# 1. Results from Prior NSF-sponsored Research

NSF has provided support for 14 projects relevant to the proposed ComNet system. The publications from these projects are given in Appendix A6. Eight of these projects have resulted in the construction of large language databases: *TalkBank* (MacWhinney), *CHILDES* (MacWhinney), *Informedia* (Hauptmann), *Universal Digital Libraries* (UDL; Carbonell, Reddy, Shamos, St. Clair), *Let's Go* (Eskenazi), *Networking Data Centers* (Cieri), *Mining the Bibliome* (Liberman) and the *Linguistic Data Consortium* (LDC; Cieri, Liberman). *TalkBank*, *CHILDES, LDC,* and *Let's Go* have constructed databases of spoken language; *Informedia* has focused on the audio-visual aspects of communication; and *UDL* has digitized and indexed books. Five other projects have constructed tools for the analysis of these databases. The *AVENUE-LETRAS* project (Carbonell, Levin, Lavie) has compiled materials and methods for the translation of indigenous languages. *Mining the Bibliome* (Liberman) has also built systems for information extraction from biomedical text. The Lemur/Indri search engine project (Callan) has constructed open-domain high-powered search engines for such databases. REAP (Callan, Eskenazi) has developed methods for configuring instructional software based on analysis of these databases. The GRASP parser (MacWhinney, Lavie) has constructed tags and parses for all of the English-language corpora in the child language database (CHILDES). ComNet will merge these NSF-sponsored research strands, along with many other resources, into a single tightly integrated, open access, sustainable, interoperable, distributed database.

2. Vision and Rationale

Each day, we generate billions of rich communication data streams in the form of conversations, meetings, broadcasts, newspapers, scientific articles, and links to scientific data. These streams of communication can be captured digitally and enriched automatically through systems for linguistic and video annotation. The TalkBank, CHILDES, Informedia, Let's Go, LDC (Linguistic Data Consortium), and UDL (Universal Digital Library) databases have demonstrated how this can be done, amassing the world's largest collections of data on spoken and written communication. However, to properly integrate these materials into a technologically sophisticated DataNet Partner, we must achieve full interoperability, universal access, deep curation, powerful visualization, and wider coverage. To this end, we propose a DataNet Partnership that will build an international database for digital data on human communication, including speech (conversations, interviews, radio broadcasts, etc.), text (historical and modern books, journals, newspapers, web pages, blogs, etc.), scientific communication (articles, letters, debates, commentary, reviews, linked to primary scientific data), video, and images. This new DataNet Partnership, called ComNet, will provide crucial scientific data for a wide set of NSF programs within SBE, CISE, EHR, and Biology, as well as for segments of NIH and IES. Beyond these specific scientific communities, the database will provide material for students, teachers, businesses, professionals, language communities, government, and the public. When fully operational, ComNet will help transform our scientific and cultural understanding of the vast landscape of human communication.

ComNet Data: The primary ComNet data structure is an alignment between a digitized transcript (or annotation) and digitized media. As Bird & Liberman (2001) have shown, linkages of annotations to media universally assume the form of directed acyclic graphs (DAGs) which they call "annotation graphs" (AG). Although this structure is universal, there are dozens of methods for characterizing and displaying these transcript-media linkages. The next page shows two ways in which a transcript can be linked to media. In the first format, from ELAN, we see the transcript displayed in a musical notation format under the QuickTime window. In the second format, from CLAN, we see the transcript in a standard textual display. In both cases, as

