Chapter 1 and Chapter 10 excerpted from
Voice User Interface Design

Michael H. Cohen
James P. Giangola
Jennifer Balogh


Addison-Wesley
Boston • San Francisco • New York • Toronto • Montreal
London • Munich • Paris • Madrid
Capetown • Sydney • Tokyo • Singapore • Mexico City
Chapter 1
Introduction to Voice User Interfaces

The following is a true account of the first experience of one of the authors (MC) deploying a commercial spoken language system:

The first application we ever deployed at Nuance was a voice-driven stock quote system for Charles Schwab & Company. The months leading up to the deployment were very exciting for me personally. I had spent more than ten years in the research labs at SRI developing spoken language technology. Finally, we had started a company to commercialize our technology. We had signed our first customer and were almost ready to deploy our first system. We were about to prove to the world that this technology had real value – that it was ready for "prime time."

I observed the early usability tests. I was behind the one-way mirror watching subjects come into the test room, watching them pick up the phone and use the system to get stock quotes and set up stock lists, and watching as the experimenter questioned them about their experience.

The first subject arrived. He was an 83-year-old man from San Francisco. He came in the door, hesitated, and said, "Oh no, I forgot my hearing aid!" Five seconds into the first real user test, and we were already hitting problems we never anticipated in ten years of research!!

The test proceeded. The subject had a great deal of trouble using the system. He had trouble hearing the prompts, was confused about what he could say, and got lost in the application. By the end of the experiment, he had never succeeded in setting up a stock list and never received a quote on a company he requested.

At that point, the experimenter asked the subject about his reactions to the application and whether he thought he would use such a system. He loved it!! He described the experience as the first time in years he had been able to have a conversation with anyone so patient, so willing to repeat themselves, and so willing to talk with him and never get frustrated. He promised he would be calling this system every day!

This story illustrates the two key themes of this book: 1) understanding basic human capabilities is key to the design of effective user interfaces; and 2) understanding the user’s needs and goals, in the context of the business goals, is key to the design of successful applications.
The two themes differ in that the first addresses general human capabilities: understanding what is easy and what is hard for all (or most) people, and how to exploit this knowledge to optimize design choices. (Here we will be concerned with human cognitive capabilities and linguistic behavior, rather than individual physical capabilities such as hearing, as suggested by the story above.) In contrast, the second theme specifically addresses the application at hand: what you need to understand about the intended users, the task, and the business goals in order to focus the design on meeting user and business needs simultaneously. In the story, we could have considered the system test a success if merely providing an electronic conversational partner were an acceptable result. The test result did not, however, meet the business goal of providing value with respect to the handling of one’s brokerage account.

In his book, *The Humane Interface* (2000), Jef Raskin draws a distinction between ‘User-Centered Design’ and ‘Human-Centered Design’. He describes ‘User-Centered Design’ as a design process focused on studying the task-related needs of the intended users of a specific application. However, he goes on to emphasize the importance of ‘Human-Centered Design’, which is focused on “making sure that the interface design accords with universal psychological facts.” His argument is that it is more important to understand the cognitive capabilities and limitations of humans in general and to apply that understanding to interface design problems.

In this book, we argue that both ‘human-centered’ and ‘user-centered’ designs are key to the creation of successful user interfaces. We will spend substantial time on core principles of design based on an understanding of human cognitive capabilities and human linguistic behavior. However, we will also examine in detail the methodologies by which you can integrate an understanding of the intended users, the application task(s), and the business goals of the system to create effective interfaces that provide successful and satisfying user experiences.

Our focus will be entirely on Voice User Interfaces, or VUIs. Moreover, given the ubiquity of the telephone and the large number of spoken language systems currently being deployed for over-the-telephone use, we will focus specifically on VUIs designed for the phone.

---

**What is a voice user interface?**

A Voice User Interface is what a person interacts with when communicating with a spoken language application. The elements of a VUI include prompts, grammars, and dialog logic (also referred to as call flow). The prompts, or system messages, are all the recordings or synthesized speech played to the user during the dialog. Grammars define the possible things callers can say in response to each prompt. The system can only understand those words, sentences, or phrases that are included in the grammar. The dialog logic defines the actions taken by the system, for example, responding to what the caller has just said or reading out information retrieved from a database. The following example is an interaction between a caller and a flight information application:

```
(1) system: Hello…and thanks for calling BlueSky Airlines. Our new automated system lets
you ask for the flight information you need. For anything else, including
reservations, just say “more options.” Now do you know what the flight
number is?

caller: No, I don’t.
```
system: No problem. What's the departure city?

caller: San Francisco

In this application, a voice actor has prerecorded everything the system says. Following the prompt “… do you know what the flight number is?” the system listens, using a grammar that accommodates inputs from the caller such as “no”, “no, I don’t”, “yes”, “yeah, it’s flight two twenty seven”, etc. The dialog logic will then decide what to do next, depending on the answer (e.g., in this case, to prompt for the departure city). If the dialog succeeds, the system will ultimately provide the desired flight information.

The methodologies and design principles for voice user interface design overlap substantially with those used for other types of user interface design. However, there are a number of characteristics of voice user interfaces that pose unique design challenges and opportunities. Two primary characteristics stand out: 1) the modality is auditory; and 2) the interaction is through spoken language.

**Auditory Interfaces**

An auditory interface is one that interacts with the user purely through sound – typically speech input from the user, and speech plus non-speech output from the system. Non-speech output (often referred to as non-verbal audio, or NVA) may include earcons (auditory icons, or sounds designed to communicate a specific meaning), background music, and environmental or other background sounds.

Auditory interfaces present unique design challenges in that they depend on the ability to communicate with a user through transient or non-persistent messages. The user hears something, and then it is gone. There is no screen on which information, instructions, or commands are recorded, as is the case with a visual web interface where items can be accessed over time at will. Users do not have the opportunity to review the system’s output or state at their own pace. Rather, the pacing is largely controlled by the system.

The ephemeral nature of output in auditory interfaces places significant cognitive demands on the user. There are a number of guidelines you can use to make sure your interface designs do not overload the user, do not unduly challenge short-term memory or learning ability, and provide a means for the user to influence the pacing of the interaction. We will cover these guidelines in Chapter 9.

Multimodal interfaces that combine speech with other modalities have the potential to mitigate some of these problems. Even a small screen, when combined effectively with a speech interface, can significantly reduce the cognitive demands on the user, thereby changing some of the design criteria and tradeoffs. However, given the immature state of the multimodal device industry and the large number of spoken language systems that are currently being designed and deployed for use in traditional telephony networks, in this book we will focus on purely auditory interfaces. Many of the same design principles, with appropriate refinement, can be applied to multimodal design. Consideration of multimodal interfaces that include speech will be left for a future volume.

Despite the challenges, auditory interfaces offer a number of unique design opportunities. People rely on their auditory systems for many levels of communication. Listeners derive semantic and other information not only from word choice and sentence structure, but also from the way a message is delivered: from prosody (intonation, stress, and rhythm), voice quality, and other characteristics. By carefully choosing a voice actor (for recording the
prompts the system will play to the user) and effectively coaching the way the prompts are spoken, you can help create a consistent system “persona” or personality. This offers opportunities for branding and for creating a user experience appropriate to the application and user population. We will discuss the crafting of persona in Chapter 6.

Auditory interfaces offer an additional opportunity based on effective use of non-verbal audio. You can use earcons to deliver information (e.g., a sound indicating “voice mail arrived”) without interrupting the flow of the application. Distinctive sounds can be used to landmark different parts of an application, thus making it easier to navigate. Additionally, non-verbal audio such as background music and natural sounds can create an “auditory environment” for the user, thereby creating a unique “sound and feel” associated with a particular business or message. Designers, through effective use of non-verbal audio, can both solve user interface problems in new ways and exploit opportunities for added value.

**Spoken Language Interfaces**

Voice user interfaces are unique in that they are based on spoken language. Spoken communication plays a big role in everyday life. From an early age, we all spend a substantial portion of our waking hours engaged in conversations. An understanding of human-to-human conversation can be brought to bear in order to improve the conversations we design between humans and machines.

Humans share many conversational conventions, assumptions, and expectations that support spoken communication, some universally and others restricted to specific language communities. These conventions, assumptions, and expectations operate at many levels, from the pronunciation, meaning, and use of words to expectations about such things as turn taking in conversations. Some expectations people bring to conversation are conscious, but many operate outside of awareness. Though largely unconscious, these shared expectations are key to effective communication.

An understanding of these shared expectations is essential to the design of a successful spoken language interface. Violation of expectations leads to interfaces that feel less comfortable, flow less easily, and are more difficult to comprehend and more prone to induce errors. Effective leverage of shared expectations can lead to richer communication, and more parsimonious and streamlined interaction. In chapters 10 and 11 we will cover many of the expectations speakers bring to conversations and show you how to leverage that understanding in the design of VUIs.

Two other realities of spoken language have major impacts on VUI design choices and design methodology. First, spoken language skills are learned implicitly at a very young age rather than through an explicit educational process. In contrast, most other user interfaces depend on some specific learned actions designed to accomplish the task at hand (e.g., choosing operations from a toolbar, dragging and dropping icons). Therefore, the VUI designer must work on the user’s terms, with an understanding of the user’s conversational conventions. As designers, we don’t get to create the underlying elements of conversation.

Second, communication through language is largely an unconscious activity in the sense that speakers usually do not explicitly think about word choice, pronunciation, sentence structure, or even turn taking. Their conscious attention is instead on the meaning of the message they wish to communicate. Therefore, despite the fact that we all engage in conversation, designers are at risk for violating the user’s conversational expectations unless they have some explicit knowledge of conversational structure, as discussed in Chapters 10 and 11. Furthermore, design approaches that make explicit the role of each prompt in the conversational context in which it occurs will maximize the ability of designers to bring to
bear their own unconscious conversational skills as they design dialogs. The detailed design methodology discussed in Chapter 8 will show how this can be done.

