Int J Speech Technol
DOI 10.1007/s10772-008-9010-8

Sequence Package Analysis: a new natural language understanding method for improving human response in critical systems

Amy Neustein

Received: 9 October 2008 / Accepted: 15 October 2008

Abstract This paper will demonstrate how Sequence Package Analysis, as a new natural language understanding method that is built on a set of parsing structures that consist of context-free grammatical units and related prosodic features for identifying affective/emotional data found in natural speech and blogs, may better accommodate the goals of crisis management and rapid decision making in critical systems. Following an in depth discussion of the genesis and development of this method for the design of voice-user interfaces and audio mining programs, debuted in an earlier issue of *IJST*, this paper will attempt to show how Sequence Package Analysis can improve human response in monitoring recorded conversations of terror suspects and the recordings of help-line desks. In both instances, effective human intervention may help avert a crisis and resulting liability. The paper's limited focus on these two respective domains does not, however, limit the applicability of Sequence Package Analysis to other critical systems, inasmuch as the parsing structures explained below are generic enough to be applied to other critical systems, such as ambulance control, aircraft or nuclear power stations, or 911 calls for help. Given that critical systems require effective human response from decision makers, natural language data must be accorded the same kind of scientific scrutiny given to graphic design and other features of human-computer interaction.

Keywords Critical systems · Natural language understanding · Sequence package analysis · National security · Help-line desks · Human-computer interaction

1 Introduction

Critical systems deserve nothing less than proficient and high-performing speech interfaces to pierce through the labyrinth of natural language dialog, which is characteristically punctuated by ambiguities, obscurities, repetitions and ellipses—of which, individually and collectively, impede Human-Computer Interaction. True, speech system developers have demonstrated remarkable progress in better understanding human dialog, but speech recognition difficulties still plague most voice user interfaces and audio (and text) mining programs, even for those that employ advanced and well thought out natural language design.

Speech Technology recently took a look at some of the practical limitations of NLP (natural language processing) as opposed to the sanguine goals of speech system developers, whose vision is to make natural

A. Neustein (✉)
Linguistic Technology Systems, 800 Palisade Avenue Suite 1809, Fort Lee, NJ 07024, USA
e-mail: lingtec@banet.net

language an indispensable feature of call centers. With fewer human agents available to handle the large volume of calls into customer-care contact centers, voice-user interfaces that are equipped with natural language understanding may be used in lieu of human agents as a cost-effective alternative. Natural language is equally important in data mining, voice search and handling directory assistance calls.

In the magazine's April 2008 cover story, writer (and managing editor) Leonard Klie meticulously surveyed the field of speech system design, from its inception several decades ago till today, after talking with some of the top industry design experts and eminent scholars in the field of natural language processing (Klie 2008). Luis Valles, principal, cofounder, and chief scientist at GyrusLogic, was quoted comparing the design of speech interfaces to a "chess game" explaining that it is "so difficult to anticipate all the possible things that someone might say or do." In the same article, Juan Gilbert, associate professor of computer science and software engineering at Alabama's Auburn University, echoed Valles' caution: "If you look at the scope of the English language, for example, there are a lot of words and a lot of different ways to say the same thing… To be effective, a speech processor must be able to recognize them all and take the appropriate action based on the parameters given." The game is apparently worth the candle: renowned speech design expert Roberto Pieraccini, who serves as chief technology officer at New York-based SpeechCycle, astutely pointed out to Klie that natural language-driven speech systems (as distinguished from directed dialog speech interfaces in which users are asked to choose from a delimited and circumscribed list of menu items) become indispensable to businesses when customers become "confused about their answers to a basic prompt" or when system designers "cannot express [their] menu choices any other way [than natural language]."

Critical systems are no exception. During a crisis, the anxious user has no time to wade through discrete voice prompts and menu options, which can be all the more confusing in a high stress environment, causing the user to err in his/her selections and prompting a never-ending return to the main menu. The human operator is likewise met with constraints that make menu choices too cumbersome, given that in a crisis—when there is a very limited window of response time—the system cannot afford to wait for the user to navigate through the menu-driven interface, even assuming that the user will seamlessly navigate the system and not be sent back to the main menu to begin his/her selection all over again. To meet such practical concerns, both for the caller and for the human operator standing by, it is important that speech interfaces used in critical systems, and especially in time-critical systems, be competently designed to process natural language input rather than rely solely on directed dialog.

This paper will demonstrate how Sequence Package Analysis, as a new natural language understanding method, may better accommodate the goals of crisis management and rapid decision making which is integral to the sound functioning of critical systems. The author will demonstrate—following an in depth discussion of the genesis and development of this new method (drawing comparisons and contrasts with existing approaches to natural language understanding and the building of speech interfaces)—how Sequence Package Analysis may improve human response in critical systems by focusing on two specific areas of application: 1) Intelligent monitoring of the recorded conversations of suspected terrorists to detect threats to national security; and 2) Intelligent monitoring of call center recordings for disturbing signs of customer anger and frustration that may portend serious financial liability for the enterprise (such as considerably increased risk to customer retention, or even the possibility of a lawsuit over unresolved customer grievances). In both instances, effective human intervention may avert a crisis and its associated consequences.