the researcher plays through the video, the respective segment of the transcript is highlighted so that the researcher can study the transcript in tight synchrony with the actual experience of the interaction.
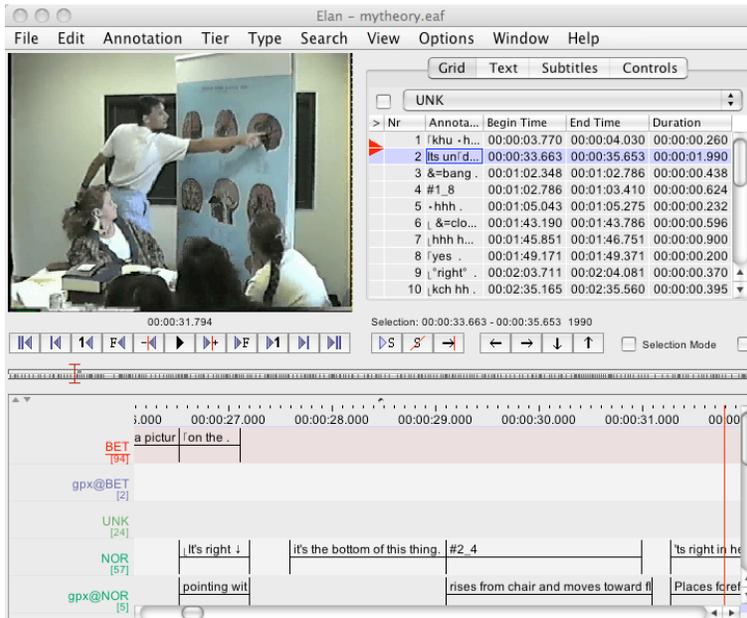
A time-aligned transcript as displayed by ELAN.
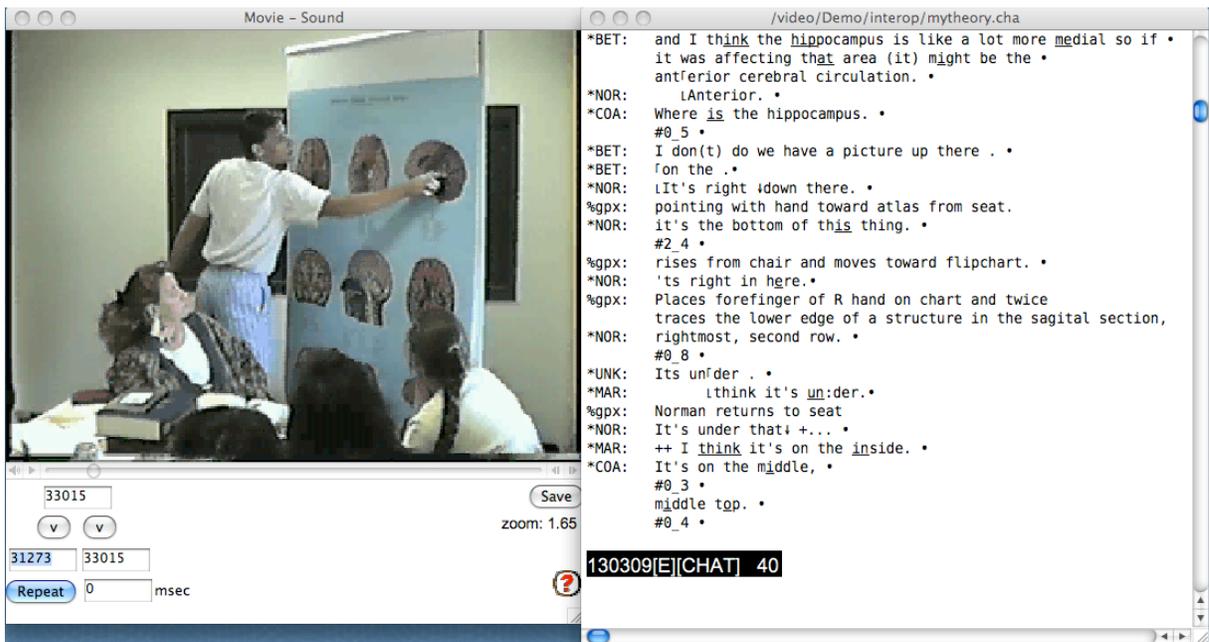
A time-aligned transcript as displayed in CLAN

This particular videotaped interaction was the subject of a special issue of *Discourse Processes* in 1999. In the interaction, medical students in a problem-based learning (PBL) class are trying to link the position of the hippocampus to observed symptoms in a neurological patient. The CD-ROM at the back of the special issue presented the transcripts linked to the media. These

2

transcripts with linked media can be browsed over the web from http://talkbank.org/browser. This particular example is at ClassBank/CogInst/mytheory.cha.

