be experienced as unnatural and confusing by the users since the output is restricted to inflexible utterances. Compared to language understanding little research has been done in the area of utterance generation. As dialogue systems become more complex, natural and flexible, more advanced techniques for adaptive utterance generation need to be investigated, which to a larger extent takes the dialogue context and the current user into account. This section brings up some of the issues which need to be considered for the generation of natural and context aware utterances in spoken dialogue systems.

3.2.1 Incrementality

When generating documents, text summarization or doing machine translation the contents is known in advance. However, speech has to be produced on-line as the dialogue proceeds and once something has been said there is no possibility to go back and make changes. Utterances in dialogue are really just streams of words from two or more people. The example below shows that a grammatical construction may be interrupted by a "sub-dialog" and then continued (taken from Skantze, 2005b).

U: I have a large building

S: building?

U: mhm, on my left

Human speech is characterised by pauses, false starts, and hesitations. These disfluencies suggest that we generate speech incrementally using information from several different sources in parallel (Brennan, 2000). For machines to produce natural, flexible speech and recover from errors in a similar way they need to generate utterances and perform interpretation incrementally. The TRIPS system with a central message-passing hub is an example of an architecture which supports incremental processing (Stent, 2001). The generation processes are distributed over several modules and since they are independent and concurrently running processes each module can start to process input immediately. This support incremental processing where the system can start to plan the next utterance even before the user stops speaking.

The realization of utterances by the text-to-speech synthesis also needs to be done incrementally. If the system is interrupted in the middle of an utterance (a barge-in) the system needs to keep track of what it planned to say and what was

actually said before the interruption (Skantze, 2005b). Most dialogue systems today only keep track of what they planned to say and have no knowledge about which information was actually passed on to the user.

3.2.2 Form and content

Humans engage in dialogue for a number of different reasons. Except for more functional motivations such as providing someone with information or requesting information we also have social reasons. An illocutionary act can be realised in a number of different ways but its realisation is not arbitrary (Clark, 1996). Depending on what we want to communicate we consider the different linguistic choices available in the current context. According to Appelt (p. 2, 1985): "the distinction between 'what' and 'how' then becomes merely a two points in a continuum between goal-satisfaction and rule-satisfaction processes, and no modularisation based on the distinction is obvious". An abstract semantic representation can have several different linguistic realisations. The surface realisation is often based only on the abstract semantic representation and completely separated from the previous utterance and the rest of the dialogue history. For the linguistic realisation to be closer related to previous utterances the surface realisation components needs access to the discourse model or the dialogue manager needs to provide the information relevant to surface realisation. Relevant information can be which words have been used and how well different concepts are grounded.

3.2.3 Referring expressions and fragmental utterances

One choice to be made when realizing a message is between using full utterances, elliptical constructions and using different anaphoric expressions. The use of pronouns is an efficient way for the speaker to refer to entities which somehow are given by context. However, pronouns can lead to errors if the listener is unable to identify the entity or misunderstands which entity the pronoun refers to. On the other hand dialogues without referring expressions would be experienced as unnatural and tedious.

Fragmental utterances such as ellipsis and fragments are very common in spoken dialogue. Like referring expressions the benefit of fragments is efficiency. A study on the prosody of fragmentary synthesized utterances support that the prosodic realisation affects the user's following utterance accordingly (Skantze, House and Edlund, 2006). This implies that such utterances can be successfully used in human machine interaction to make the interaction more efficient. According to Schlangen (2003) about 10% of the utterances in natural dialogues are fragmental (see example taken from Schlangen, 2005).

A: Who came to the party?
B: Sandy

Fragments can often be replaced with full readings. In the example above B's fragmental utterance could also be realised as "Sandy came to the party". However, in many dialogue systems the abstract representation generated by the discourse planner (typically the dialogue manager) will be the same for both of these utterances. If the surface realiser only has access to this abstract representation generation of fragments will be problematic.