The paper's limited focus on these two respective domains does not, however, limit the applicability of Sequence Package Analysis to critical systems. That is, the intricately designed table of parsing structures is generic enough to permit Sequence Package Analysis to be readily applied to the myriad domains of critical systems, such as aircraft or nuclear power station control systems, ambulance control, or 911 calls for help.

2 Background

The seminal article on Sequence Package Analysis, or SPA, appeared in the *International Journal of Speech Technology*, in which Neustein discussed how SPA can enable a speech system to detect, among other things,

the wide range of speakers' emotions found in doctor-patient recorded dialog and customer-care help-line calls—accomplishing this by relying on the *entire* sequence package, a series of related speaking turns and parts of turns, rather than on individual words or combinations of words and their attendant stress patterns (Neustein 2001a). In that article, Neustein described a variety of methodological approaches to tagging parts of speech to perform analyses of utterances, pointing out the unique methodological underpinnings of SPA when compared with other natural language understanding methods. In subsequent articles, appearing in the peer-reviewed literature and industry publications, Neustein demonstrated how SPA's unique brand of parsing structures—consisting of context-free grammatical units with notations for related prosodic features, such as elevated inflection or pitch or changes in rate of speech and their attendant intra- and inter-utterance pauses—are designed to reflect, in addition to syntax, some of the more complex and emotionally charged *semantic* aspects of communication (Neustein 2001b, 2002a, 2002b, 2004a, 2006, 2007a, 2007b). Over time, Neustein's novel research and insights on natural language understanding caught both the attention of AI researchers and speech system architects.

For example, Paprzyki, et al., members of the AI community, have referred to SPA as among the more advanced parsing methods for "captioning the text to which data mining is applied" so as to better detect the subtleties of human emotions (Paprzyki et al. 2004). Inventor Jeffrey A. Gallino of CallMiner, the leading audio search company, referenced Neustein's publications on SPA in its recently-approved patent application for an automated voice search technology, which was based in part on a speech processor that carefully divides the body of audio data into a plurality of segments (Gallino 2008). All in all, whether SPA has fueled the excogitative reflections of AI scholars or more likely buttressed the practical concerns of speech system designers, this new natural language understanding method offers an enhanced way of detecting the wide range of speakers' emotions too often obscured in user dialog. Such an enhanced analysis of speaker emotion yields undeniable benefits for critical systems, inasmuch as critical systems have come to rely heavily on an accurate grasp of speakers' emotions to perform competent crisis management and rapid decision-making in high-stress environments where emergency conditions necessitate a skilled and unerring human response.

More recently, Neustein has applied SPA to the analysis of blogs, given that blogs closely resemble the informality of conversational dialog, as opposed to the structured format of written text. Blogs are often subject to the same vagaries of natural language that present themselves to spoken dialog systems. Neustein has demonstrated how blogs may be just as likely as spoken dialog to benefit from the application of advanced NLU methods (Neustein 2007a, 2007b). Critical systems cannot ignore blogs, either. Blogs constitute a rapidly expanding medium of communication, with tens of thousands of blogs created each day, according to web archivists Aschenbrenner and Miksch (2005). Blog experts, such as Technorati's founder and CEO, David Sifry, estimate that the blogosphere will double in size once every six months (Sifry 2006).

Blogs are not only a popular medium; many in the news industry admit to using them to get leads on stories (in the world of news reporting, this has even been given a name: "consumer generated news"). It is therefore not unlikely that the very early warning signs of public safety hazards (e.g., contamination of food or water, or patients' adverse reactions to new and popular pharmaceutical products), or even threats to national security, may be gleaned at least partly from the rants of blog postings. Equipping critical systems with better NLU methods to be able to wade through the blogosphere's sea of verbiage—characteristically encumbered by the liberal use of hyperbole, digressions and irrelevancies—in order to detect the early signs of public safety hazards or national security threats may indeed improve human response in *time-critical systems*, with an overall benefit to crisis management and rapid decision making.

Understandably, as explained at the beginning of this article, speech interfaces face the fundamental challenge of trying to understand what the user is attempting to say purely for its denotative content: "what does the user mean when he uses certain words and phrases?" But upon adding the component of *emotion detection* in voice recordings or *sentiment* (or tonal) analysis of blog postings, the user interface is indisputably presented with a much more difficult task. For critical systems, emotional content, notwithstanding its presentation of an enormous challenge to the speech or text analytic program, cannot be discarded or ignored; failure to recognize the stridency and/or urgency in the speaker's natural language input, or similar emotional content in blog messages, may lead to disasters.

Speech analysts Yan Qu, James Shanahan, and Janyce Wiebe recognized the importance of the affective component in the design of speech systems, looking carefully at automatic tagging of affect, affect-based text mining, developing a semantic lexicon for emotions and feelings, among the many other design considerations for building affect-based mining programs (Qu et al. 2004). Yet, in spite of such developments, understanding emotion in dialog still perplexes many researchers and system designers. Most will agree that newer and more advanced natural language understanding methods are very much needed to process emotional data effectively.