The advantage of the ELAN musical score display format is that it accurately displays the duration of a communicative activity and the synchrony across types of communicative activities (intonation, head nodding, words, etc.). The advantage of the CLAN transcript format is that it is easier to read and presents a better overview of the conversation. Each visualization format has its own advantages and disadvantages. The important point is that, with full data interoperability, researchers can display the same underlying data in whichever viewer or editor they need for the particular project at hand. Toward this goal, we have written programs that can reliably convert data from all of the major transcript-media display programs (AG, Anvil, ELAN, EXMARaLDA, HIAT, SALT, DRT, SyncWriter, MediaTagger, Transcriber, Transana, Praat, WaveSurfer, SoundWalker, and Observer) into the ComNet XML format. Moreover, transcriptions and other annotations can be transformed from XML back into each format and compared with the original to show that no data was lost during the roundtrip conversion. The interoperability we have achieved through these programs provides an important building block for ComNet, since it allows us to integrate the hundreds of data streams from the diverse partners participating in ComNet into a single, unified database. Apart from the core data type of transcripts linked to media, ComNet will also integrate textual data streams from newspapers, libraries, national surveys, social networks, and broadcast media. These textual streams will be configured for use with the same tools for search, analysis, and visualization that we will develop for the spoken corpora.

Basic Principles: ComNet will construct a fully interoperable set of tools for data acquisition, curation, analysis, and visualization based on the ComNet (talkbank.org/talkbank.xsd) XML schema. Using these tools, we will link together an international distributed database on human communication, structured in accord with these seven basic principles:

1. **Open access**. The core components of ComNet will be freely open to all – students, researchers, government, private industry, and the public. For non-core components, access may be limited by LDC licensing, TEACH Act requirements, or IRB restrictions.
2. **Integrated structure**. Every item of ComNet data – both core and non-core – will have a unique and specifiable position within a single, comprehensive XML-based data grid. To achieve this tight integration, we will mesh existing formats into the single ComNet XML Scheme, thereby overcoming the Format Babel identified by Bird & Liberman (2001).
3. **Interoperability**. To promote interoperability, we will construct roundtrip conversions of ComNet data to other popular analytic programs for display and analysis.
4. **Powerful analysis**. Relying on ComNet XML, the project will build powerful tools for browsing, searching, visualization, statistical analysis, and report generation.
5. **Sustainability**. ComNet sustainability relies on three methods. First, we will integrate the open-access model of TalkBank with the fee-based membership and licensing models of LDC. Second, we will link with individual research communities to promote long-term research funding initiatives. Third, we rely on long-term support commitments from the University, College, and Library at Carnegie Mellon.
6. **Survivability**. We will use redundant storage, distributed backup, mirroring, migration and, most importantly, sharing to guarantee data and metadata survival. We will also develop methods for assuring managerial survival.
7. **Massive multilinguality**. ComNet will integrate data from hundreds of languages, ranging from English and Spanish to minority and endangered languages such as Mapudungún, Ojibwe, and Iñupiaq.

Intellectual Merit: Transforming the Study of Communication. What makes ComNet transformational is our commitment to the principles of open access and integrated structure. Open access to ComNet data will enable fundamental linkages between research in psychology, education, linguistics, biology, neuroscience, ethology, anthropology, sociology, political science, economics, demographics, computer science, natural language processing, human language technologies, library science, law, area studies and comparative literature. Although ComNet focuses on transcripts linked to media, the database will eventually be configured as a component of a larger DataNet grid for the social and biological sciences. Anticipating this, we will allow researchers and participants to link individual-level transcript data to related neurological, economic, medical, survey, social, and preference data. The construction of these additional linkages will be directed and controlled by the participants themselves. To maximize direct access for participants and researchers, we have built a browser-based facility at http://talkbank.org/browser for listening to interactions and reading the transcripts in synchrony with the interactions. This system allows researchers to comment on specific segments of these interactions with their comments being stored in an SQL database. For deeper analysis, unfettered by network bandwidth limitations, researchers can download transcripts and media for use with localized versions of the same or more powerful access tools. To maximize the integrated structure of ComNet, we curate all ComNet datasets in accord with the XML specifications at http://talkbank.org/software/talkbank.xsd. Tools based on this XML definition have allowed us to integrate 158 corpora into a consistent XML Schema and to build tools that automate and guarantee accurate curation for new data streams. The availability of these resources has radically transformed the way research is carried out in the four areas of child language acquisition, aphasiology, speech technology, and second language learning. The goal of ComNet is to extend the transformational model developed for these four areas to a wide set of additional DataNet Partners described below.