3.2.4 Grounding

In order for the conversational partners to understand each other *common ground* needs to be established (Clark, 1996). Common ground in conversation is the things we know about each other and that we know that the other one knows. According to Clark a proposition *p* is common ground if all people involved in a conversation know *p* and that they all know that they all know *p*… and so on. In dialogue we cannot rely on that what is said by one speaker is immediately know by the other conversational partners. To establish common ground we react to each others utterances and signal understanding or non-understanding. The processes by which speakers try to establish a common ground is referred to as *grounding*. The level of mutual understanding in a dialogue, the common ground, needs to be considered when generating utterances in spoken dialogue systems. Information about how well entities have been grounded in discourse can for example be used to make the choice between using full expressions, fragments or pronouns. Short feedback utterances such as "yeah" or "uhuh" as well as full propositional statements all contribute to the grounding of entities. The choice of feedback level, i.e. how explicit the system needs to be when confirming a user utterance, needs to be sufficient for the utterance to be perceived as common ground. A study of the Monroe corpus (Stent, 2001) showed that one third of the utterances in human human dialogue simply manages the turn or perform grounding functions. Suppose a user is talking to a train timetable information system and wants to book a train ticket to Stockholm. There are several different ways to ground the information provided by the user. Which realisation should be used depends on how confident the system is in its interpretation:

U: I want to go to Stockholm.

S-A: Ok. Where do you want to go from?
S-B: Ok
S-C: To Stockholm.
S-D: Ok, you want to go to Stockholm.
S-E: To Stockholm?
S-F: Did you say that you want to go to Stockholm?

In A the system implicitly confirms the user's previous utterance and immediately takes the initiative and introduces new information. D and E are examples of requests for the user to confirm back. The amount of feedback that is required depends on the consequences of a potential misunderstanding and on

how confident the listener is in the interpretation of the overall utterance. In Higgins (Skantze, 2005a) the discourse modeller keeps track of the level of grounding status by indicating how well established the entities are in the dialogue model. Grounding status here includes information about who added the concept, in which turn the concept was introduced and how confident the system is in the concept. The confidence score is based on the confidence scores from the automatic speech recogniser. Entities are referred to many times during a dialogue and grounding status over time can be used to adapt confirmation strategy or the level of system initiative. In the OVIS dialogue system all new information provided by the user is considered as common ground only after being explicitly confirmed by the system (Theune, 2003a). The system always uses full expressions when referring to non-confirmed information (*pending* information).

3.2.5 Speech synthesis

The acoustic realisation of utterances will not be discussed in detail in this paper. However, a few remarks are needed since the acoustic realisation of an utterances influence the listener's perception of the message. The output of the generated text is realised by a text-to-speech (TTS) engine. A simple text is the most basic input to such an engine. However, to generate more context aware speech with the appropriate stress, rhythm and vocal stress more information is needed than the simple text string. Systems which generate speech also have to deal with homographs, words that have the same spelling but with different pronunciation. To automatically generate context aware prosody in more flexible systems without a fixed set of utterances is problematic. Part of speech tags can be used to resolve homographs and distinction between questions and non-questions can be used to generate appropriate question intonation (Jurafsky and Martin, 2000). Prosody is also crucial for the generation of non-lexical utterances such as "hmm", "um" and "aha" (Ward, 2004). The grounding functionality of these utterances is mainly conveyed through prosody. Emotion is another dimension which can be expressed acoustically. Emotional speech can be valuable if spoken dialogue is to be implemented in entertainment applications.

3.2.6 Lexical entrainment

Natural language offers a vide range of lexical choices. One simple statement can be expressed in many different ways. However, in human human dialogue we tend to coordinate our linguistic behaviour. These processes in which conversational partners adopt each others terms and achieve conceptual pacts are called *lexical entrainment* (Garrod and Andersson, 1987). Zoltan-Ford (1991) studied lexical entrainment in a human machine context to see whether the system's output influenced the users' vocabulary and phrase structure. The results showed that the subjects modelled the length of the program's output and that the degree of "shaping" or modelling was not affected by mode of communication (text or speech) or output vocabulary.

