The Supreme Court of the United States (SCOTUS) has been recording its public proceedings since 3 October 1955. These recordings – now in the National Archives – span nearly five decades and consist principally of oral arguments in which justices and attorneys engage in various forms of persuasion and communication between bench and bar and, obliquely, among the justices themselves. The arguments have been transcribed professionally across the entire period, creating a matchless collection of audio materials coupled with highly accurate transcripts. The audio – along with other activities captured on audio such as the announcement of opinions – offers a unique opportunity for researchers across a wide spectrum of disciplines to engage in novel and transforming research projects that were once thought beyond the reach of investigators. This proposal will create a complete archive of 6000+ hours of SCOTUS audio, provide synchronized (i.e., time-coded) transcripts of the collection, identify and tag individual speakers, build new mark-up tools for these new domains, and share the entire corpus with investigators, students, teachers and ordinary citizens. The result of this interaction among political scientists, legal scholars, linguists, and computer scientists will yield new knowledge in the modeling of multi-party discussions with complex goals, novel strategies in small group decision process analysis, and path-breaking approaches to extended collaborative commentary addressing the dynamics of human communication.

The SCOTUS archive will be maintained as a shared public resource to enhance study and understanding of the Supreme Court of the United States. Today, more than a million unique users access the partial audio collection. With a complete and updated SCOTUS archive and improved ability to query and search, the number of users should expand substantially.

On December 13, 1971, the Supreme Court of the United States (SCOTUS) heard oral argument in Roe v. Wade. Sarah Weddington, one of two attorneys representing Jane Roe, presented a laundry list of constitutional provisions that, under her analysis, warranted invalidation of the Texas anti-abortion statute and, by implication, the invalidation of similar statutes in 45 other states. Several justices seemed unpersuaded. Then Jay Floyd, representing the State of Texas, approached the bench. He began: "Mr. Chief Justice, may it please the Court: It's an old joke, but when a man argues against two beautiful ladies like this, they are going to have the last word." What the transcript does not reflect is the dead silence that followed Floyd's effort at jocularity. That silence was but one of several signals that Floyd had failed to convince the justices of his defense. Though Weddington's argument was weak, Floyd's was weaker. The case was re-argued ten months later. The advocates were far better prepared. The rest is common knowledge. The court went on to announce its most controversial opinion in a generation, one that still roils the nation. [www links embedded in red text. Requires RealOne Player.]

This oral argument and its consequences have been the subject of hundreds of articles and books. (E.g., Pacelle, n.d.; O'Brien 2000; Lazarus 1999; Epstein and Kobyalka 1992). However, this vast published literature suffers from a certain disjunction between the original oral arguments and the myriad interpretations of these arguments and subsequent opinions. In this project we will build a system that
allows students, researchers, lawyers, and the general public access to competing
and contrasting structured analyses of cases of this type that have shaped the fabric
of our democracy. Before outlining the shape of the new system we plan to build, let
us review the details of the SCOTUS corpus.

The Supreme Court has been recording its argument sessions since October 1955.
This corpus resides in the National Archives on deteriorating analog reel-to-reel tape
that has a limited life-span. Portions of the holdings have been made accessible to
the public on line. Due to budgetary constraints, the National Archives does not have
in place a systematic preservation strategy to assure the integrity of these materials
in the digital domain over the long term. Highly accurate, human-generated
transcripts exist for all 6000 hours in the proposed corpus. It should be noted that
although the transcripts are of high quality, they are never perfect and that, if we
begin to lose the original audio sources, we may lose information needed to resolve
variant interpretations based only on the transcripts.

The purpose of this ITR proposal is to complete the creation of a digital SCOTUS
audio archive and rely on its unique features to approach novel research issues in
speech and language studies, issues in SCOTUS decision-making, and issues at the
leading edge of human communication scholarship. The proposal also offers broader
benefits through open access for the general public to the entire SCOTUS corpus,
shedding public light on an institution that has remained remote from public view. A
corpus of Supreme Court oral arguments will provide an excellent resource to
address three key questions and six major research areas, each of which we will
discuss in detail below.

1. Why corpora?

Language, like the law, is not only a logical calculus, but also a complex web of
practical precedents. As a result, linguistic knowledge includes not only syntactic
recursion and semantic composition, fleshed out by a vocabulary list, but also a
network of stochastic connections among words, structures, concepts, feelings,
styles and social contexts. In face-to-face oral discourse, our sharing of both
agreements and disagreements give rise to newly emergent complex patterns of
belief structures. These complex structures and the interaction patterns that
produce them are particularly well illustrated in the oral argumentation presented at
the Supreme Court.

