Matchmaking: dialogue modelling and speech generation meet*

Brigitte Grote
FAW Ulm
Germany
E-mail: grote@faw.uni-ulm.de

Eli Hagen
GMD/IPSI, Darmstadt and
Technical University of Darmstadt
Germany
E-mail: hagen@darmstadt.gmd.de

Elke Teich
University of the Saarland
Department of Applied Linguistics, Translation and Interpretation
Germany
E-mail: teich@darmstadt.gmd.de

Keywords: Speech generation, intonation

Abstract

This article is concerned with determining the constraints on the selection of appropriate intonation in speech generation in human-machine information seeking dialogues. The two pillars of our system—a state-of-the-art computational dialogue model and a state-of-the-art NL generator—are presented. Based on this, we determine the kinds of linguistic and pragmatic knowledge needed to sufficiently constrain choice in the intonational resources of the system. We take into consideration factors such as dialogue history, speaker's attitudes, hearer's expectations, and semantic speech functions.

1 Introduction

Task independent computational dialogue modelling (see *e.g.*, [5, 10, 35]) seldom makes contact with natural language generation (exceptions being, *e.g.*, [7, 9, 22]), and much less so with *speech* generation/synthesis. Conversely, speech synthesis, being predominantly concerned with rendering text to speech, rarely considers actual full scale generation.

In this article we introduce an approach under development in a joint collaborative project between the Technical Universities of Darmstadt and Budapest ('SPEAK!') that combines the dialogue modelling paradigm with NL generation and speech synthesis in an information retrieval system. The novelty of the approach pursued lies in the move away from *text-to-speech* and *concept-to-speech* generation towards *communicative-context-to-speech* generation (see Section 2) and the integration of dialogue representation, NL generation, and speech synthesis. Our principal concern is selection of appropriate intonation. More specifically, from our representation of communicative context, we derive constraints on *interpersonal* meaning, which are then expressed through *intonation contour* (or *tone contour* or simply *tone*).

We have taken two existing systems, the COR dialogue model ([31]) and the KOMET-PENMAN multilingual text generator [33] to build the backbone of an integrated dialogue-based interface to an information system. The linguistic generation resources of German have been enhanced by a systemic functionally [14, 15, 24] motivated grammar of speech that includes knowledge about intonational patterns [12, 34]. Section 3 presents our dialogue model and the intonational resources.

In Section 4 we first apply an bottom-up approach; we will determine the kinds of knowledge the generator needs to make intonational choices, and based on this we develop a stratified model with three strata: grammar, semantics, and extra-linguistic context. Second we apply a top-down approach; we determine how this knowledge can be obtained from the dialogue model and dialogue history, *i.e.*, from the extra-linguistic context, and thereby verify the applicability of our overall model. Section 5 concludes the paper with a summary and a number of questions that have been left untouched.

2 State of the art in speech generation

In this section we give a survey of existing speech generation systems for German, arguing that their syntax-based approach does not suffice to generate "natural" speech in dialogue systems.

In information-seeking dialogues that use spoken language for interaction, intonation is often the only means to distinguish between different dialogue acts, thus making the selection of the appropriate intonation crucial to the success of the information-seeking process (see *e.g.*, [26] for English). To illustrate this point, imagine an information-seeking dialogue where the user wants to know a specific train connection. At some point in the interaction, the system produces a sentence like *Sie fahren um drei Uhr von Darmstadt nach*

*Authors appear in alphabetical order. This work was partially funded by Copernicus, Project No. 10393 ('SPEAK!').

Heidelberg ("You travel at three o'clock from Darmstadt to Heidelberg."). There are several interpretations of this utterance, the most obvious being that the system presents some kind of information to the reader. However, the same sentence—employing a different intonation—could be part of a clarification dialogue, where the system wants to reassure that it got the user's request right. In this case, the user would be expected to react, i.e., either confirm or rebuke this statement. Only by means of intonation can the user interpret the system's expectation correctly and react accordingly.

Even though current speech synthesizers can support sophisticated variation of intonation, no existing *text-to-speech* or *concept-to-speech* system for German is available that provides the semantic or pragmatic guidance necessary for selecting intonations appropriately. The major shortcoming is that traditional text-to-speech systems (*e.g.,* [16, 18, 23]) and concept-to-speech systems [6] alike use purely syntactic information in order to control prosodic features. Moreover, with text-to-speech systems, where the syntactic structure has to be reconstructed from the written text by means of a syntactic analysis, the resulting data is seldom complete nor unambiguous. Concept-to-speech systems avoid the latter problem by generating spoken output from a pre-linguistic conceptual structure. Yet, most of the current implementations of the concept-to-speech approach use the conceptual representation only to avoid syntactic ambiguities with the assignment of intonational features still based on the written text (see [6]).

A common feature of all these systems is that they are often too expressive in that too many words are stressed, mainly due to the lack of discourse information, for instance on focus domain or the given/new distinction. A number of *discourse-model based* speech generation systems have been proposed that address exactly this problem, for example NewSpeak [17, 26]. However, the problem with these systems is that they still start from a given text, and are hence restricted to those kinds of discourse information that can be reconstructed from that text. Moreover, since they assume a one-to-one mapping between syntactic structure and intonational features, they cannot account for those phenomena frequent to our domain, where the same syntactic structure can be realized with differing intonations (see example above).

Assuming that intonation is more than the mere reflection of the surface linguistic form (see [14, 30, 19, 24]), and further, that intonation is selected to express particular communicative goals and intentions, an effective control of intonation requires synthesizing from *meanings* rather than word sequences as the discussed systems do.

This fact is acknowledged by [1], whose SYNPHONICS system[1] is based on the assumption that prosodic features have a function independent of syntax. [1] replace the idea of syntax-dependent prosody—which is implicit to all the approaches discussed so far—with the notion of the linguistic function of prosodic features including intonation. Thus, this approach allows prosodic features to be controlled by various factors other than syntax, *e.g.,* by the information structure such as focus-background or topic-comment structure. However, the function of intonation is still restricted to what is called *grammatical function*, more specifically the *textual* function of intonation, without considering aspects like communicative goals and speaker's attitude, *i.e.,* the *interpersonal* function of intonation ([14]).[2] Yet, in the context of generating speech in information-seeking dialogues where intonational features are often the only means to signal a dialogue act, these aspects have to be taken into account.

