ISCA Archive
http://www.isca-speech.org/archive

ISCA & IEEE Workshop on Spontaneous
Speech Processing and Recognition
Tokyo Institute of Technology, Tokyo, Japan
April 13–16, 2003

NATURALNESS IN SPEECH COMMUNICATIONS

Paul C. Lustgarten and B.H. Juang***
*Avaya Labs Research, Basking Ridge, New Jersey 07920, USA
**Avaya Labs Research and Georgia Institute of Technology, Atlanta, GA 30322, USA

ABSTRACT

Speech has long been considered the most *natural* form of human communications. People with normal speaking and hearing abilities use speech to exchange information every day, often effortlessly. An interesting question, however, remains rarely addressed: What constitutes naturalness in speech communications and how is naturalness achieved? In this paper, we attempt to analyze the “human behavioral components” that contribute to the naturalness or perceived naturalness in human speech communications. We further attempt to mechanize one of these components, namely use of “reference”, in an intelligent service scenario involving spoken language, to demonstrate that indeed such a component greatly adds to the natural experience in human-machine dialog. We hope to inspire further research into other naturalness dimensions that may enrich human machine interaction.

1. INTRODUCTION

Speech has long been considered the most natural form of human communications. People with normal speaking and hearing abilities use speech to exchange information everyday, often effortlessly. This is very significant and remains one of the focal points of telecommunications. During the evolution of telecommunications in the past century, telephony became a preferred method over telegraph because of its natural ease of use. It is also this natural, effortless communications ability that has inspired many speech researchers to pursue knowledge and ideas that would make a machine capable of communicating with a human like a human.

Nevertheless, not every form of speech communications is considered natural. Most of the so-called speech-enabled communication services such as an IVR (Interactive Voice Response) system perform strict conversion of spoken commands into corresponding actions. Giving an order (e.g., “call 5-8-2-2-0-0-2” as in voice dialing) is a one-way command mode that is not, strictly speaking, natural interaction. In other words, current spoken language technologies, which focus on conversion between sounds and words, have not necessarily fulfilled the original vision of natural speech communication.

Several interesting questions are rarely addressed: What constitutes naturalness in speech communications? How is naturalness encapsulated in linguistic expressions? And more generally, how is naturalness achieved in human communications, which may involve multiple modalities? In the interest of speech system developments, how can we design a machine to perform like a human in spoken dialog? These questions need to be answered for us to be able to achieve further progress toward having a truly natural human machine interaction.

In this paper, we attempt to analyze the *human behavioral components* that contribute to the naturalness or perceived naturalness in human speech communications. We further attempt to mechanize one of these components, namely use of “reference”, in an intelligent service scenario involving spoken language, to demonstrate that indeed such a component greatly adds to the natural experience in human-machine dialog.

2. A MAP OF COMMUNICATIONS

In order to stipulate on the notion of naturalness in speech communications, a review of the linguistic structure and the associated physical and cognitive behaviors is helpful. Speech is a vehicle for delivering ideas. Traditionally, realization of the speech phenomenon is embedded in layers of expressions, from the physical sound generated by our articulatory apparatus to pragmatic attributes that are signified by the prosody, such as stress, or the speaking environment. The left side of Figure 1 illustrates this traditional taxonomy of speech phenomena. Above acoustics, a speech signal encompasses and conforms to customary rules in phonology and morphology that underlie a particular language. The speaker also has to observe the syntactic rules of the language in order to be understood. Beyond syntax, the expression needs to be interpreted in a conventional semantic framework, in which many other elements of the communication are brought in to allow the listener to derive some tentative cognitive decisions.

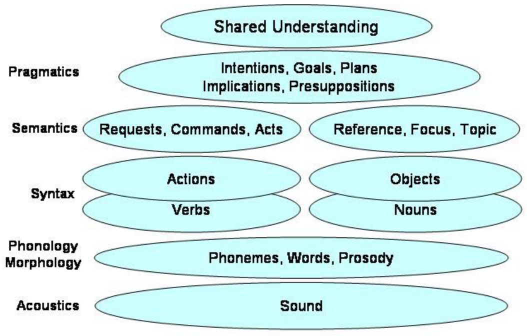

Figure 1: A map of communications

We project the analysis of naturalness in speech communications in a behavioral perspective in terms of the need towards shared understanding of the original notion that the speaker intends to convey. Specifically, as an example, while syntactic rules deal with treatment of nouns and verbs, it is the objects and the actions that the speaker really would like to deliver to the listener (see Fig. 1). Similarly, beyond the purely semantic level, the speaker's intention often focuses on things like "request", "command", and "acts", etc. The semantic and pragmatic framework is not complete without supporting circumstantial elements such as reference of expression, focus of conversation, topic of discussion, and so on. As illustrated in the figure, these are indispensable elements for a shared understanding between the communicants in a dialogue.