Over the past two decades, researchers have begun to map the geography and
ecology of this semiotic forest. Engineers have begun to learn how to build statistical
models of linguistic patterns, and how to use such models effectively in algorithms
for speech recognition, analysis of linguistic form and content, composition,
translation, information extraction and other practical applications. Psychologists
have begun to understand the development and expression of these patterns in the
human brain, and their deployment in speaking and listening. Social scientists have
begun to learn how these patterns arise in the social practice of speech communities,
and how they reflect and also create social structure.

The essential foundation of this research has been the common availability of large
volumes of linguistic data -- many terabytes of text, audio and video along with
various kinds of transcription and annotation -- as well as computer hardware and
software powerful enough to digest and analyze it. In the 1980s, these volumes of
data became available in principle to a few, and new kinds of research became
possible. In the 1990s, as such materials have become available to all, they have engendered flourishing scientific and engineering communities. General availability of large amounts of shared linguistic data provides two key benefits: it lowers the "barriers to entry" for research, and it allows replication, extension or refutation of published results. For both reasons, data publication leads to research publication and research progress.

It costs more than a hundred thousand dollars to create a million words of parsed text (because it is several person-years of skilled work). Therefore the publication of several million words of parsed text by the Penn Treebank gave hundreds of researchers access to opportunities that they could not have created for themselves, and led to many thousands of research publications in various subdisciplines of linguistics, psychology and engineering. Just as important, the fact that these researchers have been working on the same body of material has allowed them to compare results and to build on one another's work.

Similarly, recordings and transcripts of thousands of longitudinally-sampled parent-child interactions require dozens of person-years to collect, and therefore cost hundreds of thousands of dollars. The publication of such recordings and transcripts by the CHILDES Project (MacWhinney, 2000) has enabled new kinds of research in child language development, since individual researchers simply could not have undertaken the creation of such bodies of data for themselves. CHILDES data has been the foundation of thousands of research publications.

There are dozens of other examples of a similar sort, where publication of well-chosen linguistic corpora has created major new areas of scientific research and technology development. See the catalogs of the Linguistic Data Consortium, ELRA and similar organizations for examples. Typically, the influence of well-designed corpora is longer and broader than originally planned. Thus the Switchboard Corpus was designed and collected in 1990-91 to support engineering research in speaker identification, but it has been central to subsequent work on recognition of conversational speech, dialogue modeling, speech disfluencies, and many other topics. The research publications using Switchboard for these other topics now outnumber the speaker-ID publications by three orders of magnitude. As in the cases cited earlier, the Switchboard corpus cost more than a million dollars to create, but has been the foundation of so much subsequent research that its cost per resulting publication is now only a couple of hundred dollars, with the amortization still continuing more than a decade after its first publication. Other LDC corpora such as CallHome and CallFriend include free-ranging phone call dialogs between friends that have been widely used not only for the analysis of linguistic patterns, but also for analysis of higher-level discourse and content structures.

Based on existing data, the creation of the SCOTUS archive should yield similar or greater benefits. Because the transcripts are already available, the cost of creating this new archive is substantially less than the cost of building Switchboard, CallHome, or CHILDES on a per hour basis.

A. Corpus building

In 1992, a cooperative agreement grant between the Defense Advance Research Projects Agency (DARPA) and the University of Pennsylvania founded the Linguistic Data Consortium (LDC) to provide distribution infrastructure for linguistic resources. An external planning committee representing potential academic, commercial and
governmental users of LDC resource created the model that continues to govern LDC operations. Membership is open to researchers around the world. As a matter of policy, no bona fide researcher is denied access to LDC data for genuine lack of funds. Organizations join the LDC on a yearly basis where the LDC Membership Year parallels the calendar year. Members have ongoing rights to all corpora produced in the years in which they join.

During its 10 years of operation, more than 1300 organizations worldwide have used LDC data; more than 380 companies, universities and government research laboratories have joined the consortium; 920 others have licensed corpora as non-members. About half of the organizations that use LDC data are American; one third are European; Asia, the Middle East, Africa and Australia are also represented. By market segment, 76% of LDC members are non-profit organizations; 19% are commercial organizations and the remaining 5% are government research labs not only in the United States but also abroad. To date, LDC has distributed more than 15,000 corpora to its members and non-member licensees.

B. The role of LDC

LDC’s primary function is to publish and archive language resources. Each year LDC produces between 15 and 25 corpora. Outside organizations contribute about half of these asking LDC to handle final formatting, intellectual property arrangements and distribution. The other half are created by or with the help of LDC typically to support government-sponsored technology evaluation projects. At the time of writing, LDC has published more than 210 corpora covering about 35 different languages and spanning more than 850 CDs of data. These include 126 speech corpora (60%), 73 text corpora (35%) and 11 lexicons (5%).