**Why Speech?**

So, given all the challenges cited above, why bother with speech interfaces?

To answer that question, let’s begin by looking at the Schwab deployment (the application referred to in the opening story). The Schwab application was a big success – both from the point of view of the business and the end-user. Schwab’s original motivation for deploying a speech system for stock quotes was that a large percentage of callers to their touchtone system would immediately press ‘0’ on their touchtone keypad to get a live agent. Those callers were the primary targets for the new speech system.

Most of those callers ended up very happy using an automated speech application, despite their reluctance to use the touchtone system. The primary reason was that they no longer had to enter stock symbols with complex two-keystroke sequences for each letter, as they would have if they used the touchtone system. Now, they could just say the company name and get their quote. Furthermore, many customers actually expressed a preference for talking to the speech system rather than a live agent! They were no longer hesitant to ask for a large number of quotes, or even to ask for the same company twice in the same call if they had forgotten the result or were curious to see if there had been some movement in the price.

The Charles Schwab Company was also pleased with the results. They saved a lot of money, were able to reassign live agents to deal with more complex user requests, and caller satisfaction was high. They were able to reinforce their reputation as leaders in providing new innovations that bring value to their customers.

Since the Schwab deployment, there have been thousands more. You can now talk to automated systems to trade stocks, check airline schedules and book flights, access your bank account, track packages you have shipped, rent a car, check on weather and traffic conditions, get driving directions, check bus schedules, find out what movies are playing nearby, and make restaurant reservations. You can have your own automated “personal agent” that handles your phone calls, calendar, email, and voicemail. You can access voice portals which package together a variety of voice services and, in some cases, personalize them to your needs. The list goes on and on, with new ways of providing information and transactions and facilitating human-to-human communications appearing every day.

The companies deploying these systems are motivated by a number of factors:

- **Save money:** The ROI (return on investment) for speech systems deployment is typically on the order of a few months. When replacing touchtone services, abandonment rates typically go down substantially, and automation rates often rise dramatically. Decreased call durations have also been a factor, saving companies toll call fees.

- **Improve reach:** Companies want to be available to their customers in all places (home and mobile) at all times (24x7x365). In some cases, a system is deployed to complement the self-service they already provide over the web – to reach customers who either don’t have web access or desire the same self-service access when they are away from their desk or in their car.

- **Extend brand:** When you engage in spoken interaction, you don’t get “just the facts.” Speech communicates at many levels. As we listen to someone speak, we make judgments
about numerous attributes: the speaker’s emotional state, social class, trustworthiness, friendliness, etc (Soukup, 2000; Giles & Powesland, 1975). If we design a speech application with these things in mind, we can connect with customers in new ways. The system becomes an extension of the company’s brand and image. In effect, we can design the “ideal employee” – with the right voice, the right personality traits, the right mood, and the right way of handling customer needs and problems. We can present that image reliably, phone call after phone call.

• **Solve new problems:** There are many instances of problems that can be solved, or services that can be offered, by speech applications that were simply impossible in the past. We will provide two examples:
  - *Call routing* – Many companies want to combine their various access phone numbers into one convenient point of entry for servicing all of their customers’ needs. The resulting set of items (e.g., service names) from which a caller must choose is, in many cases, extremely large. Furthermore, the set of choices often do not lend themselves to an intuitive menu hierarchy that callers would be able to navigate. As a result, it has not been possible to create such systems with touchtone technology (or even with earlier generations of speech technology). Recent advances in natural language understanding technology have made it possible to allow callers the flexibility to state their needs in their own words, and then be automatically forwarded to the appropriate service. For example, a system recently deployed by Sprint PCS, with 40 different “routes” or services, allows callers to say things such as: “I just had a call dropped and I’d like to get a credit for it”, “Yes, I want that store in Rockville, on Rockville Pike” and “I’m having a problem with my voicemail, it’s locked up on my phone and I can’t get into it to get my messages.” (These were transcribed from actual calls.)
  - *Personal agents* – Most of us are not rich enough to hire someone to sit beside us in our car, office, and home, and accompany us on trips and as we walk down the street, all the while handling our phone calls, reading us our email, managing our calendar, and carrying the entire world’s white pages and yellow pages. A number of companies now offer automated personal agents that use speech technology to provide such capabilities on the user’s phone.

• **Increase customer satisfaction:** Numerous surveys and deployment studies have shown high user satisfaction with speech systems. The reasons for this can be understood by looking at the list below of end-user advantages of speech

From the point of view of the end-user, there are many appealing advantages of using speech systems, compared to other access modes:

• **Intuitive and efficient:** Spoken language systems draw on the user’s innate language skills. Many tasks can be made simpler and more efficient than with touch-tones. For example, in a travel application a caller may say things like “I wanna leave on June 5” rather than entering some awkward and unintuitive touchtone sequence (such as 0605) after hearing some longwinded instruction. Even simple things like saying the name of your destination city, rather than spelling it on a touchtone keypad, bring significant convenience to the end-user.

• **Ubiquitous:** Telephones are ubiquitous. Many people always carry a cell phone. Spoken
access makes systems available from anywhere, even when far from your desktop.

- **Enjoyable:** A well-designed speech system can provide a user experience that is engaging and enjoyable, at the same time it efficiently meets the user’s needs.

- **Hands-free, eyes free:** Activities such as driving occupy the user’s hands and eyes. Speech is an ideal solution for accessing services while engaged in hands- or eyes-busy tasks. Furthermore, mobile devices generally have very poor user interfaces – entering information of appreciable complexity is awkward with keypads, pointing devices, or handwriting recognition. Speech is the ideal solution.

3 Counting systems from all the major vendors.

---

**Where do we go from here?**

The primary goal of this book is to teach a way of thinking about VUI design. We will focus on the underpinnings of the craft, basic design principles, and design methodology. Along the way, we will present many examples, and show how to apply principles in specific instances to address specific design issues. In the end, you should have a toolbox of techniques you can apply to many design situations. However, a toolbox, by itself, is a dangerous thing. You cannot create an effective design that engages callers, meets their needs, and satisfies business goals by blindly applying rules-of-thumb. As we will see, there are two recurring issues that arise with every design decision, at every step of the design process: consideration of *context*, and understanding of the *end-user*. By context we mean everything from high-level constraints such as business goals to low level considerations such as how the wording of a prompt works in the context of the immediately preceding prompt and response. Understanding the end-user includes addressing how they will react, what their needs are, and what their preconceptions in operation are at each point in time. Consideration of context and end-users must play into every design decision, whether it is a decision to apply a principle, use a specific technique, modify it, or invent something new.

Furthermore, technology is developing rapidly, enabling new, sometimes unanticipated, possibilities. Leveraging new technological capabilities will create new challenges for the VUI designer. You must understand the fundamentals and design principles in order to be able to approach new situations. As a case in point, recent technology advances have made possible call routing applications, of the type described in the previous section. We had to learn new approaches to prompt wording and new ways of creating the appropriate mental model for the caller in order to achieve success with these systems. A **mental model** is the caller’s internal model of how they should interact with the system (Norman 2002; Preece et al. 2002). (The details of call router design will be discussed in Chapter 12.)

As we will see, there are a number of different areas of knowledge that contribute to the fundamentals of VUI design. In this book, four key areas will emerge. Three of them we have already touched on:

- **Human cognitive capabilities:** Understanding the design challenges when dealing with purely auditory interfaces.

- **Human conversational language use:** Uncovering the unconscious expectations and assumptions that humans bring to all their conversations.

- **Methodology:** Understanding the methodological principles and goals (e.g., user-centered design), and applying that understanding to the development of a detailed methodology that enables us to design effective systems that meet user needs and business goals.
The fourth area, not mentioned so far, is technology. Some understanding of speech technology is necessary and must be brought to bear in order to make effective design decisions. Our description of technology will provide the necessary background in order to understand how to leverage the possibilities of the technology and compensate for its limitations. We will see the need for technology understanding as we discuss basic dialog strategy choices as well as design decisions for optimizing accuracy and recovering from errors, developing grammars, and tuning performance.

The book is organized according to the design methodology we recommend. This methodology has been applied by the authors and many others to hundreds of successful deployments.

Each major section of the book discusses one major phase of design. The first chapter in each section describes the relevant steps of the methodology. We follow that with a number of chapters discussing design principles and issues related to that phase of design. Each section ends with a chapter containing a detailed example of a design, showing concretely how the process steps and principles for that section can be applied.

Before we dive into the methodology, a little more introductory material is needed. Chapter 2 will provide an overview of speech technology and introduce some of the technical issues that should figure into VUI design decisions. Chapter 3 will introduce the general methodology that will be covered in detail throughout the remainder of the book.
Chapter 10
Designing Prompts

A VUI dialog is a type of conversation. Perhaps this is even more the case for telephony-based VUIs – after all, the telephone itself is an icon of conversation in our culture. Even though one of the participants in a VUI dialog is a machine, both the user and the designer who engineered the dialog share a few basic expectations that also hold for human-to-human conversations. Both parties expect the use of a common language; each expects the other to be cooperative; and each expects that the other will possess basic cognitive faculties like intelligence, short-term memory, and an attention span (which are artificial in the case of the machine, but are nonetheless expected). There is another level of expectation, implicit in the way that users approach spoken language systems, which has to do with the way prompts are worded.

Humans approach all linguistic tasks with unconscious expectations and assumptions not only about content, but also about form. Since VUI prompts are communicated through spoken language, their form should satisfy the spoken language norms that users naturally expect. And the spoken language system that VUI users are most familiar with is everyday conversation. We recommend that prompts be modeled after conversational language, in the interest of promoting the design of human-centered, user-friendly, optimally comprehensible VUIs. The alternative would be a way of communicating that is less familiar and less natural, and therefore, less comprehensible.