Furthermore, in a dialogue situation as given in our approach, it is not sufficient to look at isolated sentences; instead one has to look at the utterance in its context, as part of a larger interaction. Intonation is not only used to mark sentence-internal information structures, but additionally it can be employed in the management of the communicative demands of interaction partners. Therefore, we also have to consider the function of intonation with respect to the whole conversational interaction, taking into account the discourse (dialogue) history (see also [7]). Intonation as realization of interactional features thus draws on discourse and user model as the source of constraints.

An approach to speech generation that starts from *communication context* and maps this to intonational features is the only approach that provides the intonational control needed in dialogue systems to produce speech that human hearers would find acceptable.

3 Available resources

The overall system architecture for SPEAK! is shown in Figure 1. The text generation system (KOMET-PENMAN) receives input from a dialogue module (COR, dialogue history) and perhaps several other information sources (*e.g.,* confidence measure from a speech recognition unit), which will be made more precise below (see Section 4). Together the information from these input sources controls the traversal of the grammar (see Section 3.2). The KOMET-PENMAN grammar can generate two types of output: A plain text, which can be embedded into, for instance, a dialogue box in a graphical user interface and a text that is marked up with intonational features (see Section 3.2 for an example), which is passed on to the MULTIVOX text-to-speech system [23] and presented acoustically to the user.

In this article we develop a model of how the dialogue module can control the traversal of those regions of the

[1]The SYNPHONICS system ([1]) covers the incremental generation of utterances from pre-linguistic conceptual structures to the formation of syntactic and phonological structures, with an interface to a speech synthesis module for German.

[2]See [7] for an exception.

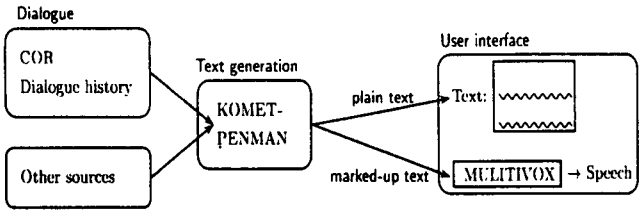


Figure 1: System architecture.

grammar concerned with intonation. As a basis for discussion, we introduce our dialogue model and the relevant parts of the grammar in detail.

3.1 The dialogue model

A dialogue model guides the interaction between a user and an information retrieval system, *i.e.*, it calculates a subset of possible dialogue acts that the user action (spoken or deictic) could correspond to, and on the system side it calculates those dialogue acts that would provide appropriate responses to a given user action. In the work presented here, we assume that a component exists that can choose one of the dialogue acts from these subsets (see *e.g.*, [13, 28].)

In the 'SPEAK!' project we have chosen to employ a modified version of the Conversational Roles model (COR) as our dialogue model (see [31]). COR is a task independent model based on Searle's speech act theory [29]. It has been modified within the 'SPEAK!' framework in order to include naturally occuring data that the original model failed to account for, but the overall speech act framework remains the same.

In the model, a dialogue is represented as a sequence of dialogue *moves* (*e.g.*, Request, Inform, Withdraw request), which are further decomposed into sequences of atomic *acts*, dialogue *moves*, and *sub-dialogues*. This recursive representation of a dialogue enables COR to account for mixed initiative dialogues, where both information seeker and information knower can employ, for instance, retraction, correction, and clarification tactics.

Below we present a simplified rewrite rule version of the dialogue model. In this version we only present the request, inform, and assert moves in detail, since the other moves are cast in the same format as the request, and one only has to insert new move names (*e.g.*, Promise → promise(K), (Dialogue(S)), *etc.*). Moves in parentheses are optional. The parameters indicate which participant can perform a given move, S=information seeker, K=information knower. *Moves* begin with upper case and *acts* with lower case. The first two rules encode the course that the dialogue is *expected* to take, while the other dialogue rules encode *exceptions*. (For a more detailed account see [31]).

```
Dialogue(S) --> (Request(S)), (Promise(K)),
Inform(K), (Evaluate(S)), (Dialogue(_))
Dialogue(S) --> (Offer(K)), (Accept(S)),
Inform(K), (Evaluate(S)), (Dialogue(_))
Dialogue(S) --> Offer(K), (Accept(S)),
WithdrawOffer(K), (Dialogue(_))
Dialogue(S) --> Offer(K), Accept(S),
WithdrawAccept(S), (Dialogue(_))
Dialogue(S) --> Request(S), (Promise(K)),
WithdrawRequest(S), (Dialogue(_))
Dialogue(S) --> Request(S), Promise(K),
WithdrawPromise(K), (Dialogue(_))
Dialogue(S) --> Offer(K), WithdrawOffer(K),
(Dialogue(_))
Dialogue(S) --> Offer(K), RejectOffer(S),
(Dialogue(_))
Dialogue(S) --> Request(S), WithdrawRequest(S),
(Dialogue(_))
Dialogue(S) --> Request(S), RejectRequest(K),
(Dialogue(_))
Dialogue(S) --> Withdraw(_)

Request(S) --> request(S), (Dialogue(K))
Request(S) --> request(S), (Assert(S))
Request(S) --> Dialogue(K)
Request(S) --> Assert(S), (request(S))
Request(S) --> Assert(S), (Dialogue(S))

Inform(K) --> inform(K), (Dialogue(S))
Assert(_) --> assert(_), (Dialogue(_))
```

Based on the dialogue model, the system builds up a tree-like dialogue history of the ongoing dialogue (see Section 4). Two central themes in our current work are to identify the relevant partial structures of such trees and to determine their semantics such that, for instance, the text generation system can search the dialogue history and interpret what it finds in order to guide the choice of intonation for the system utterances.