In response to demands from its constituent research communities, LDC has expanded its role to include data collection, corpus creation, and research on the use and structure of language resources. LDC staff has grown accordingly. Twenty-nine regular staff members now manage the research, technical, collection, annotation, publication and customer service functions of LDC’s Philadelphia office. LDC also maintains a part-time workforce that varies from 15 to 45 staff members depending upon project workload.

Access to the SCOTUS corpus will continue to be available on line through The OYEZ Project. Simple and complex search and annotation features will provides users the opportunity to comment and preserve their remarks in a permanent repository. And the development of collaborative commentaries (see below) inspired by the TalkBank framework will provide yet another access point to the collection.

2. How will SCOTUS be used?

In this section, we will review six areas in which key research questions can be addressed through examination of the new SCOTUS corpus. This survey is intended to be illustrative, rather than exhaustive, since the recent history of corpus analysis has shown that there is no finite limit to the ways in which well-constructed corpora can be used. The areas we will survey include large corpus recognition, linguistic variation, multi-party conversations,

A. Large corpus recognition
The speech research community has been clamoring with increasing vehemence for an increase to a multi-thousand-hour transcribed corpus. Corpora such as Switchboard and TDT [refs], which are the main supports of current speech recognition technology, have between 250 and 700 hours. In comparison, the SCOTUS corpus with its 6000+ hours of speech will increase the size of the database by an order of magnitude. Where text corpora of tens of millions of words once sufficed, current projects target one billion word data sets to enable robust language modeling of uncommon patterns. In speech, new approaches to re-training in speech recognition (Lamel, et. al., 2001) require thousands, not hundreds, of hours of audio data. Moreover, in constructing this larger corpus, it is important that it include a wide range of speakers and topics. In SCOTUS, some of the speech is from the justices, which allows speaker adaptation given a small number of known speakers, while the rest of the speech is from a large number of others who mostly do not repeat, offering plenty of data for speaker-independent modeling.

SCOTUS provides an ideal avenue for constructing this crucially needed large corpus. Compared to other alternatives, it has these advantages:

* the material is already recorded and carefully transcribed, so that the expense per unit time will be low;

* the material is in the public domain, so that there will not be costs or barriers associated with copyright licenses;

* the material has great intrinsic value and interest, so that a large existing community of scholars can provide (and has already provided) interpretations of its content, and can evaluate proposed analyses and interpretations;

* there is an enormous amount of background material for each recorded discussion, in the form of the judicial records at all levels for the case in question, as well as scholarly journalistic documentation and commentary.

* the SCOTUS Oral Arguments corpus will be enormously important in its own right. With over 6,000 hours of audio covering many more arguments and a span of almost 50 years, the corpus is a repository of American constitutional thought documenting the recent history of the nation’s highest court from multiple viewpoints and showing the thinking, interactions and influence that preceded the polished majority and minority opinions. As such, SCOTUS documents the history of ideas in American constitutional law over the most recent half-century.

B. Studying linguistic variation

SCOTUS will also have a major impact on the study of linguistic variation. The past two decades of experience have underlined the magnitude, extent and importance of variation in linguistic patterns across speakers, styles and social contexts. Just as contract law and tax law must each be studied on its own terms, although they share fundamental concepts, so the language of personal life narratives and of biomedical research publications must each be studied on its own terms, despite the many consequences of the fact that they are both English. The shared social practice of sub-communities creates new sub-languages, which may entail new communicative patterns at all levels from vocabulary to rhetoric. For this reason, engineers have learned that good performance requires systems to be trained using data from a
To some extent, this linguistic variation can be predicted from other characteristics of the social and communicative context, but much of the contextual variation in language is as arbitrary as contextual variation in modes of dress and other aspects of culture. Engineers are interested in building systems that can adapt across linguistic situations -- indeed, today's engineering practice employs basically the same speech-recognition programs for dictation of English business letters and transcription of Arabic news broadcasts. However, making the system work for either application requires statistical retraining with a large enough corpus of material well matched to the application. Thus even a completely adaptable approach can't avoid looking empirically at how communication works in each cultural context of interest.

Given the enormous number of possibly distinct situations, engineers and scientists have naturally selected cases on the basis of social and economic importance, or on the basis of their benefit in focusing attention on important general problems. In the speech recognition area, connected digits and radiological dictation are examples of early corpora collected because of social or economic impact, though their generality and scientific value were limited. Switchboard, which involves telephone conversations on assigned topics among strangers, was a good way to focus attention on conversational telephone speech and on variation among individual speakers, though the intrinsic value of the content was of course small.

C. New technologies for studying multi-party conversations

We have learning that new communication situations bring new problems to the fore. In the case of SCOTUS, the major new challenge for speech technology is the processing of the conversational features involved in such multi-party conversations. The analysis of task-oriented multi-party conversations offers some of the most important challenges for linguistic science and engineering over the next decade. How can we get speech analysis algorithms to deal with overlapping speech, full of interruptions and incomplete utterances, in which changing sets of speakers are sometimes cooperating in constructing a shared message and sometimes competing to determine what the message will be about? How can we do syntactic, semantic and pragmatic analysis of such highly interactive language? How can we model the way in which content emerges in such situations?