3. NATURALNESS IN COMMUNICATIONS

Speech communication is a process that involves strategy. We say someone's speech is coherent because it has a strategy to deliver the intended information consistently. Opposite to coherent speech is incoherent speech, which is hard to comprehend and often appears to be leading the listener(s) in random directions. A speech strategy is an overall schema, which comprises the intention, the goal, the plan, and the implication, that the speaker executes to deliver the information. During a discourse, the talker would normally possess a certain projection and anticipation, meaning s/he has some prior expectation as to what may come out of the other party in the next turn(s) and also prepares for it (or them).

Natural execution of the speech schema, in a broad sense, means one with *the least effort*. As mentioned above, speech is a structured expression based on commonly accepted rules in phonology, morphology, and syntax and grammar. Conformance to these rules would undoubtedly incur effort. A speech with less linguistic effort in terms of (casual) pronunciation, (fuzzy and imprecise) use of words, word grouping, completeness in sentential structure, or other usual disfluencies would consume less effort and appear to be more natural – after all, *to err is human*. Similarly, speech resulting from a relaxed speaking style in pace regulation, breath grouping, and use of expressions such as a sigh or "hmm" would also feel more natural – that is, *to rest is human*.

Another important element that contributes to naturalness in speech communications is the awareness of the context and references that surround the conversation or dialog without the need to elaborate. This includes implicit awareness of the subject domain and related topics and even the purpose of the interaction or transaction. If every subject matter referred to in the conversation needs a complete definition every time it is mentioned, the speech interaction would be unwieldy, unproductive and unnatural. The key is to successfully refer to items as needed to achieve shared understanding, without elaboration or with a minimum need to elaborate. But, how is the implicit awareness expressed syntactically and semantically? This is the question we attempt to explore in this paper.

We contend that *reference* in all its forms is a central capability of a natural spoken dialog system. For developers of automatic speech systems, the following related questions are worth addressing: Will explicit use of referential semantics (to be defined later) make human-machine interactions look and feel more natural? How do we include reference in dialog and grammar design to achieve the perceived naturalness? Answers to these questions may point to a more inviting automatic spoken dialog system that a user would find more pleasant to work with than is available today.

4. CREATING REFERENTIAL NATURALNESS

To explore the issue of communication naturalness, we adopted the task of a virtual assistant, called Daisy, which supports access and manipulation of a personal calendar, via a referentially rich, conversational style, constrained natural language spoken dialog. Furthermore, it encompasses discourse-level resolution of referential acts (as opposed to word or phrase level disambiguation). Resolution through dialog is most interesting in human machine interaction as it seeks consistency between the contextual data and the dialog in progress. The task is also inspiring as it allows creation of new referents (e.g., new appointments). To imbue a dialog system with referential naturalness is not, of course, as simple as adding a self-contained "referential naturalness" component. Rather, it is a property that suffuses multiple elements of the dialog system, in a coordinated and inter-related pattern. From our work to date, we've been able to categorize the requisite capabilities under three broad headings:

- Dialog design
- Dialog processing
- Utterance comprehension

The demands that referential naturalness places on each of these areas are described in the following sections, respectively.

5. DIALOG DESIGN

The starting point for achieving referential naturalness is to explicitly design the referential dimensions of the overall dialog. Simply put, this is to answer the questions of *what*, *how*, and *when* reference can occur in the dialog. We conceptualize that the question of **what** can be regarded as a process of specifying a set of *referential fields*, **how** the members of those fields can be referenced as an enumeration of a range of linguistic forms, or *modes*, of reference applicable to each of those fields, and **when** in the dialog these potential referential acts may occur as an aspect of the *modal structure* of the dialog.