3.2 The intonational resources of the KOMET grammar

In this section, we describe the system networks that have been introduced to the German grammar of the KOMET-PENMAN text generation component as to include specifications of appropriate intonation selections in its output ([12]). The KOMET grammar of German ([32, 11]) is a computational NIGEL-style systemic-functional grammar, based on the notion of choice. The systemic-functional framework provides us with representational means for describing available choices and for mapping (even though indirectly) communicative goals to intonational features.

According to *systemic-functional linguistics* (SFL) (see [15, 21, 7]), intonation is just one means among others—such as syntax and lexis—to realize choices in the grammar.[3] This implies that choices underlying the realization of intonation may be organized in exactly the same way as other choices in the grammar (see [14, 7]. Hence, the intonational control required for speech generation in a dialogue system has been built into the existing KOMET grammar. The added discriminations are constraints on the specification of an appropriate intonation rather than constraints on the structural form.

Treatment of intonation in SFL The three distinct kinds of phonological categories, *i.e.*, *tone group*, *tonic syllable* and *tone*, contribute to the intonation

[3] [2, 30, 4, 8] consider intonation part of phonology.

specification of a clause (see for instance [4, 30, 24]). They signal three different kinds of relation between grammar and intonation (and thus, indirectly, con-text), and hence realize different meanings. A choice from the available alternatives has to be made for each of the phonological categories in order to realize sentence intonation. The three sets of choices according to [14] are:

- **Tonality:** The distribution into tone groups, *i.e.*, the number of tone groups allocated by the speaker to a given stretch of language.
- **Tonicity:** The placing of the tonic syllable, *i.e.*, its position within the tone group.
- **Tone:** The choice of a tone for each tone group; the tone is associated with the tonic.

Choices in the systems of **tonality** and **tonicity** lead to an information constituent structure independent of the grammatical constituency, whereas choices in **tone** result in the assignment of a tone contour for each identified tone group in an utterance. From these systems, only the choices in the tone systems realize an interpersonal function[4], that of indicating a speech function or the speaker's attitude (*e.g.*, [14]). This interpersonal function is our present concern. Next, we want to investigate the tone more closely before turning to the actual system networks in the KOMET grammar.

Following [24], we assume five tones,[5] the *primary tones*, plus a number of so-called *secondary tones* that are necessary for the description of German intonation contours. These tones are: *fall* (tone1), *rise* (tone2), *progredient* (tone3), *fall-rise* (tone4), *rise-fall* (tone5), where the first four can be further differentiated into secondary *a* and *b* tones.[6] The primary tones are the undifferentiated variants, whereas the secondary tones are interpreted as realizing additional meaning. They are intepreted as follows:

1a = neutral	3a = weak contrast
1b = emphatic	3b = strong contrast
2a = neutral	4a = neutral
2b = negative	4b = negative
	5 = assertive/clarifying

Consider the following example taken from one of the information seeking dialogues: The computer has retrieved an answer to a query, and this answer is presented graphically to the user. As a default, the system would generate a neutral statement choosing tone **1a** to accompany the presentation, as in *//**1a** die ergebnisse sind unten dargestellt//*[7] ("The results are given below"). If, however, the results had so far been presented at a different position on the screen, the system would generate tone **1b** in order to place special emphasis on the statement: *// **1b** die ergebnisse sind UNTEN dargestellt//.*

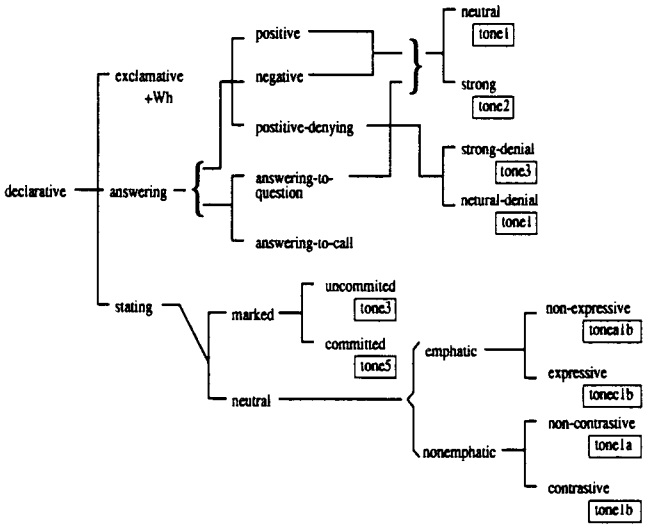


Figure 2: KEY systems in declarative clauses (simplified)

Intonational choices in the KOMET grammar
Modelling intonation in the KOMET grammar involves the introduction of more delicate systems in those areas on the lexicogrammatical level, where intonational distinctions exist, thus specifying the relation between intonation features and competing linguistic resources (like lexis and syntax). Here, we will restrict ourselves to the description of the system networks reflecting the choices in *tone*. The networks are primarily based on the descriptive work by [24].

The interpersonal part of the grammar provides the speaker with resources for interacting with the listener, for exchanging information, goods and services, etc. (see [15, 20]). On the lexicogrammatical stratum, the MOOD systems are the central resource for expressing these speech functions. More delicate speech functional distinctions—specific to spoken German—are realized by means of tone. The (primary) tone selection in a tone group serves to realize a number of speech functional distinctions. For instance, depending on the tone contour selected, the system output *//sie wollen um fünfzehn uhr fahren//* ("You want to leave at 3 pm.") can be either interpreted as a question (tone **2a**) or a statement (tone **1a**).

More important is the conditioning of the (secondary) tone by attitudinal options such as the speaker's atti-

[4] **Tone** moreover realizes the *logical* metafunction, however, we will ignore this fact for the present argument.

[5] Other approaches to intonation suggest a different number of *tones*, ranging from four to six. [8] even goes one step further in arguing that it is not sufficient to describe *tones* by a combination of *fall* and *rise*, instead, much finer distinctions have to be made (see [8]).

[6] The criteria for the distinction of primary tones is the *type* of the tone movement, for instance rising or falling tone contour, whereas the *degree* of the movement, *i.e.*, whether it is strong or weak in expression, is considered to be a variation within a given tone contour.