These are questions for which there are no satisfactory answers at present, though there are some promising lines of inquiry. The biggest roadblock is the relative lack of relevant data. Recently, several research sites in the U.S., Europe and Japan have begun collecting and publishing modest amounts of data from meeting contexts of various sorts [refs]. However, the available material is far too small for anything but very preliminary research.

The importance of these questions can hardly be exaggerated. The emergence of shared meaning and the communication of individual attitudes in discussion and debate are central to human experience, and any genuine progress in scientific understanding of this process will resonate in many areas of the social and human sciences. There are also many obvious practical applications for new science in this area, starting with an automated "recording secretary" who could prepare minutes of a meeting or search earlier discussions for relevant remarks. Other applications are harder to foresee in detail, such as a rhetorical tutor who could help students learn
to argue more effectively, or to appreciate the reasons for the rhetorical successes and failures of others.

D. Historical change

For purposes of linguistic description, especially the quantitative analysis of linguistic variation, the SCOTUS corpus will offer a unique kind of longitudinal data. Within quantitative sociolinguistics, for example, longitudinal data is considered as valuable as it is rare. Labov (1994) summarizes the need for longitudinal data.

It is obvious that distributions [of linguistic variation] across age levels might not represent change in the community at all, but instead might represent a characteristic pattern of 'age-grading' that is repeated in every generation (Hockett 1950).

The normal procedure in the quantitative study of language change, collecting data in a single epoch from speakers who represent different age groups leaves the linguist with the problem of determining when a distributional pattern observed in this "apparent time" actually represents a change in real time. Labov continues: "The obvious answer to the problems involved in the interpretation of apparent time would be to rely upon the observations in real time...." However, longitudinal data is often difficult and therefore expensive to collect. Labov explains: "There are two basic approaches to the problem of accumulating real-time data. The simplest and most efficient is to search the literature dealing with the community in question and to compare earlier findings with current ones. The second approach is much more difficult and elaborate: to return to the community after a lapse of time and repeat the same study."

Labov does not even consider the third alternative, presumably because he considers it cost prohibitive: to continually sample a community over a long duration. Nonetheless this is exactly what SCOTUS offers us. SCOTUS documents Supreme Court oral arguments from 1955 to the present. During that 47-year span, 28 justices served whose tenures averaged almost 15 years. This includes several justices with very short terms such as Arthur J. Goldberg (Associate 1962-1965) and Abe Fortas (Associate 1965-1969); others whose longer tenures overlap only partially with the SCOTUS recordings such as Harold Burton (Associate 1945-1958) and Stanley Reed (Associate 1938-1957) balancing others such as William H. Rehnquist (Associate 1972-1986 Chief 1986-), Byron R. White (Associate 1962-1993) and William J. Brennan, Jr. (Associate 1956-1990) whose tenures each exceed 30 years. Complementing other longitudinal collections that sample their population at distant intervals of several years, SCOTUS will contain frequent, often weekly, recordings of the justices permitting observation and analysis at a time granularity and over a duration not generally possible.

E. Decision process analysis

The fifth research area we will discuss deals more directly with the unique nature of the communications and arguments presented at the Supreme Court. We will refer to this line of analysis as decision process analysis. It can be illustrated most clearly through examples from actual court cases. The examples we will use come from Roe v. Wade (1973) and Waters v Churchill (1994).
During the oral argument in *Roe v. Wade* (1973) Justice Byron White questioned Sarah Weddington (counsel for Roe) about an issue not addressed by either of the parties in their briefs. He was concerned with whether or not abortions could be performed on demand throughout a pregnancy, or whether the state had an interest in restricting them at some point during the nine-month term. To determine the answer Justice White asked Weddington, "And the statute does not make any distinction based upon at what period of pregnancy the abortion was performed?" Weddington's response was unambiguous: "No, your Honor. There is no limit or indication of time, whatsoever." Justice White pursued this issue further when the Court heard rearguments in *Roe*, and insisted that Weddington provide more detailed responses to his questions. Justice White's query during the arguments was the first time the issue of time-bound competing interests entered the record in *Roe*; the parties did not address it in their briefs. While the Court asked each side to prepare written arguments about it, the above confrontation demonstrates that Justice White used the oral arguments to highlight the importance of this policy issue. [web links embedded in red text. Requires RealOne Player.]