The overarching goal in this design process is to be as inclusive and encompassing as possible on each dimension. This applies foremost to what the dialog system can **accept** from the human participant. From there, by a combination of Gricean principles of cooperative communication [Grice, 1975], lexical and syntactic entrainment, and referential grounding [Clark; Grosz], it follows that the language **produced** by the dialog system must cover (nearly) the same range. [Note that one very immediate implication of this principle is that any system whose output is confined to pre-recorded (stored) "prompts" has thereby precluded achieving a high degree of referential naturalness. We may also name such an ensemble of principles the "**protocol of reciprocity**."]

5.1. *What* - Referential Fields & Objects

A referential field is a set of possible referents that all fill the same pragmatic role in the dialog. For example, in the scheduling domain, one prominent referential field has to do with dates. To completely specify the referential field requires identifying the exact range of dates to be allowed—e.g., to allow any day within a specific calendar year (and only those days) would be one field; to include all dates within the following five-year span would make it a different (larger) field.

A dialog system, even one covering only a specific task domain, can draw from a number of referential fields. Consider the following types of field:

- Pure semantic objects (concepts)
  - *three o'clock* (from a time field)
  - *April 14th, 2003* (from a date field)
- Contingent objects (personal or communal)
  - *my three o'clock appointment tomorrow* (from a personal field of events)
  - *the all-employee broadcast on February 9th* (from a communal field of events)
- Derivative objects (attributes)
  - *the duration of that appointment* (derivative from a field of events)
  - … or *start time* or *date* or *end time* or *location* or *topic* or *participants* …
- Dynamic states and the transitions among them
  - *What did you just do?*
  - *Change it back!*
- Co-communicant's dialog model and state
  - *What are you doing?*
  - *What did you hear me say?*

As reflected in some of the examples above, a reference to a member of one referential field may be cast in terms of, or composed from, references to members of one or more other fields. E.g., a particular date might be identified through reference to a particular month, day within that month, and year (*April 14th, 2003*) .. or through reference to a month, a day of week, a recurrence count, and a year (*the second Monday of next April*), or even through reference to a particular holiday (*this Halloween*). The variety of referential forms available is itself an important aspect of designing for referential richness, to which we now turn.

5.2. *How* – through Modes of Reference

Before a system can be designed to embrace the rich use of referential semantics, it is necessary to discuss some modes of reference that often appear in our daily conversation. The first kind is a **proper name**, which presents direct reference of terms in a global sense. For example, "April fifteenth" can be regarded as a proper name for a particular day of the year; even more so, "Halloween". The second mode of reference is a **descriptive noun phrase**, which provides indirect reference of terms, often through the composition of multiple subordinate references, e.g., "the first Wednesday in April". The third mode is **deixis**, which is essentially a context-dependent reference. Examples of deixis include "next Wednesday", "tomorrow", and "the day before yesterday"—all phrases the exact referent of which depends on the day on which they are uttered—or "the next day", which identifies a day relative to another already identified. The last mode of reference is the general class of **anaphora** and **pronouns**. Terms or phrases like "it", "that Wednesday", "the day after that appointment", and sometimes a simple contextual omission in the sentence (i.e., zero anaphora), are in this class or mode of reference.

These various modes of reference identify a range of linguistic forms that a referential term or phrase may take. Having identified the referential fields that are to be incorporated in a dialog system, the dialog designer must consider and specify **how** the human communicant may express the members of each field. This can be done by comprehensively identifying the modes of reference applicable to each field, noting that this is likely to be non-uniform across the members of a given referential field. For example, some days are named holidays; most are not. Furthermore, the linguistic details of a given mode also vary with the field.

5.3. *When* – in Structured Dialog

After identifying the range of **what** can be referred to in a dialog, and **how** to do so in terms of modes of reference, the next question is **when**, within the dialog structure, those referential acts may occur. At one extreme, it is often thought desirable to impose no constraints—any entity may be referred to at any point in the dialog. Even if this were in fact an appropriate flexibility for the human communicant (which we doubt), the technical limitations of contemporary ASR accuracy make it reasonable to consider only a limited domain accommodation. Therefore, we believe it is necessary to structure the dialog, perhaps even aggressively so. This structure not only constrains the confusability, in order to achieve a useful level of accuracy by the automatic speech recognizer, but also provides information crucial to the successful (and simple) resolution of wide-ranging referential terms (discussed further below).