Since users tend to imitate the vocabulary of the system a basic design principle is to always design the system to "understand all the words it can say", i.e. put all lexical entries which the system is capable of generating in the vocabulary of the speech recogniser. Brennan (1996) discusses lexical entrainment in both human human and human machine interaction and showed that people are at least as likely to adopt the vocabulary of their computer partners as of their human partners. The subjects in the study adopted the system's lexical choices using both text and speech interfaces. They were more likely to use the same term as the system when it was presented immediately after the first term than when presented later in the context. To which extent the users were "shaped" was also influenced by how the system's terms were exposed, implicitly ("embedded") or explicitly ("exposed"). The useful aspect of lexical entrainment in dialogue systems is that the users' lexical choices will be more predictable. However, when humans coordinate their linguistic behaviour in dialogue the coordination goes both ways but within dialogue systems lexical entrainment has mainly been studied in terms of constraining the user's lexical choices (Brennan, 1996). For spoken dialogue systems to appear "natural" from a user point of view the system should also adopt the user's lexical choices. In Lemon et al. (2003) the system chooses noun phrases depending on what phrases was previously employed by the user.

3.2.7 Adaptation

A dialogue system should say the right thing, in the right way, at the right time, to the right user. The adoption of user terms is one way to adapt the system to individual user behaviour. The quality of the interaction in spoken dialogue systems differs between different users and even for the same user between different occasions. Research within the area of adaptive spoken dialogue systems have mainly been concerned with modelling and adjusting to a single aspect of the user, such as user level of expertise and user preferences. Other than adaptation of lexical choices there are dialogue systems which adapt strategies such as shifts in dialogue strategy and initiative (system/user) to adjust its behaviour to the individual user or dialogue context (Chu-Carroll, 2000), (Litman and Pan, 2002), (Komatani, Ueno, Kawahara and Okun, 2003). Extended knowledge about how humans adapt in dialogues is needed to build fully context aware adaptive spoken dialogue systems.

3.2.8 Disfluencies

Human speech is characterised by disfluencies such as pauses, hesitations and false starts. These disfluencies are often considered as "mistakes" or "interruptions" in the messages to be delivered. However, in an experimental study Brennan (2000) has studied the comprehension of disfluent and fluent speech and showed that disfluencies can bear valuable information. The subjects were presented with repair utterances with edit intervals of different lengths which consisted of either silence or a filler. The results showed that subjects' responses (the target word was the word being corrected) were faster when the edit interval was longer and contained a filler. This suggests that disfluent

speech is faster processed and may be easier to understand in certain contexts. Unlike humans, dialogue systems often speak in perfectly fluent and "grammatically correct" utterances. If disfluent speech is easier to comprehend and the system would be perceived as more natural it could be beneficial to intentionally introduce disfluent utterances in a dialogue system. Callaway (2001) mention that to implement disfluent speech in spoken dialogue systems more complex generators and statistical or syntactical models of how humans produce disfluent speech are needed.

3.2.9 Multimodality

Multimodal generation will not be discussed in detail since it is a research area all on its own and not within the scope of this paper. Issues which need to be considered in multimodal generation are synchronization and choice of modality. Modalities can substitute each other or two or more modalities can be used in combination to reinforce the message or to express different parts of it. If they are to be used in combination, especially if communicating different semantic parts of an utterance, the coordination of modalities is crucial for how the message will be perceived. The McGurk effect, which is described in (McGurk, 1976), has shown that we integrate visual articulary information into what we "hear" and that unsynchronized speech and visual articulation movements affects our perception of speech sounds. In multimodal generation there is also a choice of which modality to use.