Ultimately, the Court's opinion dealt directly with the timing of a state's interest in regulating abortions. In so doing, Justice Harry Blackmun created the controversial trimester scheme – the cornerstone of abortion rights in America for sixteen years. While it is unclear from where he derived the argument that a state's interest in regulating abortions increases during each trimester of a pregnancy, one can infer that it originated with Justice White's line of questioning at the oral arguments. In other words, the issue arose, not from the party's briefs, but directly from the exchange between Justice White and Sarah Weddington.

What transpired in *Roe* is not unique. Rather, it is just one example of how oral arguments play a unique role in the policy decisions made by the Supreme Court of the United States. Existing evidence supports this claim. Anecdotally, Cohen (1978) finds explicit instances where Justices Lewis Powell and John Paul Stevens utilize issues from these proceedings in their opinions, and Benoit (1989) shows that the Court's majority opinions include issues advanced by the winning party during oral arguments. Additionally, Wasby, D'Amato, and Metrailer (1977) indicate that the oral arguments played a key role in many of the Court's desegregation decisions.

1. **The need for systematic study**

   In the first systematic account of the role oral arguments play for the Supreme Court, Johnson (2001; forthcoming) demonstrates that the justices use these proceedings to gather information to help them make substantive legal decisions. Specifically, the Court devotes almost a third of its questions to policy issues. Additionally, almost 27 percent of the Court's questions focus on the preferences of actors external to the Court. Beyond the types of information justices gather, Johnson finds that the Court utilizes oral arguments to obtain information not provided by the parties. Indeed, over 80 percent of the Court's oral argument questions focus on issues not found in the litigants' briefs. These are important findings because they demonstrate that oral arguments provide justices with information that may help them make efficacious decisions close to their preferred outcomes.
While compelling, there are three potential problems with this initial analysis. First, the transcripts do not explicitly note which justices ask which questions during oral arguments. Instead, the transcripts only say "Court" or "Question" before each question. This forced Johnson to make inferences about justices' behavior at the aggregate level rather than at the more appropriate – individual justice – level. Second, because of the limitations of coding written transcripts without electronic search tools or voice recognition software, the analysis was limited to identifying, and simply counting, the types of questions justices ask during oral arguments. Finally, the time period (1972-1985) and the issue area (civil liberties) used in this initial study (even though theoretically grounded) limit the generalizability of the findings.

These shortfalls in no way suggest that this work is unimportant, or that the findings are invalid. Rather, these criticisms are meant to highlight the fact that this research requires further elaboration and testing. Indeed, court-watchers now know that, contrary to conventional wisdom (Segal and Spaeth 2002), oral arguments play an integral role in how the court makes decisions. The task now is to dig deeper – to determine exactly how oral arguments come into play, as well as how the efficacy of these proceedings varies from justice to justice, from court era to court era, and from issue area to issue area.

(2) Tracking decision processes in SCOTUS
The project will create a simple tool using intelligent agents and manual guidance to mark up the transcripts with new data types. The ability to electronically search for specific words, voices, and length of time devoted to the discussion of issues (at the individual justice level) will allow us to take the next step needed to test and revise previous hypotheses. Further, these new analytical tools will allow researchers to reveal, with greater probity, the extent to which, and exactly how, justices across the ideological continuum and across time (1955-2001) utilize oral arguments when making decisions. The specific research questions to be addressed by this project component include (but are not limited to) the following:

(a) Do all justices utilize oral argument proceedings to obtain similar information?
(b) Do they seek different types of information during oral arguments in different types of cases (e.g., statutory vs. constitutional)?
(c) How is this information ultimately used to shape Supreme Court policy?
(d) For issues arising during oral arguments, does use vary when a justice writes a majority, dissenting, or concurring opinion?
(e) How, and to what extent, do oral arguments help Supreme Court justices learn about the preferences of other institutions (e.g., Congress, the executive branch, and the bureaucracy)?

Without access to fully searchable, voice-identified electronic transcripts, the answers to these questions become virtually impossible. Stated another way, work in this area will not proceed without access to the technology that will allow complex searches of individual justices and litigants, what they say, and how – and how long – they say it.

Overall, the answers to these questions have important implications for the role of the Supreme Court in the democratic process. For two centuries legal scholars have debated the proper role of the Supreme Court in American democracy. Some argue that it is inappropriate for the Court to make decisions that often affect the most important and controversial issues in the United States because justices are
unelected, and therefore unaccountable to any constituency. Others argue that the Court is meant to be antidemocratic precisely to act as a check on the democratic institutions in our government. The problem is that because the Court is shrouded in secrecy we do not know the extent to which the justices are responsive or unresponsive to the majority will in the United States. Analyzing the oral arguments provides such an opportunity. Through these proceedings investigators like Johnson will be able to determine the extent to which justices seek information about Congress, public opinion, the views of particular groups, or about how the president wants them to act. This, then, will shed valuable light on an age-old question that has puzzled more than a century of Supreme Court scholars.