This structural constraint is provided on two levels: within utterance and across utterance. Within the individual utterances allowed by the dialog design, the constraining mechanism is simply that of using a grammar-based language model (vs. a statistical, or N-gram, model), expressed in a form **capable of accommodating the planned mode of reference**. Across utterances, the mechanism is to use a modal discourse structure, in which the range of utterances available on a given turn depends on **the state of the dialog**. The notion of integrating the sentential grammar (which is used by the automatic speech recognizer) and the dialog state (which is conventionally part of the dialog manager as an independent object in the dialog process) for easy *retention and retrieval of the referential fields* is one of the main contributions of this work.

The constraints provided by these two techniques are often complementary, making such integration rather natural. For example, one of the calendar management directives accepted by Daisy is of the form "Reschedule <*event_spec*> for <*date_or_time*>." Alternatively, the same objective may be fulfilled more incrementally:

H:	Reschedule <*event_spec*>.
D:	Reschedule it for when?
H:	<*date_or_time*>

In both cases, the challenge for Daisy in accepting the <*date_or_time*> construct is constrained: in the single-sentence version, by the requirement that it be embedded appropriately in a larger valid sentential form, whereas in the incremental sequence, it's constrained by the sharply-narrowed language model in effect while listening for the human's response to Daisy's "when?" question.

The art here, of course, is in the design of the requisite grammars and the modal structure of the dialog—worthy questions in their own right, but generally beyond the scope of this paper. Modality, as we use the term, can occur across a range of granularities, from a very coarse grain of separating one overall topical context from another (e.g., separating the dialogs for calendar management, call management, and message management), to the medium grain of a sub-dialog for a specific task within one such context (e.g., creating a new appointment in your calendar), to the fine grain of individual turns for collecting the arguments to a relatively simple command (such as the rescheduling example shown just above). While the grammatical separations afforded by such modality can significantly improve the recognizer's language model (e.g., in reducing confusability), it can also introduce navigational confusion or complexity for the human. Fortuitously, this usability challenge can itself be mitigated by the recursive use of referential richness in regard to the state of the dialog itself, to the degree that a human talker would not normally feel the effort. For example, Daisy provides a uniform ability, across all modes and irrespective of the local topical focus, to inquire into the current state ("What are you doing?", "What are we talking about?"), to un-do the most recent action ("No, that's wrong!"), or to back up to the (implicitly specified) most recent higher-level state in the dialog ("Back up!"). These techniques provide a "dialog stress release" for the user to re-synchronize the dialog state with the machine.

6. DIALOG PROCESSING

Having infused the dialog design with the referential dimensions described above, there are corresponding requirements on the dialog processing. The overarching principle here is actually quite simple, that being to explicitly manage the referential elements of the dialog as it unfolds so as to provide overall referential coherence and continuity. Some of the key elements for doing so are tracking the references that occur, resolving them appropriately, inter-relating those referential acts with the varying state of both the application data at issue and the dialog itself, and aligning references generated by the system with all of this.

6.1. Discourse Referents

A referentially rich dialog is one in which a variety of referential acts are possible, and in which those acts occur in a "target rich" environment—those acts might target, or (be intended to) designate, many different possible *referents*. Furthermore, at the dialog level (and sometimes even within a single utterance), a critical characteristic of the progression of those acts is the relationship among their respective referents. Most fundamentally, that relationship might be one of identity—two distinct referential acts might refer to the same object, or might have the same referent, in which case the participating referential terms (words or phrases) are said to *co-refer*, or to *be co-referential*. For example, "December 18th", "this Tuesday", "tomorrow" and "that day" might, on a given occasion, all be co-referential. Such co-referential terms might be autonomously interpretable, while, in many instances, the referent of one term instead depends, almost parasitically, on some other term with which it co-refers, as is the case with most pronouns, or, more generally (and by definition), with any instance of anaphora. In either case, conversational fluency demands that this co-referentiality be discerned and employed. It is thus necessary for the system to maintain an internal representation, or model, of the referents previously invoked in the dialog; these representations are called *discourse referents*. [Karttunen, 1976]