3.2.10 Turn-taking

The challenges of utterance generation is not only determining what to say and how to say but also about deciding when to say it. In mixed initiative dialogue systems, where both the system and the user can take the initiative, the system needs to know when it is appropriate to grab or release turn, i.e. deciding when it is appropriate for the system to say something and when to silently wait for user input. The system also needs to have knowledge about what kind of dialogue moves can be used to grab, hold and release turn. In Lemon (2003) a three valued turn marker is used to indicate who has the initiative: the system, the user or neither. Turn management relies on a number of rules related to different dialogue moves. For example questions always swap the turn while answers release it. The agent which has the turn automatically looses it after ten seconds of silence. In dialogue systems the end of user utterances are often triggered by a certain amount of silence. However, in human human dialogue there are often long silences inside utterances and a human interlocutor would not consider these silences as appropriate places to grab the turn (example taken from Edlund, Heldner and Gustavsson, 2005).

> "I am standing to the left of a <long silence> brown building"

Edlund, Heldner and Gustavsson (2005) introduce /nailon/, a component which acoustically analysis prosody and chunk the dialogue into what humans would perceive as utterance like units.

4 Approaches to utterance generation

In this section different approaches to utterance generation are discussed. The simple template filling methods which were used in the early natural language generation systems are still being used in many dialogue systems. They are straightforward methods which require little linguistic expertise. Other approaches to utterance generation are rule-based methods which originate from more traditional natural language generation, corpus based methods and trainable generation (Stent, Prasad and Walker, 2004). The focus here is methods based on more traditional natural language generation but corpus-based and trainable methods are also mentioned. First, the processes of more traditional NLG in relation to UG are considered. A reversed parsing approach, stochastic methods, trainable generation and evaluation of NLG systems is also discussed.

4.2 The processes of utterance generation

There is no general structural design of utterance generation components since most dialogue systems have individual architectural solutions. However, a rough overview of the tasks involved in utterance generation is presented in Figure 3. The process are divided into two different components; the dialogue manager which is responsible for determining "what to say" (content planning) and the surface realiser responsible for "how to say it" (surface realisation). In multimodal systems processes for determining which modality to use needs to be included in the model. Wilcock and Jokinen (2001) use a pipeline architecture similar to Figure 3. Their input message is specified in XML and the different UG processes are implemented using XSL transformations (XSLT) which operate on the message.