[7] In this paper, the following notational conventions hold: // marks tone group boundaries, CAPITAL LETTERS are used to mark the tonic element of a tone group. Numbers following the // at the beginning of a tone group indicate the type of tone contour.

tude towards the proposition being expressed (surprise, reservation ...), what answer is being expected, emphasis on the proposition etc., referred to as KEY features. If one defines KEY as the part of speech functional distinctions expressed by means of tone rather than mood alone, one can integrate the MOOD and KEY systems into the grammar by positioning KEY systems as dependent on the various MOOD systems.[8]

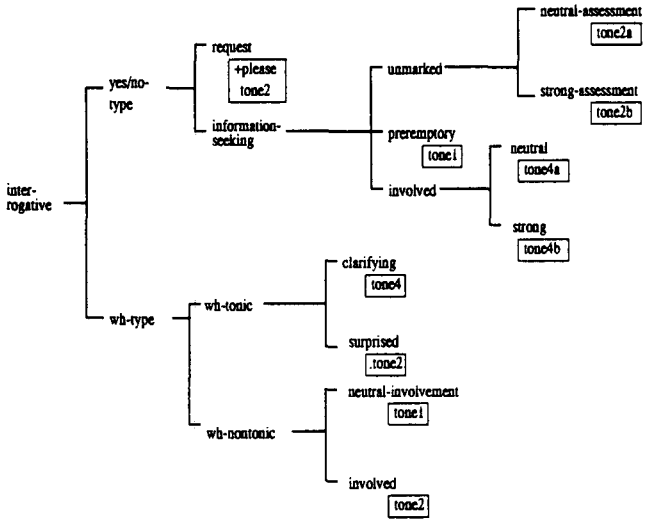


Figure 3: KEY systems, interrog. clauses (simplified)

Figures 2 and 3 give the system networks of the KOMET grammar for the declarative and interrogative sentence mood. The networks now include more delicate grammatical distinctions in order to realize the variations that have intonational consequences. The networks are restricted in that they omit some of the incongruent mood codings. The added discriminations to the KOMET grammar impose constraints on the specification of an appropriate intonation contour.

3.3 Integrating COR and KOMET-PENMAN

As illustrated in Section 3.2 the relation between dialogue moves and tone is many-to-many, hence the appropriate tone selection must be further constrained. The dialogue model provides general information about the structure of an information retrieval dialogue, hence we consider it a representation of *genre*. The KOMET grammar provides *linguistic* resources including intonational options. In the following section, we determine the kinds of information that are needed in addition to what these resources provide and suggest a method of integrating the additional resources in the overall system.

4 Constraints on choice in intonation

In information-seeking, human-machine dialogue it is crucial to signal to the user as unambiguously as possible at which stage in the dialogue she is and what action (verbal or non-verbal) she is supposed to take (see Section 2). When spoken mode is envisaged as output, the intonation contour is the major means to convey this information. The relation between dialogue moves and tone types is however not trivial. For instance, a dialogue move REQUEST—depending on the context in which it occurs—may be realized intonationally by using tone 1, tone 2 or tone 4. Hence, the selection of an appropriate tone is conditioned by factors other than just individual COR dialogue moves. When we think about the problem from the perspective of intonation, the picture becomes clearer. It is generally acknowledged in descriptive linguistics that the kind of tone attributed to an information unit encodes a basic semantic speech act or *speech function* [27, 25], such as command, question, statement and offer, even though this relation is not one-to-one. Also, it is uncontroversial to maintain that intonation potentially reflects a *speaker's attitude* towards the message she verbalizes (see *e.g.*, [24]). When looking at dialogue—rather than monologue—other factors coming into play are the *history of the dialogue* taking place and the *expectations on the part of the hearer* that are evoked at particular stages in the course of the dialogue.

In this section, we will discuss how these factors relate to the selection of tone. Our goal is to determine more precisely what is comprised by them and to arrive at a refinement of the general architecture we have presented in Section 3. More concretely, it will be shown that the factors just pointed out are logically independent parameters that in different combinations constrain the selection of a particular tone. We will then propose an organization of these different parameters in terms of stratification that allows for the necessary flexibility and brigdes the gap between the dialogue model and the generator. Discussing a sample dialogue (Section 4.2), we will then apply the model developed.

We start from the stratum of grammar and move to the other linguistic and pragmatic resources relevant to the present problem. As the starting point for discussion we take the grammatical systems of MOOD and KEY, for they grammatically encode semantic speech function and speaker's attitudes and lead directly to selections in tone.

4.1 The meanings of tone

One of the primary grammatical choices relevant for the selection of tone is the choice of *mood*, such as declarative, interrogative and imperative.[9] The relation between mood and tone is potentially many-to-many with one exception:imperative is always realized by tone 1. However, the choice of mood is crucial since it leads to a whole variety of options that are eventually realized in different tones (these are the KEY systems).

How is choice in the basic mood options constrained?

[8] [14], [7] and [21] have described this for English, [24] adapted Halliday's approach for German.

[9] We assume here that the information unit is the clause and that tonality is unmarked, *i.e.*, that there is one tone-group only. We are aware, however, that generally there is no one-to-one correspondence between information unit and clause.

Mood is in the first instance the grammatical realization of *semantic speech function*. Speech functions comprise command, offer, statement and question. Systemically, they are derived from the SPEECH FUNCTION network (see *e.g.*, [20] and Figure 4). Again, the relation between speech function and mood is potentially many-to-many:All of imperative, declarative and interrogative may for instance encode a command. For example *Schließ das Fenster!* (*Close the window!*), *Würdest Du das Fenster schließen, bitte?* (*Would you close the window, please?*), *Du sollst das Fenster nicht öffnen!* (*You're not supposed to open the window!*).