In this chapter, we will demonstrate how specific elements of human-to-human conversation can be applied to prompting. We will review some basic linguistic concepts for creating a spoken-language experience that users will find familiar and engaging. We will also see a number of cases where the intelligibility of prompts is greatly enhanced via the use of conversational language.

It is sometimes thought that “conversational” necessarily implies informality and chattiness. For our purposes, “conversational” will refer more technically to the linguistic form of everyday spoken interactions. Indeed, there are many types of conversations in real life, from chatty and informal to serious and formal. The high-level recommendations made throughout this chapter apply to the design of all VUIs, including those that require an air of formality.

It is also sometimes thought that the goal of “conversational” interfaces is to lead unwitting VUI users into thinking that they are interacting with a human being, which is not what we advocate. Our goal is simply to leverage the linguistic forms that VUI users are best acquainted with. In fact, linguistic research tells us that conversational language incorporates many design features that are specially adapted and ideally suited to the task of communicating via sound. So, many of the principles in this chapter are about improving listening comprehension and enhancing the user’s level of comfort by exploiting the linguistically familiar.
In designing speech applications, we are also designing a language experience for the user. While “speech” technically refers to such matters as acoustics, physiology, sound production and perception, “language” involves the intimate connection of structure, meaning, and use, all of which are context-sensitive. Linguistic form is shaped by forces that range from perceptual and cognitive on the one hand, to social and cultural on the other. The form of conversational language is not haphazard, but systematic and principled. Dialog designers should refer to this naturally occurring system as a resource for prompting, at least to the extent that the current technology allows. As much as possible, naturally occurring spoken language should serve as a guide for writing successful speech applications.

The ideas presented in this chapter should assist you in creating socially and linguistically appropriate characters for your application (see Chapter 6, on persona design). It is a common belief that “persona” depends on hiring the right voice talent. In fact, not even the most experienced, skilled actor with the most pleasantly resonant voice can make inherently non-conversational prompts sound natural, for example, “Please say the duration” instead of a more natural wording like “How long do you expect it to last?” As in the movies, good actors depend on good scripts. To create the sense of a familiar, engaging persona to represent your company’s brand or image, conversational norms should guide the way prompts are written. Reaping the benefits of a well-designed persona experience depends on your ability to write messages that suggest natural spoken language.

In usability studies of a voice-browser application, callers expressed a sense of loyalty to the professional, matter-of-fact, but personal and familiar character representing the system – an administrative assistant in her early 40s. They felt relatively comfortable entrusting to her their most sensitive “personal profile” information, such as credit card details. But a well-designed, likeable persona is essentially an illusion, a mosaic built up of consistent, conversational messages that collectively suggest a coherent, sociolinguistically familiar presence. Usability informants reacted favorably not only to the voice itself, but also to the way prompts were worded (discussed in this chapter) and to the way they were spoken (to be discussed in the next).

The remainder of this chapter will cover the following topics:

- Conversation as discourse
- Cohesion
- Information structure
- Spoken and written English
- Register and consistency
- Jargon
- The Cooperative Principle

Conversation as discourse

Applying elements of naturally occurring language to VUI prompting requires an understanding of conversation as a genre of “discourse,” which refers to stretches of any kind of naturally occurring, human-to-human language. Discourse is language above and beyond the use of individual words and sentences studied in isolation. Discourse analysts make some important assumptions about the nature of human language.

- Language always occurs in a context and its structure is systematically context-sensitive.
• Language is essentially a communicative phenomenon, and its structure incorporates certain design features specially adapted to this end.

In spoken language, there are discourse genres other than everyday conversation, each having their own set of requirements concerning form and content. There are, for example, fairy tales, sermons, news reports, radio advertisements, courtroom trials, lectures, eulogies, airplane safety demonstrations, and auctions.¹

There are a few basic characteristics of naturally occurring discourse, including everyday conversation, that dialog designers should consider for their relevance to prompt writing. These are important considerations even if the VUI you are designing is not an English-language application. Discourse is:

**Principled.** Language is systematic on many levels. Every spoken language has its own particular rules for putting sounds together to make up words and for putting words together to make sentences. For example, “blick,” “ving,” and “brizzle” are not words in English, but they could be. “Try new-and-improved lemon-fresh Brizzle…with grease-busting enzymes!” In contrast, “sbab,” “vlum,” or “knerpf” must all be rejected as possible words, since these forms violate the rules of sound sequencing that are particular to English. Likewise, “That dog seems to be dreaming” is a perfectly fine sentence, yet “That dog seems dreaming” is not. Conversations are like words and sentences in that they, too, conform to a highly structured system in which some creations are not allowed. In the world of VUIs, engineered conversations are at times as “ungrammatical,” in the sense of non-occurring or unacceptable, as are forms like “Sbab” and “That dog seems dreaming. In this chapter, we will see many examples of non-conversational prompts, along with examples of how they can be improved.

**Universal.** Just as all languages have complex syntactic and phonological systems, so do all languages have sophisticated discourse patterns in speech. Even if the VUI you are designing is not an English-language application, the basic concerns of this chapter are applicable to all human languages, though the details may vary from language to language.

**Conventionalized.** Another reason for taking care with the wording of prompts is that discourse is conventionalized, or shared throughout a linguistic community. Regardless of one’s educational background or socio-economic standing, native speakers of English all agree, for example, on the use of the discourse marker “By the way” to mark a conversational contribution as tangential or unrelated, but nonetheless worthy of mention. Likewise, as native speakers of English, we generally place elements of focus at or near the end of the sentence, and as native hearers of English, that is where we expect to find them. It is not possible for VUI designers to invent a new feature of discourse, but they may capitalize on those conventions that already exist.

**Unconscious.** The mechanics of discourse generally escape our awareness. Consider, for example, the use of “this” and “that” as discourse “pointer words.” As described later in this chapter, native-speakers of English use these words differently in speaking versus writing, though very few are aware that a difference exists. We might notice that something doesn’t sound right when a non-native speaker uses “this” instead of “that,” but most people are unable to articulate exactly why. As it happens, the way that pointer words are used in many VUIs corresponds to the written rather than spoken mode of communication. Since VUI

¹ Some examples of discourse genres particular to writing are: short stories, love letters, warning labels, road signs, email messages, advice columns, recipes, cover letters, offer letters, appliance warrantees, horoscopes, gossip columns, shopping lists, obituaries, and anything from the Internal Revenue Service.
prompts are spoken, however, pointer words in prompts should reflect the conventions of speaking rather than writing. Writing spoken dialog is different from writing language that is to be read.

The following sections will center on some specific features of spoken language, and how they can be leveraged to optimize comprehension, comfort, and familiarity in your VUI.

---

**Cohesion**

Cohesion is the glue of discourse. By way of example, consider the following written piece of discourse

(1) George couldn’t wait to get to France. However, he didn’t stay there long.

“He” refers to “George,” “there” refers to “France,” and “however” establishes an adversative or contrastive relationship between the proposition that George was eager to go to France and that George’s visit to France was short-lived. The example illustrates three types of cohesion devices. The term “cohesion” refers to explicit linguistic devices, like “however,” “he,” and “there,” which help bind the stretch of language into a coherent whole. Coherence, in turn, refers to the functional unity of a piece of discourse. Cohesion cannot only be found in a written monolog, as in the example above, but also across all dialogs.

Cohesion devices facilitate and reinforce comprehension of the whole, but they themselves are not responsible for creating meaning. Rather, they are natural cues that speakers and hearers use to signal and retrieve meanings that underlie utterances in context. A few cohesion devices that are particularly relevant to prompt design are pronouns, discourse markers, and special pointer words like “this” and “that”.

**Pronouns**

A special feature of human-to-human language is that information in one unit of talk often presupposes information presented in a previous one, as exemplified in (2)-(4) (Schiffrin, 1998).

(2) I saw the cat the other day. It was still wandering around without a home.
(3) I saw a robin the other day. It was the first one I saw this spring.
(4) We moved here in 1982. We didn’t even have jobs then.

In conversation, pronouns like *it* and *one*, as well as time adverbs like *then*, let the listener know that the referent at hand is the *same* as a referent uttered previously. In (2)-(4), the reader or listener can easily make sense of these sentences by recovering the missing information from an earlier point in the discourse.

---

2 In writing, however, pronouns sometimes refer to something *earlier* in the text, e.g. “As soon as she got home, Pat called her lawyer,” where *she* and *Pat* are co-referential. In conversation, we would more likely say, “As soon as Pat got home, *she* called her lawyer.”
Compare the stilted effect created by the repetition of the term “bookmark(s)” in the prompts listed in (5) with those in (6), rewritten with the pronoun one.

(5) system: You have five bookmarks.
Here’s the first bookmark...
Next bookmark.
…That was the last bookmark.
caller: Delete a bookmark.
system: Which bookmark would you like to delete?
...
system: Do you want to delete another bookmark?

(6) system: You have five bookmarks.
Here’s the first one...
Next one...
…That was the last one.
caller: Delete a bookmark.
system: Which one would you like to delete?
...
system: Do you want to delete another one?

With the simple replacement of the noun “bookmark” with the pronoun “one,” the prompts in (5) flow more naturally than those in (6). If VUI designers write prompts in context and listens to the dialog that emerges, they will be more likely to appreciate the ubiquity of pronouns in spoken language and to reap their stylistic benefits.

**Discourse markers**

Discourse markers are defined as “sequentially dependent elements, which bracket units of talk” (Schiffrin, 1987). They connect utterances (or portions thereof) by succinctly relating what has just been said to what is about to be said through time, result, contrast, and other such notions. The following is a list of discourse markers, organized according to class of use (adapted from Quirk & Greenbaum, 1973).