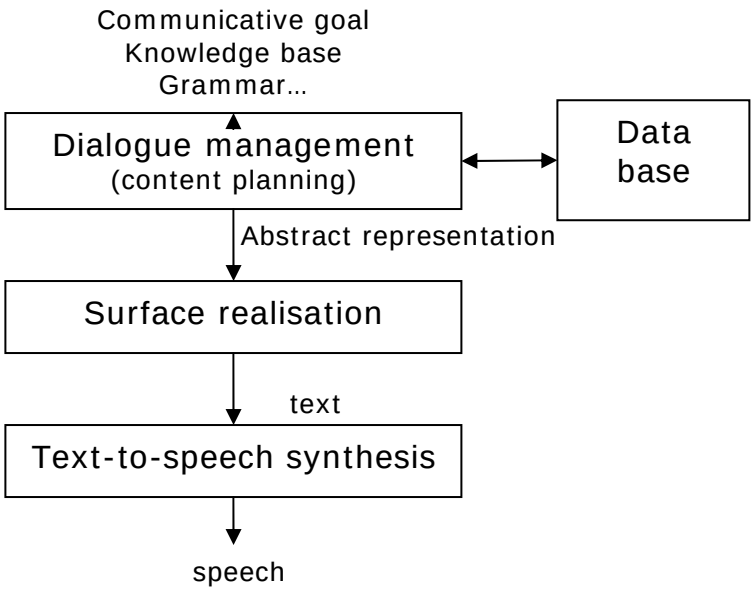

Figure 3 : NLG in spoken dialogue systems

4.2.1 Input

There is no overall consensus on what the input to an NLG system should be. In the generation of text documents the entire contents is available from start and there is a possibility to go back and make changes. However, in dialogue systems, when constructing and utterance, there is no information about the utterances which will follow and we need to rely on the previous dialogue discourse. To keep track of previous utterances the system needs to integrate these in a discourse model. In dialogue system the input for the natural language generator component is in general an abstract semantic representation of the message to be realised. This message comes from the component responsible for content planning, which is generally the dialogue manager (Theune, 2003b). This abstract message can be realised in several different ways using different lexical entities, ellipsis, anaphora or full readings. To choose between different surface realisations the natural language generation component needs access to information available in the discourse model. This information can include the user's lexical choices, user models, grounding status and ASR confidence scores.

4.2.2 Content planning

In many spoken dialogue systems the component responsible for utterance generation does not include processes for content planning. Instead, these tasks are performed by the dialogue manager and are therefore not viewed a part of the actual generation component (Theune, 2003b). Why the processes for content planning have not been generalized into a separate module for utterance generation is most likely explained by the fact that dialogue systems deal with narrow domains. The information needed and the processes responsible for content selection are therefore highly domain dependent and difficult to generalize into a separate module(s). Examples on how the semantic content can be organized are RST relations, tree structures, communicative goals, or attribute value pairs (Oh and Rudnicky, 2000) (Stent, 2001) (Skantze, 2005b).

4.2.3 Surface realisation

Surface realisation includes process for linguistic realisation and structure realisation. However, in spoken dialogue system surface realisation often refers to all processes which are not involved in content planning. The tasks performed during micro planning and surface realisation are also relevant for utterance generation. However, depending on the specific characteristics of dialogue, the motivation and challenges of these will be different. Aggregation and referring expression generation are often not considered at all in utterance generation. These processes might not be needed when the vocabulary is small, the system has the initiative and the focus of the dialogue is to solve a task within a limited domain. Still, in mixed-initiative systems which allow freer conversation, aggregation and referring expression generation are valuable.

Aggregation

According to Appelt (1985) the task of aggregation in utterance generation is to make the utterances more concise, avoid repetitious language and make the system more understandable. Since the system in general has no information about the utterances that will follow aggregation has to be done incrementally on utterance level. Lemon et al. (2003) have implemented aggregation processes in an incremental fashion. Since future utterances are unknown the only way aggregation can be performed is by "retro-aggregating" new utterances with previous ones (see example).

System:	I have cancelled flying to the base
System:	and the tower
System:	and landing at the school.

Referring expression generation

The choice of noun phrase should be done in order to provide the reader/listener with "sufficient" information to identify the intended referents. Much work on pronouns in computational linguistics has focused on anaphora resolution and the parsing of pronouns rather than how they are generated. The centering theory (Grosz and Sidner, 1986) which originates from anaphora resolution has been leading also in the development of NLG systems. According to centering theory some entities in an utterance are more central and this imposes constraints on the use of referring expressions. The leading assumption has been that a pronoun should be used whenever referring to an entity which is highly prominent in the local discourse. However, research which focus on the generation of pronouns argues that the centering theory does not account well for patterns of pronouns in natural occurring texts (Callaway and Lester, 2001) (McCoy and Strube, 1999). Sentence boundaries, distance from last mention, discourse structure and ambiguity are identified as factors which influence pronominialization. In spoken dialogue pronoun usage is also characterised by the relationship between the speaker and hearer ("you", "I"), how well entities have been established in the context and vocal stress. The acoustic aspect of speech is an extra dimension which can be used to stress prominence in an utterance which is not revealed by its syntactic structure.