How can the mapping between speech function and mood be constrained then? A major constraint on the mapping between speech function and mood is the kind of discourse or *genre*, and the type of discourse stage the message is produced in. For instance, the genre of information-seeking, human-machine dialogues is characterized by certain genre-specific stages or dialogue moves (see Section 3.1). A typical move in this genre is the REQUEST move. In terms of speech function, a REQUEST is typically a *question*, *i.e.*, [*demanding:information*].[10] The REQUEST-*question* correlation in the kind of dialogue we are dealing with here constrains the choice of mood to *interrogative* or *declarative*, *e.g.*, (1) *Wohin möchten Sie fahren?* (*Where do you want to go?*) (interrogative)— (2) *Sie wollen um drei Uhr fahren?* (*You want to go at three o'clock?*) (declarative). So, in information-seeking dialogues, the type of move largely constrains the selection of speech function, but it only partially constrains the mapping of speech function and mood.

Deciding between declarative and interrogative as realization of a move REQUEST requires information about the immediate *context* of the utterance, *i.e.*, about the dialogue history. It is in the area of combinations of dialogue moves that we find reflections of speaker's attitudes and intentions and hearer's expectations as determined by the context. The area in the grammar encoding this is *key*.

The KEY systems are subsystems of the basic MOOD options (see Section 3.2). In terms of key, example (1) would be [*interrogative:wh-type:wh-nontonic:neutral-involvement*], thus leading to an intonational realization as tone 1, example (2) would be [*declarative:answering:answer-to-question:strong*] leading to an intonational realization as tone 2. Consider the contexts in which (1) or (2) would be appropriate: (1) would typically be used as an *initiating* move of an exchange, where there is no immediately preceding context—the speaker's attitude is essentially neutral. (2) would typically be used in an exchange as the realization of a *responding to* move; in terms of the COR model, (2) would be a possible realization of a REQUEST within an INFORM or within a REQUEST—the speaker wants to make sure she has understood correctly. Only in the REQUEST or INFORM contexts of a REQUEST does it become possible to map the dialogue move/speech function correlation of REQUEST-*question* to the mood and key features [*declarative:answering:answer-to-question:strong*] (see also Section 4.2).

[10]The notation [*x:y:z*] gives a path through a system network.

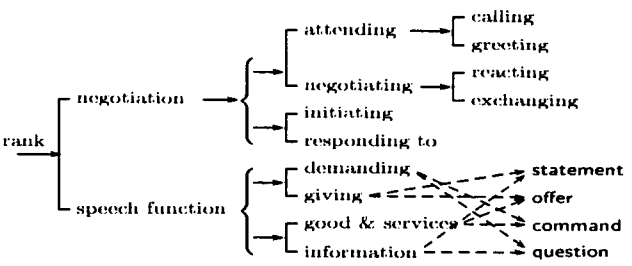


Figure 4: Speech functions—the semantic stratum.

For the representation of constraints between dialogue moves on the dialogue side and speech function on the side of interpersonal semantics and mood and key on the part of the grammar, this means that a good candidate for the ultimate constraint on tone selection is the type of move *in context* (or: the dialogue history).

Given that all of the parameters (dialogue move type, dialogue history, speech function, mood and key) are logically independent and that different combinations of them go together with different selections of tone, an organization of these parameters in terms of stratification suggests itself, for it provides the required flexibility in mapping the different categories. Such an organization is for instance proposed in systemic functional work on interaction and dialogue [3, 20, 36].

In the systemic functional model, the strata assumed are context (extra-linguistic), semantics and grammar (linguistic). On the semantic stratum, general knowledge about interactions is located, described in terms of the NEGOTIATION network (cf. Figure 4). A pass (or passes, since NEGOTIATION is recursive) through the network results in a syntagmatic structure of an interaction called *exchange structure*. An exchange structure consists of *moves* which are the units for which the SPEECH FUNCTION network holds. NEGOTIATION and SPEECH FUNCTION are the two ranks of the stratum of interpersonal semantics (see Figure 4). The MOOD and KEY systems represent the grammatical realization of a move (given that a move is realized as a clause). The ultimate constraint on the selection of features in the interpersonal semantics and grammar is the information located at the stratum of context. This is knowledge about the type of discourse or *genre*. In the present scenario, this contextual knowledge is provided by the dialogue model, reflecting the genre of *information-seeking, human-machine dialogue*. Since the stratum of context is extra-linguistic, locating the dialogue model—which has originally not been designed to be a model of *linguistic* dialogue, but of retrieval dialogue in general—here is a straightforward

step. For a graphical overview of the stratified architecture we just described briefly see Figure 5.

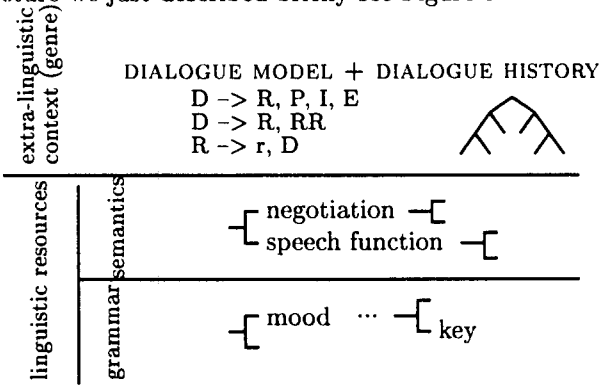


Figure 5: The stratified model.

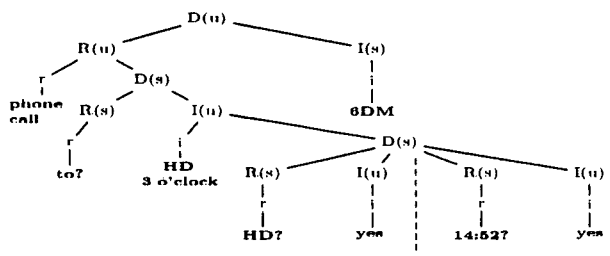
4.2 A top-down perspective

In this section, we discuss our proposal of bridging the gap between the dialogue model and the text generator KOMET-PENMAN from a top-down perspective. We develop concrete mappings between the extra-linguistic and semantic strata. Further, we show how competent choices at the semantic stratum guide the selection of features in the MOOD and KEY systems, which finally result in the assignment of a tone. We base our derivation of mappings between the strata on the following sample dialogue and its COR analysis[11] from the domain of giving out train information. In the following, we will discuss the different system utterances one by one. This discussion is summarized in Table 1 .