For example, speech analytic programs that mine natural language dialog for signs of distress, frustration, anger, and a host of other human emotions are still of very limited effectiveness. Since they simply match the speaker's natural language input against the program's list of keywords (that is, relying solely on word-spotting, and/or its supporting technology that automatically eliminates meaningless words, phrases and pauses to identify and separate the most basic sentence components: subject, verb, object), they cannot process speech as it actually occurs. If a speaker fails to use the word(s) found in the speech application's vocabulary, or its closest probabilistic equivalent, the system is "stumped," and a poor statistical word match or no match is given for the natural language entry. These same limitations apply to programs that look for changes in prosody, such as a sudden elevation of inflection and/or pitch, insofar as prosodic patterns, like lexical entries, vary across populations of speakers.

Some of the more recent mining programs claim to control for the variations in speaker prosody simply by taking into account that among angry callers, for instance, there are some who may actually lower their tone/pitch and speak more slowly, instead of raising their voices and accelerating their rate of speech (Britt 2007). However, such methods can be misleading; they can reduce interactive dialog to simplistic metrics that are woefully inadequate to the needs of speech analysis. For example, when a speaker lowers his pitch and decelerates his rate of speech he might in fact be trying to stimulate a more empathic reaction from the agent, achieved by slowing down and reducing the shrill of the complaint. Such a change in prosody would be anything but a display of anger, since the caller shows he is at pains to establish a conciliatory interaction with the customer service agent, behavior that would ostensibly not occur in the irate caller who needs to "blow off" steam.

Such shortcomings in speech analytic programs can have significant consequences. When specific keywords or changes in prosody, for that matter, are not found, the speech system can readily overlook important affective data that are critical to accurate assessments of the attitude and affect of the speaker; or, conversely, when keywords and prosodic changes give the specious appearance of caller frustration/anger, mining results can be seriously skewed. Furthermore, one must not underestimate the ripple effect of such limitations on language translation programs, which require in addition to a high word-recognition accuracy rate the correct reading of speakers' emotions and intentions to perform proper translation of natural language dialog. (Otherwise, sarcasm and irony may yield an embarrassingly mistaken literal translation in another language.)

SPA offers critical systems a new methodological approach to help surmount the inherent difficulties in speech (and text) analytic programs, such as those described above. It relies more on the sequence package in its entirety, as the *primary* unit of analysis, than on isolated syntactic parts, such as subject, verb, object. By parsing dialog for its relevant sequence packages, the SPA designed natural language interface extracts important data, including emotional content, by looking at the timing, frequency and arrangement of the *totality* of the context-free grammatical components that make up each sequence package. And since natural speech consists more of a *blend* of sequences folding into one another than a string of isolated keywords or phrases, one can plausibly argue that speech applications and text analytic mining programs equipped with SPA can better accommodate how people really talk. This then makes it possible for human response in critical systems to dovetail with the true emotional state of the user, which in turn means better crisis management and decision-making. In high stress environments, understanding human emotion (e.g., urgency, distress, agitation, fear) can make the difference between failed crisis intervention—resulting in unrecoverable business loss, personal injury or health problems, or damage to entire communities through terror attacks or public safety hazards—and a salutary outcome.

3 Methodology

The vagaries of natural speech undoubtedly become more intensified in an emotionally charged, crisis-ridden environment. SPA adjusts to this by offering a set of algorithms that can work with, rather than be hindered by, ambiguities, repetitions, ellipses, and the all too common substitution in high stress environments of enigmatic idioms, elusive metaphors, colloquialisms, shibboleths, etc., for just plain words and phrases. Ironically, SPA mines conversations and blogs to find the very sort of dialog and blog data that would have been discarded (or simply ignored) by most speech and text analytic systems as meaningless diatribes, unwieldy talk, or talk that is far too amorphous to grasp. And while some of these discarded data (such as the heightened occurrence of inter-sentential clausal connectives, multiple use of anaphors, idioms and metaphors, or deviations from normal variations in inter- and intra-utterance spacing) might appear relatively unimportant to speech and text analytic programs, these data can be very significant in properly interpreting the emotional content found in natural language dialog and blogs.

Using SPA, the author has designed a BNF (Backus-Naur Form) table consisting of 60 Sequence Packages—a typology of parsing structures representing the *semantic* aspects of communication—that capture the affective data found in natural speech and blogs. The parsing structures contained in each Sequence Package consist of a set of non-terminals—context-free grammatical units and their related prosodic features—for which there is a corresponding list of *interchangeable* terminals: words, phrases, or a whole utterance.

The SPA-based BNF table that is used to capture speakers' affect and other semantic aspects of communication—like the BNF tables that are widely used to denote *syntactic* parts of natural language grammars—consists of an elaborate formulation of parsing structures, providing for the incremental design of complex grammatical structures from their more elemental units. Many of the subtleties, convolutions and complexities of human emotion can be more effectively represented by such multi-tiered grammatical structures. A "very angry complaint," for example, could be illustrated on the SPA-designed BNF table as the natural accretion of more elemental parsing features, such as assertions, exaggerations and declarations, so as to effectively notate these semantic aspects of communication—aspects that have all too often eluded conventional spoken language systems.

It is no easy task to formally map out the conversational sequence patterns of natural language dialog and blogs that reflect such elusive, sometimes confounding, human emotions. To do this, SPA draws its methodological basis from the field of conversation analysis.

Conversation analysis provides a rigorous, empirically-based method of recording and transcribing verbal interactions by using highly refined transcription signals to identify both verbal components and paralinguistic features, such as stress, pauses, gaps, overlaps and changes in intra-utterance spacing, as demonstrated by veteran conversation analysts, Atkinson and Heritage (1984).