So the dialog system must work with discourse referents. This includes identifying each one as it is introduced during the dialog and adding appropriate information about it to the information being maintained in the processing of that dialog. This also requires assessing each referential act, as the dialog progresses turn by turn, against the collection of discourse referents accumulated thus far, in order to determine whether a given referential term is in fact introducing a new discourse referent or is co-referential with an existing one. Even when the term being processed is not directly co-referential with any of the established discourse referents, the meaning of the current term may still depend on a prior discourse referent, such as when referring to some specific attribute of an entity already referred to. For example, when discussing a particular event with Daisy, the human may ask "What's its duration?" Resolving the referent of the descriptive noun phrase "its duration" in this utterance requires first identifying the existing discourse referent invoked with the possessive pronoun "its" (i.e., which event is meant), and then composing a new discourse referent for the duration of that event (which is also reported as the answer to the query). Another important aspect of appropriately tracking and using discourse referents derives from the fact that they are introduced by **either** speaker, and not just the human (recall the "protocol of reciprocity")—so anything referred to by Daisy generates a discourse referent in the human's model of the dialog, and Daisy must be prepared for the human to subsequently appeal to that discourse referent in some subsequent utterance, perhaps in the shorthand format of an anaphoric expression.

6.2. Dialog and Referent States

The processing associated with discourse referents also extends into a broader range of issues, which have to do, on the one hand, with the states of the entities identified by those discourse referents, and, on the other hand, with the state of the encompassing dialog in regard to such matters as topic and focus. To the first of these points, note that while the entity designated by a discourse referent is in some cases quite durable (such as the stable semantic concept of the time of day "three o'clock"), others are more mutable—for example, my one-hour meeting starting at three o'clock might be revised to be a two-hour meeting starting at two o'clock. Using a system like Daisy to make such changes to one's calendar clearly requires that the system modify the calendar data itself (which is perhaps maintained in some relatively discrete calendar system office application). Less obviously, what it also requires is appropriate changes to the discourse referents modeling those entries in the calendar (or direct linkage between the discourse referent and the underlying calendar entry), since the modified attributes (here, start time and duration) might subsequently be used in a descriptive noun phrase to refer to that event (e.g., "What room is my two o'clock meeting in?"). Stated conversely, the discourse referent must not be left in its prior state of identifying a one-hour meeting starting at three when there is no longer such a meeting in the calendar. It is, however, quite useful to retain that prior state as such—as **prior** state—in order to allow such things as rescission of changes (ala an "undo" command) or explanation/review of recently enacted steps in the dialog.

The state of the encompassing dialog must also be tracked and employed in resolving certain references. In some cases, this is a relatively simple matter of focus, of holding one member of a referential field as the "current" one, such as a particular day or a particular event, which is then used to resolve subsequent references. For example, consider the following dialog segment:

H: What am I doing next Tuesday?
D: On Tuesday, the seventeenth, you have three appointments.
H: What time is the first one?

Having established next Tuesday as the day in focus, the human can felicitously refer to the first appointment **on that day** simply as "the first one". Continuing this exchange:

D: Your first appointment goes from eight-thirty to ten.

H: Where is it?

With that appointment now in focus as well (and noting that, in the current system, only appointments have locations), that first appointment is readily identifiable as the referent of "it".

In these simple cases, this focus information is used to identify a particular member of a known referential field—in the first case, **which** day is the day whose first appointment is being questioned, and in the second case, which appointment is the appointment whose location is being requested. In some cases, the focus can even determine the **type** of the referent of a particularly versatile referring expression, as shown by the contrast between the referent of "it" in the following two segments:

D: What shall I do? *[at the opening of a dialog]*
H: What day is it?
D: Today is Wednesday, the eleventh.
vs.
D: Your first appointment goes from eight-thirty to ten. *[from the segment just above]*
H: What day is it?
D: The appointment is on Tuesday, the seventeenth.

In the first of these, the referent of "it" is a temporal object—perhaps "now", or "today". In the second segment, the referent of "it" is an appointment!