DataNet Partners: ComNet will play a unique role in the construction of cyberinfrastructure for the 21st Century. It builds on the methodology of computational linguistics and data mining, as well as best practices in dozens of data intensive disciplines, to address practical and theoretical issues across a wide array of NSF programs. As summarized in Appendix 5, these areas include SBE programs in Cognitive Neuroscience, Linguistics, Developmental Psychology, Law and Social Sciences, Sociology, NSCC, and Anthropology. Outside of SBE, ComNet will provide support for work in EHR, Biology, NIH, and CISE-IIS. In addition, the deployment of ComNet tools for the analysis of interview data collected by major national surveys will provide a further level of integration with the other social and behavioral sciences. By examining interview data from surveys such as the General Social Survey (GSS) or the American National Election Study (ANES) we can link facts about human communication to geographic, economic, and demographic profiles of national populations, using "fractal" methods (Gelman, Carlin, Stern, & Rubin, 2003) for data sampling and integration.

To support this integration across the SBE Sciences, ComNet depends on input from computer scientists and computational linguists. Our partners here (as documented in the attached letters) include work groups from universities and institutes in Hamburg, Indiana University, Helsinki, Swarthmore, Greenland, Chile, Bolivia, Mexico, Peru, Wales, Colombia, El Paso, Copenhagen, NYU, Amsterdam, Columbia University, University of Washington, Stockholm, Nijmegen, Lyon, Memorial University Newfoundland, Stanford, McGill, Fairbanks, Copenhagen, Odense, Minnesota, Thessaloniki, Arizona, London, Melbourne, San Francisco, Austin, MIT, Haskins Labs, Alexandria, Beijing, Bangalore, Qatar, Beirut, Illinois, University of Virginia, Saarbrücken, Mannheim, Oxford, Hong Kong, Singapore, Purdue, Michigan State, Northwestern, Eastern Michigan, Michigan, Melbourne, and Paris. An earlier submission of this project was criticized

for focusing too closely on linguistics. The letters appended to this proposal illustrate that the bulk of our partners are, in fact, not linguists, but workers in other domains. It is important to remember that ComNet focuses not on linguistics, but on communication. ComNet applies cyberinfrastructure methods to communication data from dozens of disciplines, of which linguistics is only one of the many beneficiaries, as noted in Appendix 5, and as illustrated in the letters of support.

Data Streams: In addition to the partners who will support our research and tool development, we are partnering with a wide set of agencies who will provide streams of audio, video, and textual data for ComNet. Here and in Appendix A5, researchers and agencies that have written letters documenting these data streams are listed with an asterisk after their name.

1. LDC ingests raw data from news text, blogs, newsgroups, broadcast news and talk, biomedical text and abstracts, telephone conversations, lectures, meetings, interviews, read and prompted speech, printed, handwritten and hybrid documents and recordings of animal vocalizations (Seyfarth*, Manser*). LDC also annotates these raw data to create corpora focused on specific project needs as well as ingesting completed corpora from other sources. TalkBank ingests corpora from focused research projects. These ongoing contributions to TalkBank and LDC are leading to an exponential growth in these databases.
2. Universal Digital Library (Carbonell, 1996; Reddy, 1995; Shamos, 2005) projects in India*, China*, Egypt*, and Qatar* (www.ulib.org) are continually ingesting open-access textual materials. We will link these to open-access materials from the Internet Archive and the TEI Project http://www.tei-c.org/Activities/Projects/. Our focus here is on providing systematic access to OCR-ed versions of books with a simple, but consistent, markup that allows them to be processed by computational linguistic methods.
3. Video data will come from the satellite broadcast collection being recorded at LDC, as well as copyright free video materials from the U.S government, www.archive.org, creativecommons.org/video, open-video.org, opencontentalliance.org (Kahle*), YouTube, the BBC, and Public Television. In addition, we will include 4TB of video materials from Hauptmann's NSF *Informedia* project (www.informedia.cs.cmu.edu) composed of interviews, academic lectures, documentaries, news programs, and other high-value broadcasts (Bharucha et al., 2005; Christel et al., 1995; Pan & Faloutsos, 2001; Wactlar, 2001). Yahoo! Inc has agreed to host 10,000 hours at video.yahoo.com that will be accessible to all scholars and researchers. Through these video data streams, we expect the database to approach petabyte size in a few years.
4. The CMU-NSF *Let's Go* Project records calls to a fully automated speech dialog system that provides transit information after hours (Raux & Eskenazi, 2007; Raux, Langer, Black, & Eskenazi, 2005; Raux, Langner, Black, & Eskenazi, 2003). These conversations involve real people striving to meet real and immediate needs, such as catching the last bus home after a late shift at work. Each of the 75,000 dialogs includes multiple turns, and is automatically annotated with the expected outcome and core semantics of the information conveyed in the speech. We will extend this system to recording from additional public services.
5. Interview data from the ANES and GSS surveys conducted by the Institute of Survey Research will provide streams of conversational data linked to survey information.
6. Goldman's* Oyez Project (www.oyez.org) will provide input of transcripts of legal oral argumentation linked to digital audio, with an emphasis on the Supreme Court.
7. Our collaborators in Mexico*, Chile*, Greenland*, Wales*, and Bolivia* will provide digitized recordings and transcripts from their local indigenous communities.