Over 35 years of study of interactive dialog, conversation analysts have been able to identify and describe how participants in a dialog systematically accomplish their interactive work, while they are continually engaged in the process of making sense of the ongoing social activity. This is accomplished by breaking down natural language communication into its elemental form of conversational sequences and speaking turns within those sequences, rather than isolated sentences or utterances. Conversation analysts examine how speakers demonstrate, through the design of their speaking turns, their understanding and interpretation of each other's social actions, including the wide range of emotions embedded within those actions, such as a speaker's noticeable failure to answer a question directed at him or her, which in certain instances may indicate annoyance or irritation with the other speaker, rather than failure to hear the question.

Reduced to algorithms, many sequence packages are naturally transferable from one contextual domain to another, which means that many of the same sequence package structures found in the conversations of doctors and patients, or in patients' blogs, also appear in call center dialog between distressed customers and call center agents. In addition, by focusing on social action, rather than on grammatical discourse structure solely, this new NLU method for mining conversations and blogs can potentially be applied to a myriad of other languages, including Arabic and Farsi, because "*all* forms of interactive dialog, regardless of their underlying grammatical discourse structures,

are ultimately defined by their *social* architecture," as pointed out by Neustein (2004b).

4 Design

There are two ways that an SPA-driven speech interfaces or voice (or text) analytic mining program can work. First, SPA can serve as an "add on" layer for voice user interfaces or conventional data mining programs, including those built on vector-based models, which assign n-grams and bi-grams and hold spaces in between words and word phrases accordingly. If SPA functions as an "add on" layer, the "global weighting" to be applied for the next layer of analysis need no longer be limited to content words or their term roots; rather, it can now also encompass sequence package material. To accomplish this, SPA uses Statistical Language Modeling (SLM)—the standardized method for matching speech input to the speech application vocabularies—but instead of generating candidate words and word phrases for the speech input, SPA generates candidate *sequence packages*. Thus, using the same method of weighting possibilities used for candidate words and word phrases, SPA detects the range of possible sequence packages present at each stage of the conversational sequence, the totality of which makes up the dialog, as discussed by Neustein at the 3rd International Workshop on Natural Language Understanding and Cognitive Sciences (Neustein 2006).

As an "add on" layer, SPA can take the output of a speech engine and provide a deeper level of analysis of the patient's dialog with the physician (or healthcare worker), the customer's interactions with the call center agent, or the terror suspect's conversations with cohorts, by interpolating sequence package information into the engine's output stream. By marking sequence package boundaries and specifying package properties, the SPA-enhanced mining program gives the software downstream the contextual indicia—the precise location points in the flow of interactive dialog, signifying the different conversational activities and phases of the dialog—needed to interpret the rest of the data stream reliably.

Since much of blog material, likewise, consists of different phases of communication, including but not limited to the straightforward narrative portion of the complaint, digressions and/or complaint resolution, it is essential that a text analytic program, besides noting the descriptors used by the blogger, also provide contextual indicia for analyzing the tonal quality of blogs. For example, since the more strident tonal features are most likely to occur in the digressive phase of the blog, as opposed to the narrative complaint or the complaint resolution phase, a program designed to identify the different phases of blogs will likely give less credence to product criticism occurring in the digressive phase, where diatribes and rants are to be expected. Thus, failure to isolate the various phases of the blog can quite plausibly skew sentiment analysis, by ascribing undeserved importance to the descriptors found in the digressive portion of the blog.

SPA might also be used as a wholly integrated system rather than as an "add on" layer to conventional speech and text analytic programs. In such a case, such programs would use as their starting point sequence package grammars, represented by the specially designed BNF table of parsing structures, rather than words and word phrases. Such a use would allow the building of an entire vocabulary, by methodically uncovering the keywords characteristically embedded within these sequence package templates without necessarily having an *a priori* knowledge of the words and word phrases in the speech application.

Whether SPA is built into a system as an "add on" layer of intelligence or as a wholly integrated system, it can be argued that SPA will generally enhance the scalability of critical systems that contain speech interfaces for performing interactive voice response, data mining of recorded calls, or mining of blog messages. This is so because SPA can help to streamline the corpus of data required to build a statistical language model, by focusing on commonly occurring *sequence packages* themselves, which are seemingly more generic to a large population of speakers (and bloggers) than actual word or word phrase choices. This would eliminate the need to construct elaborate speech application vocabularies that would take apart each and every one of the user's utterances in order to discern relevant words or word phrases and to estimate the probabilities of the occurrence of various linguistic units within those phrases.

5 National security

In December 2005, *New York Times* writers Licthblau and Risen reported that officials at the National Se-

curity Agency anonymously leaked to the press that, since the September 11th attacks, "the volume of information harvested from telecommunication data and voice networks…is much larger than the White House has acknowledged" (Lichtblau and Risen 2005). Ironically, a year earlier, *Times* reporter Licthblau publicized an astounding report issued by the Justice Department's inspector general. The report revealed that "more than 120,000 hours of potentially valuable terrorism-related recordings have not yet been translated… [and] that the F.B.I. still lacked the capacity to translate all the terrorism-related material from wiretaps…" The report conceded that "the influx of new material has outpaced the Bureau's resources." Among the reasons given by the inspector general for this embarrassing backlog was the "shortage of qualified linguists and problems in the bureau's computer systems… [and] management and efficiency problems that dogged the bureau even before September 11th" (Lichtblau 2004). There is no reason to believe that these problems have been solved, despite the government's obvious determination to gather still more data.