F. Rhetorical analysis

A sixth major research area whose work will be significantly advanced by access to SCOTUS is the study of rhetorical patterns in legal argumentation. Rhetorical analysis begins with a focus on how speakers use specific linguistic structures and discourse patterns to express underlying argumentative structures. For example, Ashley’s (1990) HYPO program uses tools from Artificial Intelligence (AI) to model the use of hypotheticals (if, imagine, perhaps) in both cases involving trade secret misappropriation and Supreme Court oral arguments.

Consider the use of the word *if* in Waters v Churchill, when Richard H. Seamon (Assistant to the Solicitor General) argues that: "[U]nder this Court's decision in Mount Healthy, an employee must show both that she engaged in protected speech and that she was fired because of the fact that she engaged in protected speech. If the hospital did not know that she had engaged in protected speech, it could not have been motivated by that fact." Here, the hypothetical *if* serves to set up an imagined situation without one of the preconditions required for a First Amendment appeal in a retaliatory discharge case. However, as Justice Ruth Bader Ginsburg succinctly notes in a subsequent question: "But isn't this case different from Mount Healthy in that the question is what was the speech? ... Here the employer thinks the speech is one thing when the employer acts, but it turns out to be something else. I thought that's what this case was, and that -- that such a case has not been seen." Next, Ginsburg explores what the procedure would be if the agency were a public institution and not a private hospital. Finally, Justice Antonin Scalia explores what would have happened if the speech had been disruptive of the activities of the hospital. [web links embedded in red text. Requires RealOne Player.]

In addition to focusing on specific devices, rhetorical analysis also seeks to characterize sequential interactional patterns. This type of analysis has been particularly well developed in Mann's Rhetorical Structure Theory (RST) (1988). In Supreme Court oral arguments, there is a remarkably standardized RST pattern in which each attorney provides an opening statement and then is asked to answer a series of questions from the justices. Although this basic pattern is well defined, there are also important ways in which attorneys and justices diverge by including statements, rhetorical questions, complaints, and procedural statements.

These first two levels of rhetorical analysis require only a fairly superficial parsing of the transcripts. However, to gain a fuller understanding of the uses of particular devices and the posing of particularly difficult arguments, it is often necessary to go beyond the superficial structure to the construction of an underlying propositional network that expresses the beliefs and strategies underlying each argument or question. Tools for constructing such systems include Fisher's SemNet, VanLehn's
SIERRA, and Frederiksen's CodaZ. The SCOTUS Project will use CodaZ as an initial framework for developing a system for notating propositional networks in a way that characterizes the links between arguments, legal precedents, legal precepts, and beliefs. Once these networks are constructed, it is possible to evaluate ways in which the justices externalize their beliefs and doubts about previous opinions, details of interpretations, and the arguments presented before them. Networks can also allow us to model and understand the competition between alternative arguments and shared mental models (Clark, 1996; Toulmin, 1958).

The SCOTUS corpus will support rhetorical analysis on each of these three levels – individual words and devices, sequential structure, and propositional networks. Given the size of the corpus, it is important to make maximal use of automatic searches based on words and word sequences. Brüninghaus and Ashley (2001) and others have shown how statistically based searches of this type can be used to characterize and classify argument types. However, it will also be important to construct detailed propositional analyses of a small set of cases from the larger corpus. These cases will be the focus for the process of collaborative commentary described in the next section.

3. A new paradigm for interdisciplinary collaborative research?

In the previous section, we surveyed six ways in which the SCOTUS corpus will be used in specific research communities. This list was meant to be illustrative, rather than exhaustive. However, we believe that development of SCOTUS opens up new interdisciplinary possibilities that reach well beyond the six specific research areas mentioned above. Construction of the SCOTUS database enables a new paradigm for shared research in digital collaboratories. In order to understand the potential importance of this new approach to this new resource, we need to briefly analyze the way in which research on the dynamics of human communication is conducted today. There are currently four major methodological paradigms:

**Experimental work.** In this first paradigm, tasks and situations are contrived to elicit specific conversational patterns. The advantage of this approach is that it allows for experimental control. The disadvantage is that it tends to sample a small set of behaviors, often outside of their natural context. It is particularly difficult to apply this method to the study of political behaviors, argumentation, rhetorical structures, or persuasive discourse.

**Small segment analysis.** In this framework, a researcher applies a handcrafted analysis scheme to some small passage in a naturally occurring interaction. An enormous amount of productive work has used this method. The advantage of this approach is its applicability to all types of interactions. Its disadvantages lie in the problem of extending a detailed analysis of one passage to other passages and genres. Moreover, it is difficult for researchers to directly disprove or challenge alternative analyses across a wider data set.