8. Participants in the E-MELD* project for documenting endangered languages will provide digitized records, transcripts, and grammars from diverse language communities.
9. We will receive spoken and written data from our collaboration with the Iñupiaq* and Ojibwe* indigenous communities in the United States, as well as parallel projects with the Hispanic* and ASL* (American Sign Language) communities. For these communities, we are constructing web sites that will facilitate social networking and language preservation efforts based on computational linguistics tools.
10. We will receive digital video and transcripts of classroom instruction in math and science in Spanish from the Education ministries in Bolivia* and Chile*. We will receive parallel data in English from the newly funded 500-school Gates Foundation Project directed by Roy Pea* and from the NSF Science of Learning Centers in Pittsburgh (Koedinger*, Reznick*) and Seattle (Pea*).
11. Deb Roy's* *Human Speechome* project has been collecting video data from six cameras 24/7 across two years in a study of the language development of his son. These data will not be available for open access, but will be used to develop a follow-up project that can generate open-access data.
12. Jamie Callan will supply the ClueWeb09 corpus of one billion high value web pages from 10 languages http://boston.lti.cs.cmu.edu/Data/clueweb09/.
13. From the United Nations and the European Union, we will acquire streams of simultaneous translations with transcripts linked to audio.
14. We will include from the hundred languages being sampled by the *Million Hours of Speech Archive (MHSA)* project from Johns Hopkins University.
15. We will work to include oral history archives such as *StoryCorps*, *Rivers of Steel*, Holocaust and war archives, and similar national archive collections across the world.
16. Working with Bennett Bertenthal*, Ken Pugh*, Marcel Just, Tom Mitchell, Stephen Porges, and others, we will develop methods for linking transcripts to physiological and neuroimaging data, extending earlier work in the SidGrid and BrainMap projects.

Broad Impact: It is easy for researchers to claim that their work will be transformational. In the case of ComNet, we already have solid empirical evidence for the transformational impact of the approach. This evidence comes first from the success of the CHILDES Project database for the study of first language acquisition. The bibliography at http://talkbank.org/usage lists over 3100 articles that were published in the last 14 years based on CHILDES corpora and tools. During this period, the pace of publication has accelerated each year with half of this work published in the last five years. In addition to the bibliography, this URL presents 154 letters of support from CHILDES consortium members, 38 letters from AphasiaBank members, and 27 letters from PhonBank members. The LDC catalog has had a similar impact on the fields of speech and language technology, linguistics and language related disciplines. To date, LDC has distributed 61,000 copies of 950 titles and otherwise shared data with 2700 organizations in 68 countries. The potential impact of ComNet will be far greater than that of either TalkBank or LDC. The transformational potential here derives from the explicit articulation of a core data model that aligns transcripts and other annotations to media. When this data model is linked to the principles of open access and full integration, the result is transformational. We can then annotate, replay, analyze, visualize, and comment upon communicative, social, and educational patterns across vast collections of transcripts linked to video. By providing these materials in an integrated fashion with open access, ComNet will reach out to users across the sciences, in schools, in language communities, and in government and business.