- **Enumerative:** first, second, third; for one thing, and for another thing, to begin with, for starters; in the first place, in the second place; one, two, three...; a, b, c...; next, then; finally, last, lastly; to conclude
- **Reinforcing:** also, furthermore, moreover, then, in addition, above all, what’s more
- **Equative:** equally, likewise, similarly, in the same way
- **Transitional:** by the way, incidentally, now
- **Summative:** then, (all) in all, in conclusion, in sum, to sum up
- **Apposition:** namely, in other words, for example, for instance, that is, that is to say
- **Result:** consequently, hence, so, therefore, thus, as a result, somehow, for some reason or other
- **Inferential:** else, otherwise, then, in other words, in that case
- **Reformulatory:** better, rather, in other words
- **Replacive:** alternatively, rather, on the other hand
- **Antithetic (or contrastive):** instead (blend of antithetic with replacive), then, on the contrary, in contrast, by comparison, (on the one hand...) on the other hand
- **Concessive:** anyhow, anyway, besides (blend of reinforcing with concessive), else, however, nevertheless, still, though, yet, in any case, at any rate, in spite of that, after all, on the other hand, all the same, admittedly
- **Temporal transition:** meantime, meanwhile, in the meantime
- **Attitudinal, commenting on truth:** actually, in actuality, in (actual) fact, strictly speaking; nominally, officially, technically, theoretically

Occasionally customers resist the use of discourse markers in their VUIs on the grounds that they are perceived as “informal” or “slang.” In and of themselves, however, discourse markers as a linguistic category are neither formal nor informal. The following markers, for example, are actually more formal than what we are used to hearing in everyday conversation: *in addition, thus, therefore, hence, nevertheless, on the contrary, conversely, by comparison, in contrast, equally, all other things being equal, rather, and in conclusion.* Markers that epitomize informal conversation are, for example, *for starters, anyhow, and so.*

The benefits of discourse sensitivity, which includes the appropriate use of discourse markers, to an engineered dialog are significant. Compare examples (7) and (8), below (from Giangola, 2000). Both sets of prompts aim to collect the same pieces of information in order to schedule an appointment in a personal information manager application. The prompts in (7) conspicuously lack discourse markers, while those in (8) were rewritten to project a natural, conversational style. Here we find both kinds of cohesion devices that have been discussed thus far, pronouns and discourse markers (both underlined).

(7) Please say the date.
- Please say the start-time.
- Please say the duration.
- Please say the subject.

(8) First tell me the date.
- **Next,** I’ll need the time it starts.
- Thanks...now how long is it supposed to last?
- **Last of all,** I just need a brief description...

Discourse markers add value to prompting in a few ways. First, recall the central assumptions that discourse analysts make about human language. Language is a context-sensitive system that incorporates design principles specially adapted for communication. The redundant nature of discourse markers reinforces the functional relationship between two units of discourse. The discourse marker in the utterance “Joe is a couch potato; his brother, on the other hand, ...” induces us to expect sharp contrast, giving us a pragmatic preview of what comes next. You, the reader, might have thought to complete the sentence with “is a triathlete,” but not “hates to exercise.” In (8), discourse markers like *first, next,* and *last of all* orient listeners to their position in a sequence of questions.

Second, discourse markers suggest a human-like awareness of how the dialog is progressing. Not only does the application recognize specific responses like “yes,” “no,” “three p.m.,” and “operator,” it seems to recognize how the user’s contributions to the conversation relate to those of the system on more abstract cognitive and social levels. In these roles, discourse markers can be said to serve “conversation management” functions.
This point will be most robustly demonstrated below, in discussion of markers like *oh, by the way,* and *actually,* which are deceptively simple but functionally loaded.

As we can see from these examples, discourse markers serve a critical role in imparting naturalness to engineered dialogs, as well as their essential role in projecting a familiar and favorable persona. In both robotic, context-insensitive (7) and natural, context-sensitive (8) dialogs, the information gap that initially exists between the application and the user is resolved only one piece at a time, which has the potential to frustrate. Usability feedback reveals that the wording in (7) actually heightens the listener’s awareness that information is being collected piecemeal. One respondent said that hearing one “Please say” after the next was annoying, presumably owing to its unrelenting repetition. In general, listeners associate the dialog in (7) with such negative descriptors as “frustrating,” “impatient,” and “bored,” while the dialog in (8) is associated with the positive descriptors “attentive” and “helpful.”

Let us now turn to the use of specific discourse markers and how they can enhance the natural flow of prompts in speech interfaces. These markers are: *now, by the way, oh,* *otherwise,* and *actually.* We will also look at conversational uses of *okay* and *sorry.* These are not typical discourse markers, since they can stand alone, but they function like discourse markers nonetheless.

*now*

*Now* functions as a kind of paragraph-changer, a verbal indentation, as we saw above in “Thanks, *now* how long is it supposed to last?” (8). This discourse marker frequently introduces a new stage in a thought sequence, marking a progression of topics. For example, in cases where a single prompt has more than one topic to cover, the discourse marker *now* subtly and gracefully ushers the caller’s attention to the next topic in the sequence. In example (9), the caller has just completed a number of steps required in enrolling the caller’s “voiceprint.”

*icon*

(9) *system:* That’s all I need -- we are done getting your voiceprint. *Now,* since you’re new, would you like some help getting started?

Because “now” signals a change in topic in the discourse and does not refer to actual time, we can say things like “*Now, in the old days*...”

*by the way*

‘Just-in-time instruction’, the instructional technique introduced in Chapter 9, lends itself to introduction with the discourse marker *by the way,* as exemplified in (10).

*icon*

(10) *caller:* Call 510 465 2468

*system:* Calling 510 465 2468.

*By the way,* when you’re done with this call, you don’t have to hang up...Just say ‘come back’.
As in human-to-human conversation, *by the way* introduces a tangential thought, often of parenthetical importance.

*oh*

Such a small word, and yet volumes could be written about the discourse functions of *oh*. While most dictionaries sweep *oh* under the carpet as a mere “hesitator” or “filler” expressing surprise, Shiffrin (1987) proposes that the central, prototypical function of the marker *oh* is to signal some sort of cognitive mismatch, communicating something along the lines of “what I know or what I think does not match what you know or think.”

In (11), the caller attempts to navigate to a site that is not considered secure. The interface informs that this will not be possible and tells why.

```
icon
(11)   caller:       Go to EvilSite.org.
system: Oh, I can’t go to that site -- it’s not considered “secure.”
```

In (12), the same interface informs the caller that his personal profile cannot be “bookmarked” (i.e. referenced as a preferred voice site).

```
icon
(12)   caller:       Make that a bookmark.
system: Oh, that can’t be bookmarked…
```

This use of *oh* is related, by extension, to a “self-editing” use, as exemplified in (13) and (14), where the vertical bar represents the start of a new audio file.

```
icon
(13)   caller:       Go to sports.
system: Sure, sports! | Oh, that’s not available at the moment, but would you like to try something else instead?
```

```
icon
(14)   caller:       Get my messages.
system: OK, ‘messages’… Oh, looks like you don’t have any messages right now.
```

As a self-editing marker of cognitive mismatch, *oh* is invaluable in cases like these. The system response in (13) started off with the implication that sports was available, and in (14) that messages were. The second message in each case contradicts the initial assumption. But introduced by *oh*, the otherwise incompatible piece of information is ushered in smoothly and gracefully, cast as a sudden change in the system’s knowledge. In other words, the character the voice actor is playing only just now found out about there being no sports or messages, so it can’t be held at fault. The character was just trying to be attentive and cooperative initially, but for reasons beyond his or her control, was unable to deliver.

Be careful not to pepper the dialog with *oh* for reasons that are hard to justify. For example, the caller opts for “weather” at the main menu and the interface responds with the acknowledgement “Oh, weather.” Depending on how you say it, this phrase is open to a number of interpretations. One reading could be that the interface’s persona has temporarily forgotten that weather existed as an option. Another reading might be that the persona is surprised that the caller has asked for weather, as opposed to some other option. Yet another possibility is that the persona is secretly wishing that the caller had not made this particular selection. (Along these lines, think about how the voice actor is supposed to say “Oh,
weather.” How would you coach the delivery?) These interpretations all resonate with Schiffrin’s analysis of oh as a marker of cognitive mismatch, but this use is not appropriate here. In contrast, an appropriate context for “Oh, weather” would be to recover from a misrecognition, as in “Oh, weather,” as opposed to what had been misunderstood as a request for “traffic” or “horoscopes.” In this context, the discourse marker appropriately signals cognitive mismatch: “what I heard before isn’t what I’m now hearing.” In any case, a more appropriate wording for the original context cited above would be, “Sure, weather” or “Okay, weather,” which communicate more neutral acknowledgements.

Since different discourse markers have different context-sensitive uses, it is critical that you choose carefully.

actually

Similar in use to oh, there are a number of discourse markers that suggest that an utterance may not be completely true or exactly right in reality, e.g. officially, technically, and theoretically. Other discourse markers, such as actually (15), really, in (actual) fact, and in reality, indicate what is considered to be the “real” truth.

(15) caller: Get my personalized weather report. 
    system: Actually, you had a list of cities for me to check on, but it’s been deleted. To make a new list, go to the registration website. For now, just tell me what city and state you want the weather for…or a city and a country.

One application we have worked on makes use of actually in certain error prompts, as in (16). The caller tries out the navigational command “go forward,” although he or she has not yet visited any voice sites yet.

(16) caller: Go forward. 
    system: Actually, I can’t move ahead since we haven’t gone anywhere yet. To hear a list of places you can go, just ask what’s available.”

Actually introduces the “real truth” and prepares the listener for some detail that he or she is unlikely to have anticipated.

otherwise

This is probably the most abused of discourse markers in speech interfaces. It frequently appears in error-recovery prompts, as exemplified in (17).

(17) If you’d like to leave a message, say “yes”; otherwise, say “no.”