6.3. Referentially Sensitive Language Generation

While the dominant challenges in creating a referentially rich dialog system are centered in its language understanding capabilities, there is a corresponding dimension of referential sensitivity required in the language **generated** by the system as well. These requirements parallel those sketched in the preceding section in being centered around the processing of discourse referents and associated state of the dialog. However, whereas the overriding referential challenge in the comprehension process is reference resolution, there is a twofold referential challenge in language generation: first, providing adequate but unobtrusive **feedback** about the comprehension of referential acts committed by the human, and, second, formulating appropriate referential expressions as the system commits such acts of its own.

While providing ongoing feedback to the human is necessary for all aspects of the system's comprehension, the pursuit of referential richness entails a correspondingly heightened concern for the negotiation and assurance of referential comprehension. (We must admit that we have not been able to thoroughly implement the protocol of reciprocity, yet.) This concern reflects both a recognition of the dominant role that referential processes play in the overall comprehension process (crudely put, most utterances contain far more nouns than verbs), as well as a recognition that arbitration of shared referential understanding (are we referring to the same entities?) organically permeates the dialog, becoming the subject of explicit confirmation only in the more extreme cases of misunderstanding or repair.

Daisy employs several techniques in reflecting her understanding of key referential terms and (thereby) ensuring the alignment of that understanding with the intent of the human speaker/user. These include a number of principles from "active listening" [Gordon, 2000], such as the practice of echoing key elements of the human's utterance incorporated within her response. Furthermore, these echoes are often embellished with a small amount of additional information (e.g., when responding to an utterance that refers to "next Tuesday", using the phrase "Tuesday, the seventeenth"). Also, as part of managing the cognitive load on the human, Daisy tries to structure her sentences so that those echoed (confirmatory) referential terms occur prior to the new information that she is providing by way of response [Clark, 1977]. Finally, she also tries to adapt her terminology to that used by the human, such as in selecting a term from a set of synonyms.

Employing the human's phrasing is particularly significant in certain cases, as illustrated by the following example. If the human asks to reschedule an appointment from its current date of "tomorrow" to a date specified with some other verbal expression (say, "the third Wednesday in December", or "the eighteenth of December") that happens to be co-referential with that current date, the most cogent explanation of the unexpected co-referentiality requires using the two different referring expressions. In this case, Daisy would say something to the effect of "Excuse me, the third Wednesday in December is the same day as tomorrow, which is when the appointment was already scheduled."

When formulating referring expressions of its own, the system needs to take account of the discourse referents currently active, as well as other contextual factors such as focus. One of the goals of such expressions is that they are resolvable by the other party, ideally with little margin for confusion. However, another goal, typically with opposing implications, is that known information not be re-stated, since, at the least, it fatigues the listener, and beyond that risks actively engendering confusion by implying that the information is not in fact already known (the information wouldn't be presented unless it were relevant [Grice, 1975]). Thus, in expressing a date in a context in which the month is already well established and salient, the generated expression is appropriately limited to something like "Monday, the fifteenth", rather than one including month or even year. Adherence to this "minimality" condition becomes even more imperative when several similar referring expressions follow one another in close succession, such as listing several dates, all falling within the same month, in a single sentence.

7. UTTERANCE COMPREHENSION

With referential richness designed into the dialog, how are specific referential expressions identified and resolved within an individual utterance? We find it useful to organize the speech recognition, language understanding and other per-turn processing around a unified language model to accomplish "comprehension".

This language model is expressed as a finite state grammar, as is normally done. The rules of this grammar are semantically, or even pragmatically, oriented. So, for example, in Daisy's current grammar, there are rules for a time of day, a date, and (at a higher level) a request to reschedule a meeting to a new time and/or date. For purposes of coherence and maintainability, the rules are written as modularly as possible, with rules for more complex constructs being composed from rules for simpler constituents. The grammar is also carefully and tightly designed so that it includes as few improper English sentences as possible—with almost no exceptions, if a sequence of words is accepted by the grammar, it is a valid English sentence (or phrase) acceptable to both the application and its human users. Also, in the spirit of naturalness undergirding this work, all phrases accepted by the grammar are assessed against a criterion of human plausibility—if it would be unnatural for a human doing the same job, or discussing the same topic, to say a given phrase, that phrase is excluded from the grammar.

In building the language model, we take into account the fact that the fundamental element or token of the language model is a word. Tokens can be tagged; some carry "value" tags to denote their referential significance, such as being a member of a set (Sunday is 0, Monday is 1, etc.). A referring expression, which enunciates a referential field, is assumed to consist of a contiguous sequence of tokens.