4.3 Reversed parsing

Reversed parsing is an approach to natural language generation which utilizes the similarities between parsing and generation. In dialogue short fragmental input can only be interpreted in the context of the previous or future utterance. An example of such fragmental utterances in the travel reservation domain is:

A:	Where do you want to go?
B:	Stockholm

The full reading of these two utterances is: “person B wants to go to Stockholm”. Purver and Kempson (2004) point out that these “shared utterances” are problematic if generation and parsing are divided into separate disconnected processes. Besides the problems of parsing shared utterances separately they argue that the cognitive processes of parsing and generation in humans are closely related.

The processes of natural language generation can on an abstract level be viewed as the inverse of natural language understanding (NLU). In NLG abstract computer representations are converted into natural language while in NLU natural language is converted into abstract computer representations. In ‘Reversed parsing’ a single grammar is used for semantic interpretation of user input and for generating utterance strings from their abstract representation. According to Shieber (1998): “parsing and generation could be viewed as two processes engaged in by a single parameterized theorem prover for the logical interpretation of the formalism” (p.614). Purver and Otsuka (2003) suggest a method for generation based on incremental parsing where the generator component share the same lexical entries, the same context and the same semantic tree representations as the parser. This architecture enables a smooth transition from speaker to hearer (from generator to parser). The idea, as Shieber describes it, “is an idea with a certain elegance”, still it is not as unproblematic as it first might seam. The grammatical rules in the parser need to partly cover ungrammatical input but it is questionable if the system should generate incomplete and ungrammatical output. There is also the problem of mapping between logical forms and natural language expressions which is generally not a one to one relation (Shieber, 1988). There are often several possible linguistic realisations for one logical form and a system which always uses the same utterance for specific semantic representation might be experienced as repetitive.

4.4 Stochastic methods

Machine learning and stochastic methods are becoming widely used for a number of different tasks within spoken dialogue system research as well as in the rest of the natural language processing community. To generate output based on existing corpora seems extra tempting since the reference answer (the desired output) is explicitly available. Furthermore, rule-based and template-based systems require a lot of manual work and can be difficult to generalize and apply to new domains. Corpus-based methods are also promising if we want to model features of human human dialogue such as Disfluencies, fragments and pronoun use. Oh and Rudnicky (2000) have developed a corpus-based surface realiser which models human speech. The system is a hybrid model in which simple utterances such as greetings are “canned expressions” and more complex utterances are generated based on an n-gram language model. The n-gram language model is used to predict the next word in the utterance. In an evaluation of the system a trend implied that the subjects preferred the stochastic generation compared over a template based generation. Another

corpus based approach is a hybrid language generator which combines finite state machine grammars and corpus based language models (Galley, Fosler-Lussier and Potamianos, 2001).

4.4.1 Trainable generation

In trainable generation an utterance planner produces a candidate set of utterance surface realisations which is later ranked by an utterance-plan-ranker. The initial list is often based on general-purpose linguistic knowledge and the "training" is used to automatically adapt the natural language generator to a particular domain or a group of users. The tasks of the utterance trainer include deciding the syntactic structure of the system and aggregation. SPoT, "Sentence Planner, Trainable", is trainable generator which is automatically trained based on feedback provided by human judges (Walker, Rambow & Rogati, 2001). SPoT learns to select a sentence plan which average is only 5% worse than the top human-ranked sentence plan which is on average 36% better than a random selector.

4.5 Evaluation

Evaluation of natural language generation methods is problematic for a number of reasons. First, the NLG processes are difficult to disconnect and evaluate separately from the rest of the system. The performance of the generation component will therefore be highly depending on previous processes such as language understanding and dialogue management. Dale and Mellish (1998) discuss the difficulties related to evaluating natural language generation components. Some of the issues which are brought up are: (1) the fact that there is no objective criterion of "goodness" which can be used to assess the natural language output, (2) the non-agreement and time consuming aspects of using human judges and (3) the problem of how to get adequate training and test data. Evaluation of generation components in spoken dialogue systems is even more complicated. The content planner is often an integrated part of the dialogue manager which makes it difficult to evaluate separately. Moreover, the language generation is difficult to separate from the quality of the speech synthesis. Oh and Rudnicky (2000) evaluates their stochastic language generator in a comparative study by running two identical systems varying only in the generation component. To evaluate their surface realization they presented calls to subjects where the human operator's utterances were substituted with two different versions of systems responses, one template-based, generation and one stochastic generation.