In real-life conversations, the information that follows otherwise is not strictly predictable from what came before, in spite of this marker’s literal “if not” definition. The natural, informative use of otherwise is illustrated in (18) and (19).
(18) If you’re running late, call me. Otherwise, I’ll come by at eight-thirty.
(19) We better leave now, otherwise we won’t get good seats.

It would be conversationally abnormal for someone to say “If you’re running late, call me. Otherwise, don’t call me” or “We better leave now, otherwise we’ll leave later.” These “otherwise” alternatives are so predictable from their antecedents that the phrase as a whole is devoid of informational meaning. There is a conversational principle that can be roughly paraphrased as “Be informative.” (This principle is a corollary to the Cooperative Principle, see below.) The use of otherwise in prompts like (17) does not move the hearer into a sufficiently informative space, as would be expected in authentic discourse.

Prompts like (17) may sound unnatural to users for another reason. “Yes” and “no” were natural responses to the state-initial prompt, which would have been something like “Do you want to leave a message?” some two turns earlier in the conversation. What (17) is really saying is, “If you want to leave a message, answer as directly as possible my last question, because we’re in the same recognition state and so the same recognition grammar is active.” So the wording in (17) is designer-centric because it assumes familiarity with the concept of recognition states and active grammars.

In user-centric prompting, in contrast, “otherwise” would direct the caller into a more informative space, as in (20), or the whole prompt could be reworded in order to avoid its use altogether, as in (21).

(20) If you want to leave a message, start talking after the beep. Otherwise, … feel free to hang up.
(21) Sorry, did you want to leave a message?
   Yes or no, did you want to leave a message?
   Did you want to leave a message？ [parenthetically] (Just say ‘yes’ or ‘no’.)

The next two items are not discourse markers in the technical sense, but they serve a similar transitional function.

okay

In a certain application we have designed, whenever the caller answers a question with “no,” or a synonym of “no,” the following prompt begins with okay, or some other acknowledgement, as exemplified in (22).

(22) system: Did you want to review some more of your personal profile?
caller: No.
system: Okay, what’s next?

Okay acknowledges the caller’s refusal, and smoothly paves the wave for the next dialog state. Notice that by removing okay (23), the result is not only a more stilted-sounding dialog, but also the vague impression that the system might not have heard what the caller just said.

(23) system: Did you want to review some more of your personal profile?
caller: No.
system: What’s next?
In (22), okay could be substituted with all right, no problem, or not a problem, depending on the persona.

**sorry**

A common criticism of the use of sorry in prompting is, “But, come on, now...how can a computer be sorry?” However, sorry is useful not so much for apology, but rather as a signal that the upcoming contribution will somehow fail to meet the listener’s expectations. It can be employed as a mild bracing advisory. When your local video store clerk informs you that the movie you seek to rent week after week is yet again unavailable, there may be no hint of regret in the tone of “Sorry, it’s out” or “Sorry, we don’t have it.”

In the recording studio, skilled voice actors can read “sorry” more as a transitional device than as a genuine expression of full-fledged, heartfelt regret. For example, in one application we have worked on, second-time recognition errors trigger the generic prefix-type prompt in (24), as well as a number of randomized paraphrases. The voice actor’s delivery of “sorry” is professional and neutral-sounding, a transitional device that mirrors the conventionalized use of “sorry” in day-to-day public life.

Avoid using discourse markers just to make the interface sound more like a real person. In our view, the primary purpose of using discourse markers in a VUI is to facilitate comprehension, thereby reducing cognitive load, not to create the deception of a human attendant. Discourse markers, after all, are a design feature of natural discourse that allows speakers to reinforce meaning relationships between utterances, in turn allowing listeners to decode them more easily. Prompt writers should take advantage of this “design feature” of natural discourse, but with care.

**Information structure**

For the purposes of this book on VUI design, let “information structure” simply refer to the placement of contextually-determined “old” versus “new” information, whether in recorded messages or in natural conversation. By way of example, consider the following interaction between one of the authors (JG) and a local directory assistance service.

```
(25) live 411 Operator: Hi, this is Joan. What city?
caller: Mountain View.
live 411 Operator: All right?
caller: CostPlus.
recording: The number you requested, six, five, zero, nine, six, one, six, zero, six, six, can be automatically dialed by pressing one now. An additional charge will apply.
```

Except for the prosody of the phone number (see Chapter 11), there is nothing strange about the recorded message, as far as individual words and sentences go. Yet the way the system gives the requested number is nonetheless uncomfortable. Although we have not conducted a formal study, anecdotally people say, “I have to hang up in the middle of the sentence if I want to remember [the number].” The problem has to do with the way this
recording figures in context – specifically, the way that the “old” versus “new” information has been laid out for the listener. From the point of view of most callers, the new piece of information, or “focus,” should be the requested phone number (“650 961 6066”), yet this information is buried in the part of the sentence that is conventionally reserved for old or non-focal information. Such placement actually facilitates our forgetting what the number is, though finding it out was the purpose of the call to begin with.

VUI designers and prompt writers should be aware that there is a neutral (default) position of focus in English, as determined by what is called the “End-Focus Principle” (Quirk & Greenbaum, 1973). That is, native speakers of English naturally place new or focal information at or near the end of the sentence, and this is where native hearers of English naturally expect to retrieve it. An example of a message that is conversational, that not only complies with the End-Focus Principle but also facilitates short-term memory, would be: “Here’s the number: 650 961 6066. For automatic dialing, press ‘one’…”

The way people naturally structure information is highly context-sensitive. In the case of the directory assistance recording, it is context that makes the positioning of the phone number inappropriate. It is easy to imagine some other context where a phone number occupies this position appropriately, as in frequently heard messages of the type “The number you have dialed, 444-4444, is not in service. Please check the listing and dial again.” In this case, context establishes the already-dialed phone number as old information, and so its non-focal placement is justified.

To be technically accurate, “end focus” is given to the last “open class” item or proper noun in a clause (Quirk & Greenbaum, 1973). Open-class items are nouns, adjectives, verbs, and adverbs. These categories are called “open” because they are expandable, open to new members (witness the creation of such verbs as “downsize,” “email,” and “dis”), while closed-class items are articles, demonstratives, prepositions, and conjunctions (these are categories that reject new members – no one can invent alternatives to words like “the,” “that,” “of,” or “and”). In the sentence, “Sean Connery was born in Scotland”, the last open-class item is the noun “Scotland.” By default, it is the focus, the new piece of information in this sentence. In contrast, “Sean Connery” is the “topic” of the sentence, or the old piece of information upon which the speaker makes some comment. Old information is generally placed in the subject, while new information is generally housed in the predicate.

Because prompts are sometimes written out of context, it is easy to find examples that violate principles of information structure, which is context-sensitive. In a certain application that recognizes dates, there are special messages associated with months that have only thirty days. These messages are played only when the recognizer returns, for example, “June thirty-first,” or “February twenty-ninth” in a non-leap year, with high confidence. To comply with the End Focus Principle, we would expect such messages to be worded as in (26), but in fact, they were worded as in (27).

![icon](26)  
**caller:** Make it for June thirty-first.  
**system:** Actually, June has only thirty days.

![icon](27)  
**caller:** Make it for June thirty-first.  
**system:** There are only thirty days in June.

---

3 Focus can occur at earlier points in the sentence, but this requires the use of special, “contrastive” stress and intonation, as in “Sean CONnery was born in Scotland,” which would answer the question of who was born in Scotland.
The wording in (26) appropriately casts “June” as the subject of the sentence. In this position, it assumes the role of “topic,” upon which some new comment will be made. The new information is the “only thirty days” part, which is appropriately cast in the predicate, where it receives end focus.

It seems plausible that the End-Focus Principle has something to do with the short-term memory phenomenon known as The Recency Effect, as we discussed in Chapter 9. Elements that come at the end are more prominent and easier to remember precisely because they are the last thing one hears. To exploit these linguistic and psychological principles, it makes sense to structure directions such as the novel element, such as what key callers should press or what special voice command they should say, follows the topic, which is some objective or task. Preferably, the touchtone or voice command should be placed at the end of the prompt.

Careless handling of focal information can lead to ambiguity and wrong inferences. For example, a certain application allows callers to choose from a list of song titles and to download the musical selection to a cell phone. To select a song, the caller is supposed to say “That one!” upon hearing the desired title. In the sample interaction in (28), the system prompts the caller to say whether or not it got the selection right, simultaneously informing the cost:

(28) ([System plays list of Britney Spears selections] …, “Oops, I Did It Again,” …
caller: That one! system: Would you like to buy the song “Oops, I Did It Again” for $4.99?

So what is being questioned here? Is it the song selection (“Oops, I Did It Again”) or the price ($4.99)? Context and world-knowledge likely indicate the former, but information structure, with price in end-focus position, suggests the latter, as if the price were open for negotiation. In other words, the prompt is ambiguous, permitting two valid interpretations. As a legitimate response, a haggle-prone caller might legitimately reply, “No, but I’ll give you a buck fifty for it.” The prompt can be recast unambiguously as something like: “Good choice… ‘Oops, I Did It Again’ is only $4.99. Do you want to go ahead and buy it?” Here, the price has been altogether removed from the scope of interrogation, and the prompt has only one possible interpretation.

In light of the End-Focus Principle, we should have a renewed appreciation for the communicative function of the passive voice in conversation, as well as in prompts. Consider the active and passive sentences in (29) and (30), respectively. In both cases, “Leif Erikson” is the “agent” (that is, the do-er), while “North America” is the “patient” (the affected element) of the verb “discover.”

(29) **ACTIVE VOICE** (answers “What did Leif Erikson discover?”)

Leif Erikson discovered North America.

AGENT PATIENT

TOPIC/OLD INFO FOCUS/NEW INFO

(30) **PASSIVE VOICE** (answers “Who discovered North America?”)