During operation, the ASR system returns a full parse tree. Currently, this is realized by tagging each token with one or more tags indicating the production in which they appear. These tags become the non-leaf nodes of the tree, with the recognized words being the leaf nodes of the tree. Looking at the tags on each token from right to left, we can envision that the tags are being peeled off the tokens, with each maximal sequence of tokens that share the same rightmost tag being clustered together under that tag. (See Figure 2.)

Since we are dealing with limited domain comprehension at this moment, the language model is rather tightly integrated with the application, namely a calendar assistant, and the intended result of comprehension is simply a set of parameters and actions that define some operation on an engagement or appointment. Therefore, the flow of control, which is modulated by the need to acquire this set of parameters, is defined at the application level, in terms of the sequence of semantic states. Some states flow unconditionally into the next. Otherwise the branching from a given state is determined by a small production system, where the trigger of each production is a possible node in the parse tree, and the action of each production is simply the name of a state to go to.

Under certain limitations, multiple productions can be triggered on a single utterance, which causes the target states of those additional productions to be pushed on a stack. Correspondingly, the transition from a state can be specified indirectly, through popping the top-most state from this stack. Thus, while a cluster of application states with only simple transitions among themselves constitutes a finite-state automaton (FSA), this stack capability provides the effect of being able to push an entire FSA. An entire FSA thereby becomes the unit of programming modularity and re-use—a capability used pervasively throughout Daisy's code. Abstractly, this elevates the computational power of the aggregate application processing from that of a simple FSA to that of a full Turing machine, without departing from the elegance and dialog-compatibility of the state-transition control model.

To prepare for a generalized semantic processing, a collection of re-usable semantic analysis modules has been implemented. Each module receives a subtree of the current utterance parse tree (and possibly some context). The structure of a given type of subtree is determined by the grammar. Each semantic processing module returns some appropriate value(s), according to its semantics; the data type returned by each module is determined independently in designing that module, unconstrained by the design of other modules.

Finally, for system response, language generation is performed by and in individual states. It can incorporate the leaves of the particular subtree(s). In general, a generated utterance is composed programmatically by the application state. Only the simplest ones are fixed and invariant strings.

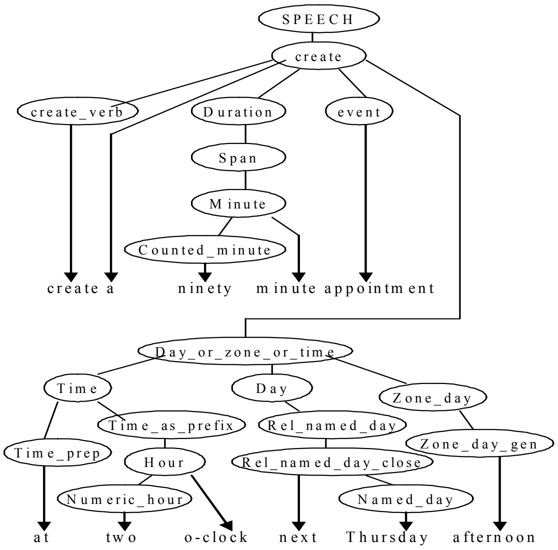

Figure 2: Parse tree returned by ASR

References

1. Clark, H. and Haviland, S. Comprehension and the given-new contract. In Freedle, R., ed. Discourse production and comprehension. Erlbaum, Hillsdale, NJ, pp. 1-40, 1977.
2. Dennett, Daniel C. The Intentional Stance. The MIT Press, Cambridge, MA, 1987.
3. Gordon, Thomas. P.E.T.: Parent Effectiveness Training: The Proven Program for Raising Responsible Children. Three Rivers Press, New York, 2000 (30th Edition).
4. Grice, H. P. Logic and Conversation. In Cole, P., Morgan, J.L., eds.: Syntax and Semantics: Vol. 3: Speech Acts. Academic Press, San Diego, CA, 1975.
5. Karttunen, Lauri. Discourse Referents. In J. McCawley, ed.: Syntax and Semantics 7: Notes from the Linguistic Underground. Academic Press, New York, NY, pp. 363-385, 1976.