5 Final remarks

Natural language understanding has gained far more attention than the area of natural language generation and compared to monologue generation the effort put on dialogue generation is negligible. However, recently there has been an increased research effort on utterance generation which considers the specific characteristics of dialogue (Theune, 2003), is more sensitive to the local dialogue context (Lemon, 2003) and with better suited architectural solutions (Stent, 2001). Studies of human human dialogue has also contributed with important knowledge (Brennan, 2000), (Schlangen and Lascarides, 2003). Further research on human human dialogue is needed to obtain better models of human communication which can be used to generate fragmental utterances and disfluencies in spoken dialogue system. Better methods for evaluation, domain independent solutions and standardized processes are also needed in order to persuade designers of commercial dialogue system to abandon pre-recorded speech or template based methods. More intelligent utterance generation will improve the overall quality of the dialogues and make interaction more efficient and natural. Context aware output avoids telling the users what they already know, help the users grasp the conversational status, and provides them with a sense of a mutual understanding. Intelligent output can also be used to detect and diagnose errors and help the users understand the potential as well as the limitations of the system.

References

Appelt, D. E. (1985) Planning English referring expressions, *Artificial Intelligence*, 26(1):1-33

Brennan, S. E. (1996). Lexical entrainment in spontaneous dialog. In *Proceedings of ISSD*, 41-44.

Brennan, S. E. (2000). Processes that shape conversation and their implications for computational linguistics. In *Proceedings of the 38th Annual Meeting of the ACL*. Hong Kong.

Callaway, C. and Lester, J. (2001) Pronominalization in generated discourse and dialogue, In *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*. Philadelphia, Pennsylvania

Callaway, C. (2003) Do we need Deep Generation of Disfluent Dialogue?, In *Proceedings of the 2003 AAAI Spring Symposium on Natural Language Generation in Spoken and Written Dialogue*, Stanford University.

Chu-Carroll, J., (2000) MIMIC: An adaptive mixed initiative spoken dialogue system for information queries. In *Proceedings of the 6th ACL Conference on Applied Language Processing*.

Clark, H. (1996) Using Language, Cambridge University Press.

Dale, R. and Mellish, C. (1998) Towards Evaluation of Natural Language Generation, In *Proceedings of the First International Conference on Language Resources and Evaluation*, Granada, Spain.

Edlund, J., Heldner, M. and Gustafson, J. (2005) Utterance segmentation and turn-taking in spoken dialogue systems. In B. Fisseni et al., Eds., *Sprachtechnologie, mobile Kommunikation und linguistische Ressourcen*, Frankfurt.

Galley, M., Fosler-Lussier, E., and Potamianos, A. (2001) Hybrid natural language generation for spoken dialogue systems, In *Proceedings of Eurospeech*, Scandinavia.

Garrod, S. & Andersson, A. (1987) Saying what you mean in dialogue: A study in conceptual and semantic co-ordination. *Cognition*, Vol. 27, 181-218.

Grosz, B. J. Sidner, C. L. (1986) Attention, intentions and the structure of discourse, *Computational Linguistics*, 12(3):175-204

Jurafsky, D. and Martin, J. (2000) Speech and Language Understanding: an introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition. Prentice-Hall.

Komatani, K. Ueno, S. Kawahara, T. & Okun, H. G. (2003) User modelling in spoken dialogue systems for flexible guidance generation. In *Proceedings of Eurospeech*.