North America was discovered by Leif Erikson.

PATIENT AGENT

TOPIC/OLD INFO FOCUS/NEW INFO
These examples illustrate a fundamental difference in the use of the active and passive voices. In the active voice, the agent is the topic or old information while the patient is new or focal. In the passive voice, these roles are reversed: the patient is the topic (or old), and the agent is the focus (or new). So the passive voice is useful because it allows the speaker to reverse the usual order of agent and patient, and this is done in order to reverse their default informational status.

The choice of the passive voice is more than appropriate – it seems to be required – in the recorded announcement in (31).

(31) This program has been made possible by a grant from the Ford Foundation.

If we attempt to rewrite the sentence in order to “avoid the passive”, as suggested in some texts on writing style, the result is strangely off-balance: “A grant from the Ford Foundation has made this program possible.” In the context in which we would hear (31), the piece of information that requires focus is “the Ford Foundation.” What is old, topical information is “this program,” the one we’ve been watching for the past hour. The activized version of (31) strikes the ear as “off balance” precisely because it flouts the End Focus Principle.

The passive is used appropriately in the following two examples from a telecommunications VUI. The company that owns the application wanted to make it very clear to callers who exactly would be handling their request. Some requests would be handled by the new, automated speech system in question, while others would be routed to customer service representatives, the same as before. The exchanges in (32) and (33) illustrate these two dialog possibilities.

(32) caller: Copy of my bill.
    system: Your request for a copy of your bill will be handled by our Automated Express Service.

(33) caller: Clarification, please.
    system: Your request for clarification of an item on your bill will be handled by the next available representative.

In the system’s responses in (32) and (33), the old information, “your request for a copy of your bill” and “your request for clarification,” respectively, is maintained as old information in the subject slot, while the new pieces of information, “our automated express service” and “next available representative,” are accordingly given end focus. If the prompt writer had put these sentences in the active voice in order to satisfy his grammar-checking software, the result would have been a stilted, context-insensitive messages such as “Our automated express service will handle your request for a copy of your bill.”

This discussion of information structure reveals the significance and the implications of speech’s linear organization. The prompt-writer must consider the role of context in judiciously sequencing elements for a familiar, comprehensible listening experience.

---

**Spoken and written English**

In both touchtone and speech applications, it is easy to find messages that reflect the norms of written rather than spoken language, e.g. “I must first verify your account number,” and “You may now record your message.” Spoken and written languages differ for a number of reasons (Crystal, 1992). Speech is dynamic and transient. No sooner is a
word or phrase spoken and it’s gone for good. Also, in spoken language, participants are present at the communicative event. Prototypically, they interact face to face. The speaker usually has a clear notion of whom he or she is addressing. Writing, in contrast, is static and permanent. Written words are there for as long as the medium lasts. Because of this permanence, written language is generally more formal than spoken language, and it ensures the survival of certain more conservative words, phrases, and grammatical structures, for example, “To whom it may concern.” In addition, there is always a lag, the duration of which is often unknown, between the time of the writing and the time of the reading, as well as different settings, which may also be unknown, where the writing and the reading take place.

Another important difference is that the writer is distant from the reader, cannot see the reader, and may not know who the reader is. This is of course the case with speech applications as well. However, all speech applications crucially depend on spoken language, which has evolved in a different direction than writing because speech is prototypically face-to-face and prototypically more personal. To find yourself participating in a conversation, even one that has been engineered, with a persona who is reading formally written text at you, the content of which depends on your own unrehearsed, spontaneous responses, is unprecedented in authentic discourse. Because it cannot be likened to any real-world experience, we can actually consider it an “anti-metaphor,” and so it is undesirable for interface design. The highest-level metaphor that will most easily usher the VUI user from recognition state to recognition state is that of “an everyday conversation.” VUI prompts should therefore adhere to spoken rather than written norms, still guiding and directing the user through the interaction.

In the sections that follow, we will examine a few differences between speaking and writing in English and demonstrate their relevance to prompting. These differences motivate the methodological suggestion in Chapter 8 to read your prompts aloud – as you listen to them, you will react differently than when you merely read them on the printed page.

**Pointer words**

One manifestation of the many differences between spoken and written language is in their use of “pointer words,” for example, this, that, here, and there. The most basic use of pointer words is to indicate position or location of things in physical space, e.g. “This car isn’t for sale, but that red one over there is.” As this example illustrates, the distinction is roughly equivalent to “near” versus “far.” It is typical, however, for languages to extend the use of the physical pointers to the discourse domain, in other words, to point either forward or backward in a stretch of written or spoken language, e.g. “Here’s what happened…” (pointing forward, to what is about to be said) and “There you go again!” (pointing backward, to what someone else has just said).

Prompt writers should be aware that the directionality of this vs. that, in particular, differs depending on whether they are used in spoken versus written language.

(34) SPEAKING, BACK-POINTING
I tried to open the door, but that didn’t work.

(35) WRITING, BACK-POINTING
I tried to open the door, but this did not work.
These examples demonstrate that the back-pointer of choice in speech is *that* (34), while the back-pointer of choice in writing is *this* (35).

The next pair of examples (Quirk & Greenbaum, 1973) relates to speaking only. *That* points backward (36), while *this* points forward (37).

(36) SPEAKING, BACK-POINTING

They get Newsweek. I’d never subscribe to a magazine like *that.*

(37) SPEAKING, FORWARD-POINTING

He told it like *this:* George was running down the road and all of sudden,...

In conversation, therefore, statements like the ones in (38) – (40) are most easily interpreted as forward pointing. (In other words, “this is” is synonymous with “here’s.”)

(38) This is the last item on the list.
(39) This is your last message.
(40) This is the end of your favorites list.

Though these statements are used and interpreted as forward-pointers in everyday conversational contexts, we have observed that they are sometimes used in VUIs to point backwards, in compliance with the norms of written English. The unfamiliar use of spoken *this* as a back-pointer in (38)-(40) would be likely to impede listening comprehension.

An example of a prompt that uses both *this* and *that* appropriately is shown in (41).

(41) *This* is the number I heard: 408 749 9591. Did I get *that* right?

*This* points forward to the phone number and could be replaced with ““here.” *That* points backward to the number and could be replaced with “it.”

**Contraction**

Spoken language tends to favor contractions (e.g. *you’re, can’t, it’s, don’t*) while written language tends to avoid them (e.g. *you are, cannot, it is, do not*). Example (42) is a conversational prompt with three contractions.

(42) ..Finally, if *you’re* finished with your bookmark list, just say “I’m done”, and

*I’ll* put it away.

Do not expect a professional voice actor to take the liberty of substituting contractions for uncontracted forms. While your voice actor may do you this occasional favor, in fact, those “in the business” consider it unprofessional to edit scripted material extemporaneously, unless there is some understanding or agreement to the contrary.
Contractions have occasionally been discouraged in VUI designs based on the incorrect assumption that they are characteristic of “lazy” speech. If contractions were lazy, then people who are feeling lazy should be able to respond affirmatively to a question like “Are you hungry?” with simply “Yes, I’m.” Native speakers of English unanimously reject the contraction in such cases. Another impossible contraction is “you’re” in “Sandy’s tired, and you’re, too.” Examples like these suggest that the distribution of contractions in speech probably has more to do with syntactic and rhythmic constraints than on undesirable personality traits like sloth or carelessness.

In other cases, contractions have been avoided on the grounds that the content of the VUI is, for example, financial, and banking demands a certain level of formality. There is no evidence, however, that people stop using contractions in speech just because they work in a bank or are trading shares in a volatile market. Contractions should be used in your VUI to the extent that they are a defining characteristic of everyday spoken language, regardless of the conversation’s level of formality. The point is simply to provide users with a familiar, comfortable language experience.

**Must and may**

*Must* and *may* each have two basic meanings. One can be loosely described as social/interactive and the other one as a logical/probability meaning (Celce-Murcia & Larsen-Freeman, 1999). The social/interactive meaning of *must* is “obligation” or “necessity” (43), while the logical meaning is “deduction” or “inference” (44).

(43) **SOCIAL must**

Passengers seated in an emergency row must read the safety information card.

The applicant must fill out form 27B and retain the yellow copy for his or her records.

Fully qualified candidates must hold a bachelor’s degree.

(44) **LOGICAL must**

Who’s at the door? ...Oh, that must be the plumber.

It must’ve rained last night.

You must think I’m crazy for pulling a stunt like that.

Melrose (1983) found in the speech of native speakers of North American English that *must* is reserved for inferences, which is the logical/probability meaning. In other words, although people use “must” to signal obligation in writing, they avoid it in speech, even in formal settings. Instead of *must*, obligation or necessity is generally expressed with *have to* (e.g. “Sorry, but I have to leave now.”).\(^4\) Finally, Melrose notes that social *must* seems to be

\(^4\) Melrose also found that (*have) got to/gotta* is used in both social and logical contexts, especially when there is some sense of urgency, e.g. “You gotta lend me ten dollars” (social use signaling obligation) and “You’ve gotta be kidding me” (logical use signaling inference).
more common in British English, e.g. “You must come over for dinner soon” and “We must correct that problem as soon as possible.”