**Automatic corpus analysis.** Instead of focusing on just small segments, some analysts attempt to automatically analyze discourse patterns in larger corpora. One advantage of this approach is that it can achieve greater generality across interactions and genre. A second advantage is that the analyses involved can easily be replicated and varied. The disadvantage of this approach is that only the most superficial aspects of communicative behavior can be coded automatically across large corpora.
Corpus analysis through coding. In an attempt to deepen the type of conversational and rhetorical patterns that can be studied across corpora, some researchers have attempted to code fairly large quantities of spoken interaction for specific speech act or discourse patterns. The advantages of this work lie in its theoretical and conceptual grounding. The disadvantages lie in the time-consuming nature of the effort and the fact that only a few isolated dimensions of interactions can be exhaustively coded, often failing to generate a complete picture of the interactions involved.

The SCOTUS corpus will provide researchers with an opportunity to elaborate a new and more powerful methodological approach to rhetorical and discourse analysis, which we will call, extended collaborative commentary. In order to understand this new method, let us first take a look at some recent related advances. For example, consider a collection of articles edited by Mann and Thompson (1992). In this collection, 12 discourse analysts examine a two-page fund-raising letter mailed out in 1985 by the Zero Population Growth (ZPG) group after publication of their Urban Stress Test. This work examined the rhetorical structure of a letter to the editor on the subject of population growth. The articles in the book provided alternative ways of viewing the structure of the letter. Despite the high quality of the analyses, the fact that the text and commentary were not computerized meant that it was difficult to follow links between comments and the original text and virtually impossible to track agreements and disagreements between the commentaries. Thus, this commentary failed to give rise to productive competitive argumentation.

More recently, there have been a series of attempts to develop a fuller framework for collaborative commentary. These attempts have been conducted within the context of the TalkBank project, a joint effort of Carnegie Mellon University (Brian MacWhinney and Howard Wactlar) and the University of Pennsylvania (Mark Liberman, Steven Bird, and Peter Buneman). The first TalkBank effort in this direction was a special issue of the journal Discourse Processes published together with a CD-ROM (1999, T. Koschmann, ed.). The special issue was composed of analyses by six authors of the same digitized six-minute video segment from medical school problem-based learning, along with a set of three commentaries. This CD-ROM and the videos it contained became the focus of the first TalkBank meeting held at CMU in the Fall of 1999. Based on the experience gained from the Koschmann special issue, TalkBank organized and produced the CD-ROM for a special issue of the Journal of the Learning Sciences (2002, A. Sfard & K. McClain, eds.). Here, six different students of educational practices commented on a videotaped 7th grade classroom interaction exploring the role of designed artifacts in mathematics learning. This new JLS special issue was planned enough in advance to give the authors access to transcripts linked to the video in the TalkBank CLAN program. Relying on the final PDF versions of the articles from the publisher, experts edited the PDF files using Acrobat to create links from specific portions of text to segments of the QuickTime video. Finally, at the very last stage in the process, MacWhinney read through the analyses of each of the authors and placed their commentary back into the original CLAN transcripts with links to the video. Unfortunately, the tight publication schedule made it impossible to include detailed commentary directly from each author. However, this JLS issue nonetheless stands as a first case of a primitive form of digital collaborative commentary. TalkBank is now pursuing similar CD-ROM-based projects in the areas of Emergency Medicine (Colin Mackenzie), High School Mathematics (Dave Carraher), and aphasic discourse (Audrey Holland).
These first forays into CD-ROM-based publication have piloted a basic procedure for conducting a limited form of collaborative commentary. The methods developed in this initiative can now be linked to the concept of a Digital Collaboratory. Collaboratories are designed to allow researchers ongoing, interactive access to a core set of data that form the empirical basis for much of the work in the field. Collaboratories have made great progress in areas such as field biology, seismology, and genomics. In the field of child language acquisition, the CHILDES (Child Language Data Exchange System) Project, organized by Brian MacWhinney has come to serve as the focus for most new empirical work on the development of spoken language. However, there has not yet been a project that has linked together the notion of a digital collaboratory with the notion of collaborative commentary. SCOTUS provides us with an excellent opportunity to develop the first such system for extended collaborative commentary.

There are four reasons why the SCOTUS corpus is ideal for this purpose:

Size. Because SCOTUS will contain more than 6000 hours of material, it provides an enormous resource for the study of variation in rhetorical patterns across decades, speakers, cases, and legal issues. Finiteness. Despite its massive size, the scope of SCOTUS is extremely well defined. It represents a unique genre of conversation, interaction, and argumentation that is tightly constrained by both the rules of legal analysis and the particular protocol of the Supreme Court. Quality. The relatively high quality of the recordings and transcripts, as well as the linkage of the transcripts to the recordings provides researchers with excellent linked access to particular segments. Social and academic relevance. Above all, the fundamental social and political relevance of the issues decided by the court guarantees that these dialogs will remain of interest to all generations. This means that, once a collaboratory and procedures for collaborative commentary are established, the system will be self-perpetuating.