Stages of Transformation: The impact of a major new Cyberinfrastructure initiative like ComNet moves through a series of predictable stages. In the first stage, which is now complete

in child language studies and well underway for aphasiology, the members of a tightly integrated field are able to access a shared database that refocuses their research agenda. During the second stage, advances in cyberinfrastructure allow researchers to integrate data across time and space. Examples include gesture-speech linkage (McNeill, 1992), the neurodynamics of perspective matching (Greeno & MacWhinney, 2006), and phonetic compression (Bybee & Scheibman, 1999). During the third stage of transformation, integration reaches out across data types to link in other sciences. During this stage, ComNet will provide the springboard for the construction of an internationalized "dream database" that integrates linguistic, survey, neuropsychological, medical, and genetic data at both the individual and group level. Researchers will be able to compare language use across large, rich multimedia and textual corpora from fields as diverse as national survey data, minority-language preservation, rural farming communities, and classroom instruction in science and mathematics. This dream may seem far away, but there is already movement in this direction within the NSF, the ESCR, and through new groups such as the International Data Forum. ComNet will fit in ideally with this emerging new framework of internationalized data sharing.

XML integration: As we noted earlier, ComNet depends on two core principles: open access and complete integration. These new developments all rely on the full systematization of the database through XML (Extended Markup Language) – a self-extensible language for metadata definition, communication, and archival formatting. ComNet XML has a comprehensive semantic structure that can support detailed scientific searches. The fact that all ComNet datasets will be formatted in this common language is what drives the overall synergy of the project. Several companies are building suites of tools for conversion of other data formats into and out of XML, for organizing and viewing XML, and for archiving heterogeneous XML-compliant data and its associated XML metadata. CMU has secured the collaboration and support of IBM for open source access to its UIMA (XML-based Unstructured Information Management System), along with funds for support of its use. We have also secured support from 3KSoft, a smaller company dedicated exclusively to XML tools and middleware, who is making all of its XML tools and middleware freely available for this project. Finally, our work builds on years of NSF-sponsored support for the Indri and Lemur open-source search-engine projects. Together, this data-centric computational middleware will provide a solid software footing for ComNET, enabling our researchers to build atop the ComNet middleware, rather than re-inventing or buying the requisite software infrastructure. By removing the lower-level infrastructure risks and costs, we can focus on the goal of opening up the study of human communication to our whole society.

Example Projects. When ComNet resources are in place, it will be possible for users to address hundreds of types of research questions. Here are some illustrations of the types of questions that researchers have been asking, sampled from this much larger space:
1. Analyze the contents of newspapers in six different languages for their reaction to the terrorist attacks of September 11, 2001, using LIWC (Pennebaker, Francis, & Booth, 2001).
2. Find out what children in varying social groups have to say about going to the doctor and how parents respond to these expressions (Steward & Steward, 2006).
3. Track uses of 鄙人 (bi3ren2) or 在下 (zui4xia4) for "your humble servant" in Chinese texts from the Ming and Ching dynasties in order to understand the time course of changes in class structure (R. Brown & Gilman, 1960).
4. Examine the extent to which high school classes in mathematics in eight different countries engage in the process of "accountable talk" (Michaels, O"Connor, & Resnick, 2008) and how this discussion plays out in terms of whole class involvement (Pea, Mills, Hoffert, & Rosen, 2002).