The stilted quality of the prompts in (45) and (46) can therefore be repaired by replacing social *must* with more conversational alternatives, as in (47) and (48), respectively.

```
(45) You must say your PIN one digit at a time, for example, two one zero zero.
(46) We must first get your starting point.
```

```
(47) Go ahead and say your PIN one digit at a time, for example, two one zero zero.
(48) First let’s get your starting point.
```

Like *must*, the verb *may* also has social/interactive versus logical/probability uses, and also like *must*, it is only the logical/probability use that we find in everyday conversations. The social/interactive use of *may* is permission (49), while the logical use is possibility (50).

```
(49) SOCIAL may
You may leave the room.
May I be excused?
Cell phones and two-way pagers may not be used during flight.
```

```
(50) LOGICAL may
It may rain tomorrow.
I may have a couple of twenties in my wallet.
You may not be getting enough sleep.
```

With social *may* (49), the speaker is implicitly referencing a position of authority and an inequality in status, so this is a relatively formal way to grant or request permission. In these cases, *can* is more egalitarian and friendlier. In North America, there is less social stratification than in other language communities around the world, so, not surprisingly, *can* is more prevalent than social *may*, and is preferred in contexts like You can (may) go now (Bailey, 1997; Melrose, 1983)). Presumably, North Americans are put off by the implicit social inequality, formality, and coldness of *may* in its social use. In contrast, social roles and setting are irrelevant to the use of *may* in the logical/probability context. In (50), the speaker is assessing the probability of rain, independent of the social aspects of the conversation (Celce-Murcia & Larsen-Freeman, 1999).

In the interest of projecting a more favorable persona, social *may* in (51) and (52) should be reworded as in (53) and (54), respectively.

```
(51) ...At any time, you may also ask for help.
(52) When you are finished recording, you may hang up or press pound for more options.
```

```
(53) ...At any time, you can also ask for help.
(54) When you’re finished recording, feel free to hang up – or to hear some more choices, press the pound key.
```

Chapter 10, 18
Note in (54) the contraction “you’re” and the restructuring of information to accommodate the End Focus Principle.

**Will versus going to**

In general, *will* is more formal and impersonal as an indicator of futurity, while *going to* is more informal and personal. Because it is more formal, *will* is often preferred in writing. For native speakers of English, however, there are a few cases in which these verb forms cannot be interchanged. In conversation, *will* is commonly used to signal spontaneous volition (55)-(56) and future contingency (57)-(58), while *going to* signals planned intention (59)-(60) and “the future on its way” (61)-(62).

**Will**

(55)  
A:  That was my mother’s favorite vase.  
B:  Oops! I’ll buy you another one.

(56)  
A:  I can’t really talk right now.  
B:  Okay, I’ll call you tonight.

(57)  
If you buy a house, you’ll get a huge tax write-off.

(58)  
If you put your pawn there, he’ll win the game.

**Going to**

(59)  
A:  What’s Harvey doing with my pliers?  
B:  He’s going to fix the TV antenna.

(60)  
A:  What’s on your agenda for tomorrow?  
B:  I’m gonna get a haircut.

(61)  
Stop the car! I’m going to be sick!

(62)  
Look, it’s going to rain.

If *going to* in these examples is replaced with *will*, the result is very unnatural, yet *will* is often used in VUI prompts where we would conversationally expect *going to*. Take, for example, the prompt in (63), which is part of a lengthy enrollment process in a subscriber-based service and is followed by some instructional information.

(63)  
I will now record your account number. Say the number one digit at a time…

This prompt can be rewritten to conform to conversational norms by replacing *I will* with *I’m going to*. In addition, the placement of the adverb *now* between the two verbs (*will* and *record*) yields another structure particular to written discourse, or in some formal spoken genres, such as when a professional magician addresses his audience, “*I will now saw my lovely assistant in half.*” We can rewrite (63) as (64).

(64)  
Now I’m going to record your account number. Tell me the number one digit at a time…

Another example comes from a VUI in which several steps are required in order to provide the caller with driving directions. At one point in the sequence, we receive some helpful orientation, in (65).
We will now collect your address.
Though this prompt is intended to be heard over the telephone, it is representative of written discourse. First, we hear *will* instead of *going to*, even though planned intention is a better metaphor in this context than spontaneous volition. Again, we find the placement of “now” between the auxiliary (*will*) and the main verb (*collect*). Finally, the use of *collect* in the sense of “elicit information” is unnecessarily technical. This prompt may be recast as something like this:

icon

Now I’m going to ask you a few questions to find out where you’re going.

This is longer than the original, but it is easier to listen to, since it avoids those discourse features that are alien to spoken language and instead exploits familiar linguistic conventions.

“Romans perspire, Anglo-Saxons sweat”
Above it was noted that written language often favors certain words, phrases, and grammatical structures that are avoided in speech. In this section, we will focus on spoken versus written vocabulary.

The title of this section, “Romans perspire, Anglo-Saxons sweat,” alludes to the richness of the vocabulary of the English language. Owing to our dual inheritance of words from the Romance and Germanic language families, we often have recourse to two words that are roughly synonymous, as in the case of verbs like *sweat/perspire, put out/extinguish, leave/exit,* nouns like *fish tank/aquarium, crowd/multitude, lunch room/cafeteria, drink/beverage,* adjectives like *kingly/regal, funny/humorous, friendly/amicable, fat/corpulent,* and adverbs like *yearly/annually, et al.* In terms of usage, however, it is usually the case that Latin-derived alternatives (like *perspire* and *extinguish*) have a more formal, technical ring to them, while their Anglo-Saxon counterparts sound more informal and colloquial. When we write carefully, we gravitate towards more formal (Latin-derived) alternatives, given the relative permanence or persistence of written messages. The writing of prompts, which are intended to be spoken, should avoid this tendency. The list of verbs in (67), on the left, was culled from a review of several speech and touchtone applications, with more colloquial suggestions provided in (68), on the right.

<table>
<thead>
<tr>
<th>(67)</th>
<th>(68)</th>
</tr>
</thead>
<tbody>
<tr>
<td>acquire</td>
<td>get</td>
</tr>
<tr>
<td>activate an account</td>
<td>set up an account</td>
</tr>
<tr>
<td>create an account</td>
<td>set up an account</td>
</tr>
<tr>
<td>create a bookmark</td>
<td>make, add a bookmark</td>
</tr>
<tr>
<td>encounter [information]</td>
<td>find, come across</td>
</tr>
<tr>
<td>encounter [difficulty]</td>
<td>have problems</td>
</tr>
<tr>
<td>exit list</td>
<td>be done with a list</td>
</tr>
<tr>
<td>experience difficulties</td>
<td>have problems</td>
</tr>
<tr>
<td>obtain</td>
<td>get</td>
</tr>
<tr>
<td>pause</td>
<td>take a break</td>
</tr>
<tr>
<td>provide</td>
<td>give</td>
</tr>
<tr>
<td>receive</td>
<td>get</td>
</tr>
<tr>
<td>request, collect</td>
<td>ask for</td>
</tr>
<tr>
<td>respond</td>
<td>answer</td>
</tr>
<tr>
<td>return</td>
<td>go back</td>
</tr>
<tr>
<td>select</td>
<td>choose, pick</td>
</tr>
<tr>
<td>terminate</td>
<td>end, finish</td>
</tr>
</tbody>
</table>
“Register” has to do with the level of formality of a piece of discourse. Compare, for example, (69)-(71).

(69) To whom do you wish to speak?
(70) Who would you like to speak to?
(71) Who do you wanna talk to?

(69) is the most formal and (71) is the least. The specific linguistic features that differentiate the register of these questions are:

- Formal *To whom...?* vs. informal *Who...to?*
- Formal *Wish* vs. less formal *would like vs. want to*, pronounced colloquially as *wanna* *Speak* vs. more informal *talk* to mean “have a conversation”

For Halliday (1994), register is a multi-dimensional construct, which actually consists of three components.

- **Mode** refers to the channel of communication, e.g. written vs. spoken, in person vs. remote. This parameter is invariable in VUIs.

- **Field** has to do with the content of the discourse, as well as the social setting in which the language is being used. Field is often reflected in word choice.

- **Tenor** involves the roles and relationships of the participants. In other words, who is talking to whom? Imagine, for example, that the caller has just heard “Sorry, but there’s a problem with that passcode,” and is then asked to visit the registration website. In (72)-(75), we have four wording possibilities. Although their function in context is synonymous, they nonetheless differ in tenor.

icon

(72) You must visit the registration website at phone dot ACME Widget dot com.
(73) Please visit the registration website at phone dot ACME Widget dot com.
(74) ...but why don’t you visit the registration website at phone dot ACME Widget dot com?
(75) You might want to visit the registration website at phone dot ACME Widget dot com.

Because (74) and (75) are less direct as requests, they imply a social relation between the speaker and listener different from that of (72) and (73); the speaker in (74) and (75) is more polite and more deferential. Further, these wordings illustrate an important linguistic concept – the literal, superficial meaning of an utterance is often different from the meaning it conveys in context. Only literally is (74) a question, and only literally does (75) assert the possibility of a desire (“might want”). In actuality, these are *socially intelligent* ways of telling someone to go do something. In any case, the tenor of messages in an application

---

5 As we have seen, the use of social *must* in (72) is atypical of spoken English, though popular in touchtone and VUI applications.
depends largely on the persona chosen for that application, and the persona’s intended social role vis-à-vis the user.

Mode, field and tenor must all be considered when writing prompts.

Whatever you decide in terms of the most appropriate register for your application, make sure it’s exercised consistently throughout your dialog. Inconsistent register riddles the first draft of an otherwise impeccable dialog design for a money transfer demo, (76)-(79):

(76) What account do you want to transfer from?
(77) What account do you wish to transfer to?
(78) I’m sorry, I didn’t understand. Say the name of the account you wish to transfer money to. For example, you could say ‘savings account’.
(79) You wish to transfer five hundred dollars from your savings account to your money market account. Is this correct?

These prompts transmit social cues that clash. This particular use of wish, in (77)-(79) is literary, and in speech it conveys social distance and impersonality. In addition, (79) reflects formal, written discourse in its use of this as a back-pointer. On the other hand, sentence-final prepositions in (76)-(78) are conversational. Example (78) uses the formal, distant, impersonal-sounding wish, but the use of “could” (in “you could say ‘savings account’”), to make a request is relatively personal, indirect, and gentle. (Compare, for example, “You must see a doctor” with “You could see a doctor”. Dissonant social cues in these prompts undermine the establishment of a single voice belonging to a coherent personality.