Lemon, O., Gruenstein, A., Gullett, R., Battle, A., and Peters, S. (2003) Generation of collaborative spoken dialogue contributions in dynamic task environments, In *Proceedings of the AAAI Spring Symposium on Natural Language Generation in Spoken and Written Dialogue.*

Litman, D. & Pan, S. (2002) Designing and evaluating an adaptive spoken dialogue system, *International Journal of User Modelling and User Adapted Interaction*, 12.

McCoy, K. and Strube, M. (1999) Generating anaphoric expressions: Pronoun or definite description? In *Proceedings of the Workshop on the Relation of Discourse*.

McGurk H, MacDonald J. (1976) Hearing Lips and seeing voices, In *Nature*, 23-30, 264(5588):746-8.

Oh, A. H., and Rudnicky, A., I. (2000) Stochastic language generation for spoken dialog systems. In *Proceedings of the ANL/NAACL 2000 Workshop on Conversational Systems*, pages 27–32, Seattle. ACL.

Pickering, M., and Garrod, S. (2004) Toward a mechanistic psychology of dialogue. *Behavioural and Brain Sciences*.

Purver, M. and Otsuka, M. (2003) Incremental generation by incremental parsing: Tactical Generation in dynamic syntax. In *Proceedings of the 6th CLUK Colloquium*.

Purver, M. and Kempson, R. (2004) Incrementality, alignment and shared utterances. In J. Ginzburg & E. Vallduví (edd.), *Catalog '04 Proceedings of the 8th Workshop on the Semantics and Pragmatics of Dialogue*, Barcelona , 85-92’

Schlangen, D. and Lascarides, A. (2003) The interpretation of non-sentential utterances in dialogue. In *Proceedings of the 4th SIGdial workshop on Discourse and Dialogue*, Sapporo, Japan.

Schlangen, D. (2005) Modelling dialogue: Challenges and approaches, Künstliche Intelligenz

Shieber, S.M. (1988) A Uniform Architecture for Parsing and Generation. In *Proceedings of the 12th International Conference on Computational Linguistics*.

Skantze, G. (2005a) Natural language generation in spoken dialogue systems, Term paper, Course in Natural Language Generation, GSLT, Sweden.

Skantze, G. (2005b) GALATEA: A Discourse Modeller Supporting Concept-level Error Handling in Spoken Dialogue Systems", In *Proceedings of the 6th SIGdial Workshop on Discourse and Dialogue*, Lisbon, Portugal

Skantze, G., House, D., and Edlund, J. (2006) User responses to prosodic variation in fragmentary grounding utterances in dialog. Submitted to *Interspeech*

Stent, A., (2001) Dialogue Systems as Conversational Partners. Ph.D. thesis, University of Rochester.

Stent, A., Prasad, R. and Walker, M. (2004) Trainable sentence planning for complex information presentation in spoken dialog systems. In *Proceedings of ACL-2004*.

Theune, M. (2003a) From Monologue to Dialogue: Natural Language Generation in OVIS, In *Proceedings of the American Association for Artificial Intelligence (AAAI'03) Spring Symposium on Natural Language Generation in Written and Spoken Dialogue*, Palo Alto, CA, USA.

Theune, M. (2003b) Natural language generation for dialogue: system survey. *TR-CTIT-03-22*.

Reither, E. and Dale, R. (2000) Building Natural Language Generation Systems, *Studies in Natural Language Processing*, Cambridge University Press.

Walker, M., Rambow, O. & Rogati, M. (2001). Spot: a trainable sentence planner. In *Proceedings of the North American Meeting of the Association of Computational Linguistics.*

Ward, N. (2004) Pragmatic Functions of Prosodic Features in Non-Lexical Utterances. In *Speech Prosody 04*.

Wilcock, G. and Jokinen, K. (2001) Design of a generation component for a spoken dialogue system. In *Proceedings of the 6th Natural Language Processing Pacific Rim Symposium (NLPRS-2001)*, Tokyo.

Zoltan-Ford, E. (1991) How to get People to Say and Type what Computers can Understand. *International Journal of Man-Machine Studies*, 34(4), 527-547.