A) usr <Calls the train information>
B) syst Wo möchten Sie hin?
C) usr Heidelberg (HD) um 3Uhr.
D) syst Sie wollen nach Heidelberg?
E) usr Ja.
F) syst Wäre 14:52 heute Nachmittag OK?
G) usr Ja.
H) syst Eine einfache Fahrt kostet 6DM.

English translation:

A) usr <Calls the train information>
B) syst Where do you want to travel?
C) usr Heidelberg (HD) at 3 o'clock.
D) syst You said Heidelberg?
E) usr Yes.
F) syst Is 14:52 this afternoon OK?
G) usr Yes.
H) syst A one-way ticket costs 6DM



Initial requests Utterance B) results from a real information need on the system part. In order to do anything at all, the system must know where the user wants to travel. Unless the user volunteers the destination, it must request this information from her.[12] The user did not say where she wanted to travel, hence the system initiated the exchange, this is represented by the following path through the NEGOTIATIONha network: [*negotiation:negotiating:exchanging/initiating*].

In terms of speech function, we realize this request as a question ([*demanding/information*]). Other possible realizations of a request would be command, offer and statement, though none of them applies in the given context. The scenario itself excludes the command and the statement option, since the system is in need for information. Finally, since the system is incapable of handing over, say, a ticket, this request cannot be realized as an offer ("Let me give you a ticket to your destination").

Knowing that we have to realize a question, we have three MOOD options available: [*declarative*], [*yes/no-question*], and [*wh-question*]. Keeping in mind that we want our system to be user friendly, we do not want it to realize this request as a yes/no question ("Do you want to go to Heidelberg?"), or a statement ("You want to travel to Heidelberg?"), since it would then exhaustively have to search through its knowledge base in order to find the right destination to include in its utterance. Hence we conclude that requests that are not in response to a user utterance should be realized as a wh-question.

In terms of KEY, we do not want our system to be overly [*involved*] in the conversation or [*surprised*] by the fact that it has to request some information and there is nothing to [*clarify*], hence the only accessible KEY feature is [*neutral-involvement*], which implies that utterance B)—an initiating, neutral, wh-question—should be realized as tone 1.

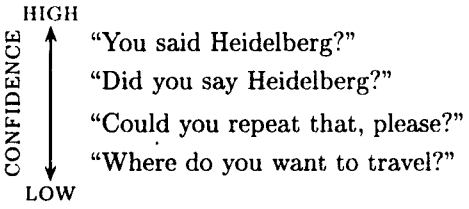
Responding requests Utterance D) is a request in response to the destination that the user informed. In terms of semantic choices it is [*initiating*] a new embedded exchange, while it is [*responding to*] a user move in the embedding exchange. The speech function is question since the system wants to initiate a response.

We suggest that the linguistic realization of this question depends on how confident the system is about what the user informed, hence in order to choose appropriate MOOD and KEY features, we argue that we need access to an additional resource—a confidence measure.[13] For the current example, we suggest the following alternatives:

[11] In the analysis: D=(sub-)Dialogue, R/r=R/request, I/i=I/inform.

[12] We assume that the system has an abstract internal specification of its information needs and that it keeps a record of the information it has already received.

[13] This is highly relevant if the input channel is spoken, since speech recognizers cannot achieve a 100% recognition rate. Technically, the confidence measure would come from the speech recognition unit.



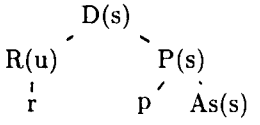
If the system is confident that it has understood what the user said, it would ask only to confirm what it believes to know, hence it would choose a declarative with tone 2 ([*answering:positive/answering-to-question:strong*]). If the confidence is somewhat lower, there are two ways of realizing a yes/no-question: tone 2a ([*interrogative:yes/no-type:information-seeking:unmarked:neutral-assessment*]) or tone 2 ([*interrogative:yes/no:request*]). Finally, if the system has not at all understood what the user said, it could indicate this by using a clarifying wh-question with tone 4 ([*interrogative:wh-type:wh-tonic:clarifying*]).

Utterance F) is also in response to a user inform, but what makes this situation different from the response above is that here, there is a mismatch between what the user wanted and what the system can offer (User wanted 3 o'clock, while system can only offer 14.52). Hence the system must offer the user an alternative and the linguistic form of this utterance might differ with the "closeness" of the alternative to the original demand.

If the alternative is reasonably close (In our example, there is a time difference of 8 minutes, which, for this scenario, might be considered a good alternative), we find it appropriate to generate a yes/no-question with tone 2b ([*interrogative:yes/no-type:information-seeking:unmarked:strong-assessment*]). The lack of good alternatives, however, might condition a wh-question ("What is your next preferred departure time?") with tone 1 ([*interrogative:wh-type:wh-nontonic:neutral-involvement*]).

Inform The system answers the user's question, *i.e.*, it is [*giving/information*], and hence the speech function is statement. Statements of this type do not need any particular intonational marking, since at this point they are expected, hence we choose the features [*declarative:stating:neutral:nonemphatic:non-contrastive*], *i.e.*, tone 1a. *E.g.*, ("Eine einfache Fahrt kostet 6DM" (="The ticket costs 6DM.").

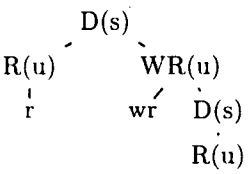
Promise The information knower can utter a promise when she wants to signal the information seeker that she is considering the request. For instance, "Ich DURCHSUCHE die datenbank." (= "I am searching"). A promise move is always in response to a request move and the relevant partial structure is:



The speech function is statement since the system is [*giving/information*], and as MOOD and KEY features we choose [*declarative:stating:neutral:nonemphatic:non-contrastive*], hence tone 1a.

As indicated in the partial structure above, an assert move can follow a promise act. This is additional information that the system volunteers the user, which often take the form of a polite command, *e.g.*, "Bitte warten Sie."(= "Please wait."). Linguistically, commands are realized as imperatives and hence tone 1.