When reviewing dialog specifications, we occasionally get feedback about the use of prepositions at the end of sentences. Consider the following:

(80) Auto-Attendant: “Who would you like to speak to?”
(81) Traffic: “What ‘hot spot’ do you want a report for?”
(82) Banking: “What account are you transferring from?”

Because these are thought to exemplify “bad grammar,” they are then “corrected,” as in (83)-(85).

(83) Please say the name of the person to whom you’d like to speak.
(84) Please tell me the “hot spot” for which you would like a traffic report.
(85) Please say from which account you would like to transfer funds.

In actuality, the use of who instead of whom (80) and the sentence-final placement of prepositions (80)-(82) is not only acceptable in conversation, it is characteristic of conversation.

---

6 There is a conversational use of wish, but it expresses “contrary-to-fact” desire, as in “I wish it would stop raining,” or “I wish I had a million dollars.”
Jargon

“Jargon” is defined as “the technical or specialized language of a trade, profession, or similar group.” As is the case with most other professional groups, there is a fair amount of jargon used by speech technologists, a group that includes dialog designers, linguists, software engineers, and other speech scientists. Jargon is acceptable if the speaker is communicating with others of the same “trade, profession, or similar group.” Otherwise, the use of jargon can induce reactions ranging from non-comprehension to mild alienation.

In a demonstration of a VUI designed to furnish driving directions, we happened to hear the prompt in (86), though the jargon-free wording in (87) would have been just as effective.

(86) An error has been generated. Returning to Main.

icon

(87) Sorry, there was a technical problem, so we’ll have to go back to the main menu.

The wording of the first part of (86) reflects the technical concept of “error generation.” In the second part of the prompt, “main” is a truncation of “main menu.” This truncation is frequent in the conversation of dialog designers, who may easily overlook the association between “Returning to Main” and homophonous “Returning to Maine.” Despite the disambiguating role of context, this humorous connection comes more readily to listeners who are unfamiliar with dialog design.

In a dialog specification document for an application intended for airline employees, we encountered the message in (88). In this context, however, a more appropriate, user-friendly message would be as in (89).

(88) No travelers are defined for this employee.
(89) Sorry, but it looks like the employee hasn’t designated anyone for flight benefits.

The element of jargon is this particular use of “define,” which is usual for programmers and database specialists, but unusual for typical users of this travel industry VUI.

Note, however, that jargon in itself is not always undesirable in a VUI. In this same interface, an example of appropriate jargon is the frequent use of the verb phrase “list for (a flight),” as in “Do you want to list for this flight?” When an airline employee “lists for a flight,” he or she is making an unconfirmed booking. This is different from “waitlisting,” since employees can “list” for a flight that has plenty of availability. In any case, “list” is appropriate here because it is the jargon of airline employees, in other words, jargon that the users themselves will expect to hear and want to use in the interface.

The Cooperative Principle

We have been looking at how to word prompts so that they sound natural in context. In this section, we will turn our attention to the natural interpretations that listeners make when they hear messages in context.
Conversation is more than just people exchanging information. When we participate in a conversation, we bring shared assumptions and expectations regarding the topic, how the conversation should develop, the quality as well as quantity of the contribution that each participant is expected to make, and so on. Conversational participants share common principles that enable them to interpret each other’s utterances. Consider the examples in (90) and (91)

(90) A: I saw Mark having dinner with a woman last night.
(91) A: What a day! I need a drink.
B: Have you ever been to the Eagle?

Assuming that we all know that Mark is a married man, the sentence in (90) means that Mark was dining with a woman who was not his wife, even though his wife is of course a woman. Presumably, if Mark had been dining with his wife, the speaker would have said so. In (91), if speaker A acts on the assumption that B’s reply is relevant, A will assume that B’s question is intended to function as a recommendation to go to a place called the Eagle. Again, in the spirit of cooperation, A can further interpret that the Eagle is nearby, is open, and is a place that serves drinks. Imagine if one of these conditions were not true. If, for instance, the Eagle were hundreds of miles away, B’s conversational behavior would indeed be inappropriate.

Because the meanings of messages in conversations are not tied to their literal, superficial meanings, prompt-writers must take care with the interpretations that listeners will infer in context. In example (92), the welcome message permits an inference that is later contradicted by the main menu prompt.

(92) welcome prompt:
“Hello, and welcome to the Frequent Buyer Rewards Line. You can now redeem points on-line, at ‘www dot frequentbuyer dot com, forward slash Points’…”

main menu prompt:
“How can I help you? <pause> You can say ‘buy points’, ‘redeem points’, ‘transfer points’, etc”

The wording of the welcome message allows callers to infer that they must go on-line in order to redeem points. In other words, it seems as though this feature is not supported by the speech interface - otherwise, why would the VUI’s persona be providing us with the web address? Despite this legitimate interpretation, “redeem points” turns up among the main menu options in the very next prompt.

A corollary of the Cooperative Principle requires that speakers be informative. Conversation analysts refer to this principle as the Maxim of Quantity (Grice, 1975), which holds that a speaker’s contribution to a conversation is as informative as is required to advance the perceived purpose of the conversation. This principle is sometimes violated in speech applications. In (93), we have an example from a voicemail system. The context is as follows: the subscriber has just logged in with a passcode and is receiving his or her new message count.

(93) The following...{two, three, four...}...new messages have not been heard.
Assuming that the system’s contribution is informative, the caller has no choice but to infer that sometimes new messages can already have been heard. The message strikes a chord of dissonance because, for most people, a “new message” is one that has not been heard yet. After using this voicemail system for a while, you may (or, sadly, may not) realize that a “new message” is intended to refer either to a message that you have never heard or to a message that you have heard, even if in part, but have not yet saved or deleted (e.g. when you skip a message). A more straightforward version of the message would be simply: “You have {two, three, four...} new messages”. Ascribing new and counterintuitive definitions to commonplace vocabulary can impede comprehension of the message and adversely affect usability of the VUI.

From these examples involving The Cooperative Principle and corollaries like the Maxim of Quantity, we see that conversations are more than literal messages designed to exchange information. For Richards (1980), conversation “consists of exchanges which are initiated and interpreted according to intuitively understood and socially acquired rules and norms of conversational cooperation, which can in turn be manipulated to create a wide range of meanings beyond the level expressed directly by the utterances in the conversation themselves.” In this chapter, we have argued that interactions with speech applications must not be treated as exceptions in this regard.

---

**Conclusion**

The main idea of this chapter has been that we approach all verbal tasks with certain expectations having to do with the linguistic form of utterances in context. Speech interfaces are not exceptional in this regard and should be regarded by the design team as conversations. Dialog designers and prompt writers can and should take advantage of the fact that everyday conversation is the communication system most familiar to our users.

We occasionally encounter some resistance to conversational prompt design on the grounds that spoken language is purported to be inherently inferior, careless, or too relaxed for the customers of Company X. In these cases, prompts that lack contractions or everyday cohesion devices are endorsed under the rubric of “better writing.” As we have seen in this chapter, however, conversational language is a sophisticated system of its own, having developed for reasons of its own. Better prompting is not a question of “better writing,” in the traditional sense. Unlike written language, spoken language has evolved from prototypically face-to-face interactions over the course of millions of years. Spoken language incorporates features and principles that are specially adapted for this mode of communication and that we should observe and exploit in prompting.

In order to leverage users’ familiarity with spoken language, we have made the following recommendations, based on observations of the form of naturally occurring language:

- **Attend to the differences of spoken vs. written language.** They have evolved differently owing to the nature of the signal (e.g., ephemeral vs. persistent). What people are used to hearing is not what they are used to reading.
- **Use cohesion devices** (like pronouns and discourse markers) appropriately. They enhance the functional unity of a dialog, reinforcing relationships of meaning between messages. Listeners rely on them for comprehension.
- **Observe and exploit the principles of information structure.** They dictate where listeners naturally expect to retrieve contextually determined old vs. new information.
• **Register** is a sociolinguistic phenomenon that will reflect directly on your brand and corporate image. In particular, consider **mode**, **field**, and **tenor**, which are essential ingredients of persona design. Inconsistency in any of these areas can leave users with an unfavorable impression. Stick with a style that befits the content and the desired social role of the VUI persona vis-à-vis the user.

• **Jargon** is another sociolinguistic phenomenon that at times should be avoided, at other times exploited. Appropriate use of jargon begins with knowing who your user is.

• The **Cooperative Principle** accounts for how and why language users “read” between the lines. Consider the possible inferences that users will draw from the messages you compose. Be informative, but not in excess, and anticipate how users will be informative in responding to prompts.

Successful prompt writing depends on a view of the VUI as much more than a mechanism for collecting information via acoustic signals. A user-centric speech application is really a **language** interface, where “language” implies both a cognitive dimension (e.g., use of cohesion devices, principles of information structure, and pointer-word directionality) and a social dimension (e.g., jargon, vocabulary and grammar choices, and register). VUI designers who are mindful of these dimensions of linguistic experience are well on their way to creating a VUI that will engage users and project a favorable corporate image or persona.

Another recurring theme of this chapter has been the paramount importance of considering the form of prompts **in context**. A prompt is not a simple, isolated translation of a recognition return, such as `<action GET_QUOTE>`, or “Too Much Speech”. Every prompt is like a small piece of a tile in a mosaic, in that the whole is perceived as greater than the sum of the parts. When writing prompts, always consider the context of the dialog, since the structure of naturally occurring language is itself context-sensitive. Context-sensitivity is what listeners expect, whether they are participating in a human-to-human conversation or a dialog that is engineered.

Conversational language is one of Mother Nature’s greatest masterpieces. Prompt-writers and dialog designers should regard it as a source of inspiration for user-centered design, tempered with an awareness of the limits of the current technology.