5. Tabulate and graph the frequency of verb overregularizations such as *goed* and *runned* in child language samples between ages 1;6 and 3;10 (Marcus et al., 1992). Pass the matches to either Excel or R for further statistical analysis.
6. Compare videotaped story telling and route descriptions in cultures that construct geocentric spatial terms, such as Tzeltal (P. Brown, 2001) or Guugu Yimithirr (Haviland, 1993) with those that use allocentric reference. Relate these to ways in which people describe geophysical features of their environment (Mark, Freksa, Hirtle, Lloyd, & Tversky, 1999).
7. Extend the analysis of Seyfarth & Cheney (1999) for vervet monkey vocalizations to vocalizations from groups of meerkats (Hollén & Manser, 2007), using data currently in TalkBank.
8. Use scene segmentation and gaze alignment to measure how people maintain eye contact (Vertegaal, Slagter, van der Veer, & Nijholt, 2001) and postural direction during survey interviews (Groves, 2004; Schober & Conrad, 2006) when they are producing the disagreement signal *well* or its equivalent in French (*bien*), Hungarian (*hát*), German (*ja, na ja*) or Spanish (*pués*).
9. Link ComNet data to ALFRED (alfred.med.yale.edu) to locate possible genetic markers for indigenous groups whose languages lack recursive syntactic devices (Everett, 2007; Hauser, Chomsky, & Fitch, 2002).
10. Extend the method of Mitchell et al. (2008) that predicts fMRI brain activity from the meaning of nouns to the prediction of brain activity from the reading of types of passages. Extend this method across languages.
11. Create a concordance of terms referring to sustainable energy generation across European and Japanese parliamentary debates and output a set of audio files including the sentences with these references. These materials will support a wider comparison of differences in national energy policies (Jacobsson & Bergek, 2004).
12. Study the ways in which Supreme Court justices signal their intentions to vote on a given case to other justices (Johnson, 2004) during the years 1957 and 2008. Link these signals to cognitive analyses (Ashley, 1991).
13. Study the way bias and perspectives are expressed in social media web artifacts such as blogs or Youtube (Lin & Hauptmann, 2008).

These research themes are sample illustrations taken from a much larger space. Focusing just on the one field of child language, MacWhinney (2008) lists 50 such research theme types and the wider literature demonstrates the importance of at least 100 more. When we look at other fields, we find a similar diversity. In each field, however, the barriers to analysis are the same. Absence of published metadata makes it difficult to find, group, select, integrate and analyze appropriate data. Moreover, once located, there are often the barriers of Licensing and Format Babel. By lowering these barriers, ComNet will allow researchers to ask questions in ways that are currently impossible. By allowing researchers to address these new issues, ComNet will enable a transformation in the study of human communication.

Sustainability: DataNet Partners must present a plan that can be shown to guarantee sustainability. To achieve sustainability, ComNet will rely on three methods.
1. First, ComNet will rely on the LDC licensing model. LDC has a 16-year history of funding of database growth and development through corporate and university membership licensing fees. These fees, coupled with funding for specific corpus creation projects, have allowed LDC to support 45 personnel working full-time in a large and well-structured research facility. Within this model, we will distinguish the core open access segments of ComNet from the restricted access segments in the following ways. All corpora created within ComNet will be open access without fee. In addition, LDC will provide open access without fee to some

existing corpora such as CallFriend or CallHome that are of broad interdisciplinary interest. Finally, all ComNet data will be redundantly provided under the LDC model, available to LDC members at no additional cost and available for free to organizations not affiliated with LDC. All ComNet data will be provided in ComNet XML. In the case of LDC data, the ComNet XML format will be one of multiple formats available. This redundant approach will accommodate multiple user preferences and enhance sustainability. LDC and TalkBank have cooperated in this way before for the development of open access corpora such as ANC (American National Corpus), SLX (Sociolinguistics), SBCSAE (Santa Barbara), and various AnimalTalk corpora. In addition, all LDC data will be searchable and browsable using ComNet methods. This will provide improved usability for LDC members and will help in the recruitment of new LDC subscribers, thereby further strengthening the LDC licensing model.

2. Second, we will continue the current TalkBank emphasis on securing separate funding for particular community-based initiatives. NIH has provided \$6 million of funding for the CHILDES, AphasiaBank, PhonBank, and TalkBank Projects. These applications have all been funded during the first round of review with priority scores in the 99th percentile, because of the strong support they receive from the relevant scientific communities. However, the disease-based modularization of NIH makes it impossible to provide cross-domain funding of the type required by DataNet. This is why NSF DataNet funding is needed to jump-start ComNet. However, once configured, ComNet will be able to compete successfully for funding of domain-based projects in areas such as Classroom Discourse (NSF-EHR), Second Language Acquisition (Office of Education), and Endangered Languages (several sources).

3. Third, we have secured pledges for long-term data survival from the University, College, and Library at Carnegie Mellon (see attached letters). Using mirror sites, we will establish four fully operational non-development mirrors of the database and programs outside of the United States and three within the United States. For each of the seven sites, we will run yearly tests to demonstrate survivability.