But exactly how will researchers work with this new material and how can we maximize the process of collaborative commentary? To organize this effort, we plan to implement a five-stage process.

1. Group formation. Before we can talk seriously about the formation of a collaboratory, we need to contact a large number of researchers in this area with the goal of inviting them to participate in the formation of first work groups and then a collaboratory. However, because a very large group would be unwieldy, we will constrain our invitations initially to researchers in two classes. The first group focuses on legal scholars and active litigators with a specific interest in Supreme Court advocacy. Three leaders have expressed their interest in the collaboratory and their willingness to participate in its work: Former Solicitor General Charles Fried (who returned to the Harvard Law School faculty), Tom Goldstein, Esq. (partner in Goldstein & Howe, a boutique law practice specializing in Supreme Court litigation), and David Frederick, Esq. (a Supreme Court litigator and author of *Supreme Court and Appellate Advocacy*). The second group includes students of discourse structure, who will view this material as a test-bed for their systems of rhetorical analysis. Among the viewpoints included here are rhetorical structure theory (RST), centering theory, conversational analysis, sub-goal architectures, and conflict resolution theory.
2. **Focal segments.** The first activities of this group will be to identify a small number of cases. Within those cases, 5-minutes segments will be further selected for detailed analysis.

3. **Collaborative publication.** The first product of this group effort will be a publication of a set of collaborative and competitive analyses for publication as special issues in one or more of these relevant journals: *Discourse Processes, Language and Society, Rhetoric and Public Affairs, Journal of Communication, Law and Society,* and *Journal of Appellate Practice and Process.* These publications will include digital versions of the transcripts linked to the audio. The transcripts will be elaborated with commentary illustrating each form of analysis. In addition, the digital resources (transcripts, comments, audio) will include PDF copies of the articles with links to the relevant segments of the audio and transcripts. In some cases, it may be possible to provide links to other portions of the SCOTUS corpus and additional electronic materials. The transcripts will be formatted in XML and will be played through the TalkBank XML editor.

4. **Collaboratory creation.** In order to broaden the system of collaborative commentary across the research community, we will need to open the database to commentary from all researchers. This will be done through web-based editing using a variety of Java XML tools. The clarification and testing of competitive analyses will be further promoted through mailing lists and commentary organized through ZOPE.

5. **Data-mining.** An important step in this project involves the transition from a focus on small segments of SCOTUS argumentation to an examination of rhetorical patterns across the larger corpus. This type of work involves the deployment and construction of data-mining tools designed to locate interactions with specific properties. For example, we may want to find instances in which the justices ask the attorneys to clarify their interpretation of previous court decisions. Or we may seek to locate cases involving arguments regarding national security issues as limitations on particular constitutional rights. Within these larger search sets, we will then want to further analyze the structure of these arguments and the judges' questions with uniform coding systems.

This work should lead to major advances in our understanding of the shapes of rhetorical, logical, and legal analyses presented to the court and how these arguments vary as a function of changes in legal practice and the larger society.

This proposal echoes several elements in the recently released Atkins Report by relying on cyberinfrastructure to exploit "mutual self-interest and synergy among...research communities...." Atkins and his colleagues envision "comprehensive libraries of digital objects and multidisciplinary, well-curated federated collections of scientific data...and the ability to collaborate with physically distributed teams of people using...these capabilities." We believe this proposal is fully consistent with the report's highest aspirations. (Atkins, 2003).

Prior NSF Support

Support from NSF enabled the creation of The OYEZ Project [<http://www.oyez.org>](http://www.oyez.org) – an online multimedia database devoted to the U.S. Supreme Court -- beginning with a small instrumentation grant in 1996 and followed by a larger grant from NEH and another NSF grant as part of the "National Gallery of the Spoken Word." More than 1 million unique user sessions access OYEZ data each month. Approximately
one-third of the SCOTUS audio corpus has been digitized selectively by The OYEZ Project with additional access to complete SCOTUS opinions provided by FindLaw <http://www.findlaw.com>. NSF has supported both the CHILDES and TalkBank Projects, although the majority of the support for CHILDES has come from NIH. These projects have both succeeded in producing large databases which are used each year by thousands of researchers, producing hundreds of published articles. The software from these projects is also used by hundreds of other researchers to produce new data sets which will eventually be contributed to either CHILDES or TalkBank. The complete set of published articles for CHILDES can be found at <http://childes.psy.cmu.edu/pubs>. The bibliography for TalkBank-related articles is still being compiled. NSF has supported the Linguistic Data Consortium with large grants from 1995-1999. An examination of LDC's online catalog is evidence of its results <http://www.ldc.upenn.edu/Catalog/>