Indeed, it should be asked whether there may be another persistent reason for the discrepancy between data collection and analysis: namely, that many government translators and linguists are skeptical about finding important *clues* to terror-related activities in recordings of conversations with terror suspects, without which effective human response and/or intervention are rendered moot. Such skepticism, after all, is at least partly justified. Most audio data mining programs that parse recordings in search of "keywords" can be stymied by speakers who deliberately avoid the use of keywords—names of persons, locations, landmarks or references to times and calendar dates—that might serve as "red flags" to anyone listening in on the call. As a result, clever terrorists can outsmart a conventional mining program that relies on word-spotting techniques in parsing recorded dialog.

Against this background, some members of the intelligence community have noted the benefit of exploring newer and more efficient data mining methods. In the wake of 9/11, the National Law Enforcement Technology Center, a special program within the National Institute of Justice's Office of Science and Technology that provides information as a service to law enforcement and forensic science practitioners, devoted part of one of its weekly newsletters to Sequence Package Analysis, as a new AI-based natural language understanding method. The publication pointed to the utility of SPA as "a new voice technology tool" to "help law enforcement better weed through wire-tapped conversations to learn of possible terrorist plots" (National Institute of Justice 2001).

SPA would enhance human response in critical systems by pointing out to the human intelligence officer or agent—either retrospectively, by dissecting recorded conversations, or in real time, by analyzing a conversation in progress—those precise portions of the terror suspects' conversations that require particularly close human analytic inspection, thus sparing the officer or agent the need to listen to entire conversations, in real time, or subsequently comb through a completed transcript of the recorded conversation, before making accurate crisis intervention decisions.

Here is a hypothetical example of a conversation between two terror suspects taking place in Brooklyn shortly after 9/11. Although the dialog is a hypothetical construction, the sequence patterns contained in the dialog example below are themselves empirically derived from close analysis of actual conversations, as shown by Sacks and Schegloff (1979).

In the example below, Speaker "A" is trying to inform Speaker "B" about an important meeting to take place at a new location, which is right at the foot of the Brooklyn Bridge. However, Speaker "A" is confronted with two difficulties: First, he must try to avoid any direct reference to Brooklyn Bridge—one of the most heavily surveilled locations for terrorist activities in the United States—because it could arouse the suspicions of an intelligence agent who might be listening in on the call or examining a subsequent recording or transcript of the call.

Second, Speaker "A" must try to maintain an air of nonchalance, refraining from making any prefatory remarks to the other speaker that could convey a sense of urgency about the imminent meeting, which would naturally stir the suspicions of an intelligence agent or officer scrutinizing the call. As part of this air of nonchalance, the speaker must also prevent any sudden and marked changes in prosody (vocal stress patterns) that could draw the attention of a third party listening in on the call, especially one who is attuned to noticeable elevations in pitch or an increase in volume, or other prosodic features associated with an imperatival, anxious or urgent tone.

Yet, in spite of these constraining conditions placed upon Speaker "A," he must try to accomplish the work

at hand of unequivocally conveying to Speaker "B" where to meet—making sure Speaker "B" definitely understands the important plans so that they won't be foiled.

Here is how Speaker "A" might accomplish this delicate task:

Speaker "A": Come to the intersection near River Cafe? (The question mark shows an upward intonation.) [0.2–0.5 second pause]
Speaker "B": [1.6 second pause]
Speaker "A": You know the busy street with the big traffic light?
Speaker "B": River Café, yeah.

Although both speakers avoid any reference to Brooklyn Bridge, as well as any reference to the importance of getting these directives straight (thus attempting to stump conventional natural language understanding programs, which are attuned to listening for keywords and phrases), the SPA-driven speech engine, which looks for more generic forms of linguistic data that conform to the templates of sequence packages, could have detected the speakers' intent to firm up their plans. To do this, the SPA speech interface would have mapped for the human agent the six-part Sequence Package described below. This was used by the interlocutors to make arrangements to meet. This example in particular calls for one to pay close attention to the spacing of inter-utterance and intra-utterance pauses, which emerge as important clues to the interactive work of the speakers in accomplishing, albeit in covert fashion, the making of logistic arrangements, while avoiding using any names that would have triggered warnings in conventional audio search engines.

Speaker "A":

(1) A noun referent ("River Cafe") with an upward intonation:
"Come to the intersection near River Cafe?"
(2) A brief pause, giving the listener the opportunity to show recognition or in the alternative, ask for clarification:
[0.2–0.5 seconds]

Speaker "B":

(3) A long pause by the listener which indicates his lack of understanding or possible confusion:
[1.6 seconds]

Speaker "A":

(4) A clarification of the noun referent ("River Cafe"):
"You know the busy street with the big traffic light"

Speaker "B":

(5) A repetition of the noun referent, which had been the source of the recognition trouble:
"River Café."
(6) A "recognition marker" immediately after the repeat of the noun referent, which had been the source of the recognition trouble:
"Yeah."