Request in Withdraw A request in the context of any of the unexpected dialogue moves (*e.g.*, withdraw-request) mostly serves as confirmation question similar to the responding requests in inform that we discussed above. This is, however, an unexpected move on the part of the user, hence we suggest that these requests, again mapping to question on the speech functional level, are realized as yes/no-question (as opposed to declarative with tone 2, see above) *i.e.*, "Do you want to quit?" *vs.*"You want to quit?". Which tone one chooses for this type of question depends on how involved one wants the system to appear. Tone 4a indicates neutral involvement, while tone 4b signals strong involvement. The partial structure of this type of requests is as follows:



Offer In an information retrieval system, the system often offers the user a list of alternatives from which she has to choose one. If we consider the appearance of a list on the screen a metaphor for actually handing over an object, this situation corresponds to [*demanding/goods & services*], *i.e.*, the speech function is command, hence we suggest that offers are realized as imperatives with tone 1. *E.g.*, "Bitte wählen Sie eins." (="Please choose one.")

Summary The above discussion is summarized in Table 1 . Further, from the data that we have collected so far we observe:

- The dialogue move guides the selection of speech function, *e.g.*, request corresponds to speech function question, whereas *offer* maps to command.
- The dialogue history, or context, guides the selection of semantic choices, *i.e.*, pure initiating moves (*e.g.*, request) correspond to [*exchanging/initiating*], while responding initiating moves (*e.g.*, inform(request)) correspond to [*exchanging/responding*] in a first grammar traversal and [*exchanging/initiating*] in a second.
- Choices in MOOD and KEY systems can often not be made unless we have access to additional knowledge sources as, for instance, a confidence measure.

Future empirical studies will determine whether these generalizations hold.

	Genre level	Exchange move	Speech function	Mood/ Key	Tone
B)	request	[exch/init]	question	WH	1
D)	inform(request)	[exch/resp] [exch/init]	question	A	2
D)	inform(request)	" "	question	Y/N	2b
D)	inform(request)	" "	question	WH	4
F)	inform(request)	[exch/resp] [exch/init]	question	Y/N	2a
H)	inform	[exch/resp]	statement	S	1a
	promise	[exch/resp]	statement	S	1
	promise(assert)	[exch/resp]	command	I	1
	withdraw(request)	[exch/resp] [exch/init]	question	Y/N	4a/b
	offer	[exch/init]	command	I	1
	reject request	[exch/resp]	statement	A	2

Table 1: Notation: '/' and '//' indicate choices in paralell systems, ':' indicate refinement of the previous choice. WH = wh-question, Y/N = yes/no question, A = answering, S = stating, I = imperative.

5 Conclusions

In this article we have developed a model for guiding the selection of intonation in a system supporting human-machine interaction in retrieval dialogues with spoken output. We have concentrated on the choice of tone as a major signal of interpersonal semantic features, such as speech acts and speaker's attitudes. To express these appropriately is crucial especially in human-machine dialogue, since they contribute to the success of the interaction in a major way.

As a starting point we have taken two existing systems—the COR dialogue model and the KOMET-PENMAN generation system. On this basis, we have determined a number of factors that contribute to the selection of appropriate tones, such as speech function, speaker's attitudes and hearer's expectations, and types of dialogue moves in context. Finally, we have proposed a stratified model that includes all of the relevant kinds of information to guide the selection of tone.

Even though our dialogue model was originally not designed for language, we have shown that this relatively simple model provides useful information for intonation selection. A linguistically based discourse model would be able to provide more information, but in the context of an *interactive* conversational system in which there are practical limits on how long it can take to produce a response, we believe that a full fledged discourse analysis system would be too slow.

We are aware that we have left untouched a number of problems that are involved in the generation of appropriate intonations. These include:

- accounting for the *textual* meaning of intonation encoded in information structure and thematic development/progression (realized in tonicity; see Section 2);
- We handle only situations in which there is a one-to-one corresponds between tone group and clause.
- We can only make predictions about complete clauses, hence the grammar prevents the generation of utterances with ellipses. This is relevant for geographical clarification question, *e.g.*, "Wollen Sie nach Frankfurt am Main oder Frankfurt an der Oder?". In many contexts it is more natural to use just a phrase "Frankfurt am Main oder an der Oder?"

  Similarly it is unnatural to generate the evaluate moves as complete clauses. It suffices to generate simple phrases like "Thanks" or "OK".

Also, the method we have applied here to develop our model has been solely qualitative. For a proper validation we need to analyse larger quantities of dialogue in order to have an empirically sound foundation. The same is true for the classification of intonation which was developed by Pheby in the late sixties. Here, the collaborative work with speech synthesis will provide us with empirical data that can then be used to refine the classification.

REFERENCES

[1] B. Abb, C. Günther, M. Herweg, C. Maienborn, and A. Schopp. Incremental syntactic and phonological encoding - an outline of the synphonics formulator. In *Proceedings of the Fourth European Workshop on Natural Language Generation*, pages 19–29, 1993.

[2] H. Altmann, editor. *Zur Intonation von Modus und Fokus im Deutschen.* Tübingen: Niemeyer, 1989.

[3] M. Berry. Systemic linguistics and discourse analysis: a multi-layered approach to exchange structure. In Malcolm Coulthard and Michael Montgomery, editors, *Studies in Discourse Analysis.* Routledge and Kegan Paul, London, 1981.

[4] M. Bierwisch. Regeln für die Intonation deutscher Sätze. In *Studia Grammatica VII: Untersuchungen über Akzent und Intonation im Deutschen*, pages 99–201. Berlin: Akademie Verlag, 1973.

[5] E. Bilange. A task independent oral dialogue model. In *Proc. of the European Chapter of the ACL*, pages 83–87, 1991.

[6] G. Dorffner, E. Buchberger, and M. Kommenda. Integrating stress and intonation into a concept-to-speech system. In *Proc. of the 14th Intl. Conf. on Computational Linguistics (COLING'90)*, pages 89–94, 1990.