In this example, an SPA-driven mining program would have uncovered the term "River Café" upon its search of the dialog for sequence package templates that form the most likely match for the sequences found in the speech engine's output stream. Here's how:

First, the speech mining program would look for a noun referent marked by an upward intonation followed by a brief pause. Second, the program would identify the deviations from the norm in inter-utterance spacing—*i.e.*, wherever the gap between speaker "A" and speaker "B" exceeds what veteran conversation analyst Gail Jefferson called the "tolerance interval" (Jefferson 1989, p. 170) an interval "which marks the acceptable length of absent talk in conversational interaction,' as discussed by Wooffitt et al. (1989, p. 144). The consensus among conversation analysts is that silences exceeding 1.2 seconds signal trouble in the dialog. In this example, we have an inter-utterance pause lasting 1.6 seconds, which would definitely be noted by the SPA mining program.

Third, the program would look for a clarification of the noun referent that caused the recognition trouble displayed by the other speaker. Since the clarification attempt is constructed as an anaphor ("...the busy street with the big traffic light"), the program must search solely within the boundaries of the sequence package to link the anaphor correctly to its antecedent referent. In so doing, the program would locate the prior utterance that begins the sequence package. At that point in the dialog, the speaker raises his inflection when identifying a new meeting place, pausing slightly to give the other speaker the chance for feedback ("Come to the intersection near River Café?" [0.2–0.5 seconds]).

It should be noted that in this example, the program's decision to link the anaphoric expression to its

antecedent referent in the prior utterance is not governed by grammatical rules, which might dictate the linking of anaphors to their antecedent referents in the immediately preceding sentence. Sequence package configurations work differently, recognizing the patterned regularities of talk as a socially organized activity. In light of such regularities, anaphoric connectivity may in fact deviate from strict grammatical rules—as in the case of an enraged speaker who fails to identify the subject or object of his ire until *after* several speaking turns of "venting" which have been punctuated by anaphoric expressions. (Often, the subject or object is not named at all, unless the listener, in a state of exasperation or confusion, will demand that of the speaker: "who or what are you talking about??")

The last part of this sequence package template indicates that the trouble, which provoked a long silence and a subsequent clarification, has been resolved. The speaker's repetition of the noun referent that had been the source of the trouble ("River Café"), followed by a recognition marker ("Yeah"), ends the sequence and, in so doing, ends the phase of the dialog in which arrangements to meet are made.

A mining program that uses SPA to uncover critical information about suspects' activities (such as their meeting places) would now have the option of adding "River Café" to the speech application's vocabulary as an important word to look out for in the future because of its close proximity to Brooklyn Bridge. In short, an SPA mining program would work in two phases: first, it would generate candidate sequence packages for the speech input found in the speech engine's downstream; second, it would extract from these sequences packages "new" references to persons or places so that they can be properly added to the speech application vocabulary. In this way, one can empirically design an application vocabulary that better matches the reference terms (names and locations) that suspects actually use, when discussing terrorism-related activities, than a vocabulary that is derived from a list of "keywords" that one thinks they will use.

The six-part sequence package analyzed above consists of a concatenation of utterance components that are *commonly* found in dialog, whether or not the conversation revolves around the activities of terror suspects. A mining program can expect to see this linguistic pattern with some degree of predictability when one speaker in the course of making arrangements introduces a new term (such as a name of a person or a place) to another speaker—and where the "uninformed" speaker seeks, for whatever reason, to minimize his "ignorance" of the new term, by allowing the conversation to continue without stopping first to "topicalize" his lack of recognition of the new term ("Oh, I had not heard of River Café before now!"). This shows that the algorithmic design of sequence packages, including those that underlie the conversational activity of "making arrangements," is generic enough to be detected not only in conversations of terror suspects but across other domains.

This illustration shows that SPA technology brings a new method of parsing dialog to data mining, one that examines conversation for its relevant sequences, consisting of clearly defined sets of sequence packages. By breaking up dialog into discrete sets of sequence packages—which often include linguistic data discarded by most mining programs, such as intra and inter utterance pauses—SPA-driven automated mining programs may help intelligence practitioners decipher the covert dialog of terror suspects, characteristically ambiguous and elliptical, resulting in more effective crisis management and decision-making. This new method of natural language understanding can enhance the mining of important information that is all too often masked by terror suspects who carefully avoid the use of names, dates, locations, among other things. Thus, it may offer intelligence agencies a constructive solution to mining suspected terrorism-related calls. This would no doubt serve as an incentive for the F.B.I. to reduce its enormous backlog of untranscribed and unanalyzed calls, knowing there is an NLU method available to them to glean critical data from the mass of recordings. This could only help to paint a more encouraging picture of homeland security, and enhance human response in those critical systems that are designed for the purpose of insuring the security of its citizenry and the public as a whole.

6 Private industry

The call center industry has long recognized that emotion detection presents one of the greatest challenges for speech analytic programs. Judith Markowitz, one of the preeminent voices and pioneers in automated speech recognition and speech understanding, examined the difficulties in monitoring the *dynamic* flow

of emotions in call center interactions, in which a relatively placid caller might suddenly become acrimonious (Markowitz 2006). A pivotal issue for a call center is the ability to know when to "*escalate*" a customer complaint call by immediately transferring the call to a supervisor, or on some occasions even to a manager. Failure to recognize signs of caller distress and to properly route the call to a supervisory agent may seriously compromise customer loyalty. Yet, as long as call center mining programs merely calculate the occurrence of specific keywords/phrases such as the number of times the caller requests a "supervisor" or asks for the "cancellation of an account," those callers whose word choices do not conform to this set of expected keywords will most likely go unnoticed, and that could be a large portion of callers.

This is so because, as part of the complexity of human emotion, behavioral responses to stressful events may be hard to predict. All too often customers, embroiled in an angry exchange with a call center agent, will not have the presence of mind or the knowledge to request a supervisor; and when they do, they are often told that there is no supervisor available. This is where a critical system, equipped with a well-performing natural language speech interface to detect the generic features of caller distress, may alert a supervisor or manager to intervene on the spot before the enterprise risks losing the customer. These features are more likely to be detected in sequence package data than in a keyword search, for reasons already explained above. Nevertheless, for the intercession of a supervisor (or manager) to be most effective, the critical system must make sure that its speech interface is neither too lax in spotting distressed calls nor too vigilant in "red flagging" calls as warranting the spontaneous intervention of supervisors or managers; the latter could turn a call center into an oppressively micro-managed environment, restricting the relaxed flow of dialog between agents and customers, which is necessary not only to resolve complaints but also to potentially turn a customer complaint call into an opportunity to market new products and service upgrades to the caller.

Below is a real example drawn from a software help-line for some of the earlier versions of the Microsoft Windows program, as discussed by Emmison (2000):

Caller: Absolutely unbelievable! What is your name?
Agent: Mr. Smith.
Caller: Well! I intend to take this much further; this is just absolutely ridiculous!

In this illustration, while there are no standard "catch" phrases or keywords to signify an irate caller, the caller's exasperation with the customer service agent can nevertheless be detected by an SPA-enhanced mining program, which begins by identifying the sequence package boundaries—and the specific properties found within the parameters of such boundaries—for anger and/or frustration. That is, by looking for specific sequence packages—an organized arrangement of specific (context-free) parsing structures, which more *broadly* reflect the semantic features of communication than a limited set of keywords—the SPA-enhanced program can better detect the caller's true emotional state.

Here's how. First, SPA would detect signs of caller anger/frustration in unusual features that show aberrations in the dialog. In this example, across several speaking turns, SPA could identify a set of "exaggerative qualifiers." The first part of the set is the opening exaggerative qualifier ("Absolutely unbelievable!"); the second part of the set is the exaggerative qualifier ("This is just absolutely ridiculous!") that closes the set. Such a set of qualifiers is *particularly* noticeable when dialog has been progressing up to this point more or less routinely, as in this example.

Second, SPA would look within this set of exaggerative qualifiers for the occurrence of any aberrations in the natural progress of dialog. In this example, an SPA-enhanced mining program would have detected a sequential "non sequitur," an utterance that is in a sequentially "displaced" position within the dialog: an interposition of a question between this set of exaggerative qualifiers: "What is your name?"

Since requests for speaker identification are generally made at the beginning of a conversation rather than halfway through a call, a caller demonstrates his/her attendance to the social organization of dialog by prefacing a non sequitur request for speaker identity with an apology or excuse ("Pardon me, but I didn't get your name when you first came on the call" or "What is your name again, I seem to have forgotten?"). It is the noticeable absence of such apologies/excuses preceding the caller's request for the agent's name midway in this dialog that raises the caller's request for the agent's name from that of a simple inquiry to that of a confrontational argument.

Third, following the non sequitur request for identity information, SPA would look for the occurrence of a declarative assertion, or a threat to take action ("Well! I intend to take this much further…") which appears prior to the second part of the exaggerative qualifier set, closing off the sequence package for caller anger/frustration.

By mapping out sequence package boundaries, as in this illustration, a speech analytic program can take what ordinarily might appear to be disorganized dialog or simply "shoot-from-the-hip" talk—common when a caller is preoccupied with "venting" his frustration at the contact care center agent—and find the indicia of caller emotion that portend serious consequences for the enterprise, such as an increased risk to customer retention. In such instances, direct and effective intervention, by permitting an improved human response to crisis management and rapid decision making, can make a significant difference for the viability of an enterprise and also avert further liability and or financial loss.

7 Conclusion

Human-Computer Interaction is gaining momentum for those engaged in the interdisciplinary study of the design, development and evaluation of user interfaces, as evidenced by the introduction of a first time track on HCI at the upcoming 24th Annual ACM Symposium on Applied Computing (March 2009). Meanwhile, critical systems have emerged as a topic of study in their own right. Engineers and researchers are increasingly aware of the urgent need to perfect the design of the user interface in order to enhance the performance of critical systems. Speech interfaces that use natural language can no longer be left out the equation. Sequence Package Analysis offers a new and advanced natural language understanding method that, as part of speech interfaces and audio and text mining programs, allows intelligence agents/officers, call center supervisors/managers, and others in a supervisory capacity to perform more successful intervention and crisis management—whether in medical triage in a disaster zone, ambulance control, nuclear power station control, national security, public safety or commercial operations of call centers. Since critical systems require effective human response from decision makers, natural language data must be given the same kind of scientific scrutiny given to graphic design, programming languages, operating systems, and the many other features essential for the well-functioning design of a user interface. In the final analysis, at the very core of human response, especially in a high-stress, crisis-ridden environment, is a penetrating and accurate understanding of human dialog. And that deserves a science of its own.