[7] R.P. Fawcett, A. van der Mije, and C. van Wissen. Towards a systemic flowchart model for discourse. In R.P. Fawcett and D. Young, editors, *New Developments in Systemic Linguistics*, volume 2, pages 116–143. Pinter, London, 1988.

[8] C. Fery. *German Intonational Patterns.* Tübingen: Niemeyer, 1993.

[9] M. Fischer, E. Maier, and A. Stein. Generating cooperative system responses in information retrieval dialogues. In *Proceedings of the International Workshop on Natural Language Generation*, pages 207–216, Kennebunkport, Maine, 1994.

[10] D. Frohlich and P. Luff. Applying the technology of conversation to the technology for conversation. In P. Luff, N. Gilbert, and D. Frohlich, editors, *Computers and Conversation*, pages 187–220. Academic Press, 1990.

[11] B. Grote. Grammatical revision of the german prepositional phrase in KOMET. Technical Report, GMD/Institut für integrierte Publikations- und Informationssysteme, Darmstadt, 1994.

[12] B. Grote. Specifications of grammar/semantic extensions for inclusion of intonation within the KOMET grammar of german. COPERNICUS '93 Project No. 10393. Deliverable R2.1.1, 1995.

[13] E. Hagen and A. Stein. Automatic generation of a complex dialogue history. In *Proc. 11th Canadian Conference on Artificial Intelligence (AI96)*, page forthcoming. Canadian Society for Computational Studies of Intelligence, 1996.

[14] M.A.K. Halliday. *Intonation and Grammar in British English*. The Hague: Mouto, 1967.

[15] M.A.K. Halliday. *An Introduction to Functional Grammar*. Edward Arnold, London, 1985.

[16] J.P. Hemert, U. Adriaens-Porzig, and L.M. Adriaens. Speech synthesis in the spicos project. In H.G. Tillmann and G. Willee, editors, *Analyse und Synthese gesprochener Sprache. Jahrestagung der GLDV*, pages 34–39. Hildesheim: Georg Olms, 1987.

[17] J. Hirschberg. Using discourse context to guide pitch accent decisions in synthetic speech. In G. Bailly and C. Benoit, editors, *Talking machines: Theory, Models and Design*, pages 367–376. Amsterdam: North Holland, 1992.

[18] K. Huber, H. Hunker, B. Pfister, T. Russi, and C. Traber. Sprachsynthese ab Text. In H.G. Tillmann and G. Willee, editors, *Analyse und Synthese gesprochener Sprache. Jahrestagung der GLDV, 1987*, pages 26–33. Hildesheim: Georg Olms, 1987.

[19] K.J. Kohler. *Einführung in die Phonetik des Deutschen*. Berlin, 1977.

[20] J.R. Martin. *English text; System and structure*, chapter 7, pages 493–573. John Benjamins Publishing Company, Philadelphia/Amsterdam, 1992.

[21] C. Matthiessen. Lexicogrammatical cartography: English systems. Technical Report, University of Sydney, Linguistics Department, 1993.

[22] M. O'Donnell. A dynamic model of exchange. *Word*, 41(3):293–327, 1990.

[23] G. Olaszy, G. Gordos, and G. Nemeth. The multivox multilingual text-to-speech converter. In G. Bailly and C. Benoit, editors, *Talking machines: Theory, Models and Design*, pages 385–411. Amsterdam: North Holland, 1992.

[24] J. Pheby. *Intonation und Grammatik im Deutschen*. Akademie-Verlag, Berlin, 1969. (2nd. edition, 1980).

[25] J. Pheby. Intonation. In K.E. Heidolph, W. Flämig, and W. Motsch, editors, *Grundzüge einer deutschen Grammatik*, pages 839 – 897. Akademie-Verlag, Berlin, 1980.

[26] S. Prevost and M. Steedman. Specifying intonation from context for speech synthesis. *Speech Communication*, to appear.

[27] R. Quirk, S. Greenbaum, G. Leech, and J. Svartik. *A comprehensive grammar of the English language*. Longman, London, 1985.

[28] N. Reithinger, E. Maier, and J. Alexandersson. Treatment of incomplete dialogues in a speech-to-speech translation system. In *Proc. ESCA Workshop on Spoken Dialogue Systems; Theories and Applications*, pages 33–36. ESCA and Center for PersonKommunikation, Aalborg University, Denmark, 1995.

[29] J.R. Searle. *A Taxonomy of Illocutionary Acts. In: Searle, J.R. Expression and Meaning. Studies in the Theory of Speech Acts.*, pages 1–29. Cambridge University Press, Cambridge, MA, 1979.

[30] M. Selting. Phonologie der Intonation: Probleme bisheriger Modelle und Konsequenzen. *Zeitschrift für Sprachwissenschaft*, 11(1):1993, 99–138.

[31] S. Sitter and A. Stein. Modelling the illocutionary aspects of information-seeking dialogues. *Information Processing and Management*, 8(2):165–180, 1992.

[32] E. Teich. Komet: Grammar documentation. Technical Report, GMD/Institut für integrierte Publikations- und Informationssysteme, Darmstadt, 1992.

[33] E. Teich, J.A. Bateman, and L. Degand. Multilingual textuality: Experiences from multilingual text generation. In Zock M. and G. Adorni, editors, *Selected Papers from the Fourth European Workshop on Natural Language Generation, Pisa, Italy, 28-30 April 1993*. Springer, Berlin, New York, forthcoming.

[34] E. Teich, A. Stein, E. Hagen, and J.A. Bateman. Meta-dialogues implementation. Technical report, GMD/IPSI, Technical Universities of Darmstadt and Budapest, 1995. Speech Generation in Multimodal Information Systems, Copernicus Project No. 10393, SPEAK! deliverable P3.2.3.

[35] D. Traum and E. Hinkelman. Conversation acts in task-oriented spoken dialogue. *Computational Intelligence*, 8(3):575-599, 192.

[36] E. Ventola. *The Structure of Social Interaction: A Systemic Approach to the Semiotics of Service Encounters*. Frances Pinter (publishers), London, 1987.