References

Aschenbrenner, A., & Miksch, S. (2005). Blog mining in a corporate environment. In *Vienna University of Technology, Institute of Software Technology and Interactive Systems and Research Studios Austria, Smart Agent Technologies* (Vienna Technical Report). Asgaard-TR.

Atkinson, J. M., & Heritage, J. (1984). Transcript notation. In J. M. Atkinson & J. Heritage (Eds.), *Structures of social action: studies in conversation analysis* (pp. ix–xvi). Cambridge: Cambridge University Press.

Britt, P. (2007). Advanced analytics offer greater precision. *Speech Technology Magazine*, *12*(4), 32–35.

Emmison, M. (2000). Calling for help, charging for support: some features of the introduction of payment as a topic in calls to a software help-line. In *Symposium on help-lines*, Aalborg, Denmark, September 8–10, 2000.

Gallino, J. A. (2008). Software for statistical analysis of speech, *CallMiner*, U.S. Patent 7,346,509, Patent and Trade Office, Washington, D.C., March 18, 2008.

Jefferson, G. (1989). Notes on a possible metric for a 'standard maximum' silence of approximately one second in conversation. In D. Roger & P. Bull (Eds.), *Conversation: an interdisciplinary perspective* (pp. 166–196). Clevedon, Philadelphia: Multilingual Matters.

Klie, L. (2008). NLP of fertile ground: but despite its promise going natural will take work. *Speech Technology*, *13*(3), 14–19.

Lichtblau, E. (2004). F.B.I. said to lag on translations of terror tapes, *New York Times*, September 28, 2004, pp. 1, 22.

Lichtblau, E., & Risen, J. (2005). Spy agency mined vast data trove, officials report, *New York Times*, December 24, 2005, pp. 1, 20.

Markowitz, J. (2006). Detecting emotion, *Speech Technology, 11*(1).

National Institute of Justice (2001). Linguistics expert predicts voice technology will play pivotal role in spotting terrorists. In *JustNet-law enforcement and corrections technology news summary*, October 18, 2001.

Neustein, A. (2001a). Using sequence package analysis to improve natural language understanding. *International Journal of Speech Technology*, *4*(1), 31–44.

Neustein, A. (2001b). Why linguistics is important for the design of a non fictional 'hal'. White paper, published *proceedings of SpeechTEK 2001*, October 24–26.

Neustein, A. (2002a). Sequence package analysis: a new data mining tool to speed up wiretap analysis. Published *proceedings of AVIOS (applied voice input/output society)*, May 10, pp. 263–267, 2002.

Neustein, A. (2002b). 'Smart' call centers: building natural language intelligence into voice-based apps. *Speech Technology*, *7*(4), 38–40.

Neustein, A. (2004a). Sequence Package Analysis: a new natural language understanding method for performing data mining of help-line calls and doctor-patient interviews. In B. Sharp (Ed.), *Proceeding of the first international workshop on natural language understanding and cognitive science, ICEIS 2004*, University of Portugal, April 13, pp. 64–74.

Neustein, A. (2004b). Sequence Package Analysis: a new global standard for processing natural language input? *Globalization Insider, X111*(1,2).

Neustein, A. (2006). Using Sequence Package Analysis as a new natural language understanding method for mining government recordings of terror suspects. In B. Sharp (Ed.), *Proceedings of the 3rd international workshop on natural language understanding and cognitive science*, *ICEIS 2006*, Paphos, Cyprus, May 23–24, pp. 101–108.

Neustein, A. (2007a). Sequence Package Analysis: a new natural language understanding method for intelligent mining of recordings of doctor-patient interviews and health-related blogs. In *Proceedings of the fourth international conference on information technology: new generations*, ITNG 07, Las Vegas, Nevada, April 2–4, pp. 441–448, 2007.

Neustein, A. (2007b). Sequence Package Analysis: a new method of intelligent mining of patient dialog, blogs and help-line calls. *Journal of Computers*, *2*(10), 45–51.

Paprzyki, M., Abraham, A., & Guo, R. (2004). Data mining approach for analyzing call center performance. In *Lecture notes in computer science*, The 17th international conference on industrial engineering applications of artificial intelligence and expert systems, Ottawa, Canada. Berlin: Springer.

Qu, Y., Shanaha, J., & Wiebe, J. (2004). Exploring attitude and affect in text: theories and applications. In *AAAI spring symposium*. Stanford University, March 22–24, 2004.

Sacks, H., & Schegloff, E. A. (1979). Two preferences in the organization of references to persons in conversations and their interactions. In G. Psathas (Ed.), *Everyday language: studies in ethnomethodology* (pp. 15–21). New York: Irvington Publishers.

Sifry, D. (2006). State of the blogosphere. *Sifry's Alerts*, August 7, 2006.

Wooffitt, R., Fraser, N. M., Gilbert, N., & McGlashan, S. (1989). *Humans, computers and wizards: analysing human (simulated) computer interaction* (pp. 166–196). London: Routledge.