3.   Activities

ComNet activities will be organized into seven thrusts: ingestion, curation, interoperability, search and visualization, speech technology, multimodal analysis, and language communities. ComNet will be centered at Carnegie Mellon with a secondary site at the University of Pennsylvania. We will devote a significant portion of our resources to inter-institutional projects that fall within the scope of the thrusts described below. For each year of the project, we will focus on two of these thrusts. During this focus period, we will invite the relevant specialists in the thrust areas to visit CMU and LDC to develop ways of integrating their work with ComNet. They will receive specific programming support from our staff and their work will be a focus of seminars during the relevant year and corresponding sessions in the ComNet Summer School. These activities will support long-term collaborations with our research partners.

Thrust #1: Ingestion

Thrust #1 will ingest streams of digital data from projects, organizations, research groups, universities, libraries, companies, and governmental agencies, both nationally and internationally. ComNet will not engage in the collection or digitization of speech, text, or video. ComNet will not be involved in corpus building. Our role is to ingest existing archives and streams of digital language data and to curate these to form a unified multimedia distributed database. Above, we listed the 16 major data stream types we plan to ingest and curate. Deposition of raw data will rely on these methods:

1. For the hundreds of researcher-based corpora that form the backbone of research in areas like bilingualism, aphasia, gesture, or animal communication, we will receive transcripts through email. Procedures for contribution, including IRB release, are at http://talkbank.org/share/irb/contrib.html. These procedures and IRB guidelines parallel those found at https://www.icpsr.umich.edu/ICPSR/access/deposit/ for contribution of data to ICPSR.
2. Contributors will use FTP to ship us the digitized media accompanying these data streams. For sites with slow Internet connections, we will provide help in transferring data using sets of small 500GB external drives.
3. Contributed data sets will be formatted in CHAT and tested for accuracy using the CHAT2XML program available from http://talkbank.org/software/chat2xml.html
4. Data that are not in correct CHAT format will be curated as described below.
5. Metadata will be structured in accord with the OLAC standard (www.language-archives.org). We will configure a web page at talkbank.org to allow contributors to input data.

Some of the 16 major data streams listed earlier are already configured to supply data in ComNet's XML format. For example, projects like Oyez, AphasiaBank, and TalkBank already subscribe to the correct format. We have well functioning translators for data from many of the other formats, including Informedia, LDC TRS, Let's Go, and ELAN. For other data streams, we will need to write routines to convert to ComNet XML.

Efficient ingestion of data on human communication depends on establishing good understandings with specific communities and research groups regarding the importance of supporting the overall data-sharing effort. Currently, granting agencies require researchers to share their data, but fail to specify exactly how this should be done. ComNet will fill this gap by providing detailed procedures for ingestion for each particular data type and project type. We will then provide detailed reports to granting agencies documenting how grant recipients have fulfilled their commitment to share their data.

Thrust #2: Curation

The curation of ComNet data involves the six activities of protection, provenance registration, validation, organization, metadata generation, and documentation.

Protection: Over the course of 24 years, TalkBank, CHILDES, and LDC have aggregated the world's largest databases of spoken language materials. The fact that this work has never triggered a violation of privacy is indicative of the attention we have paid to this issue. IRB committees across the country are now using our standards in determining their approach to making spoken language materials available. These policies are available at http://talkbank.org/share and have gone through repeated cycles of discussion and refinement from communities as diverse as teachers, parents, aphasiologists, computer scientists, lawyers, and patient advocates. Some of the procedures involved include:
1. Online storage of IRB materials and releases from data contributors.
2. Removal and bleeping of last names and addresses.
3. Password and encryption protection of sensitive materials.
4. Repeated checking with contributors to make sure that the current level of access is in accord with their wishes and those of the participants.
5. Contacting individual children when they become adults to verify that they wish to allow continued access to their data.
6. Targeted de-accessioning from the database of segments or whole transcripts that seem embarrassing or which could lead to identification of the participant.

IP, Copyright, Provenance: In addition to issues of confidentiality, ComNet will deal with issues of intellectual property, copyright, provenance, and authenticity. Both LDC and TalkBank publish all of the software we have produced under the GNU Library GPL, or equivalent open source license, and we will continue to do so. We choose the GNU Library GPL because it maintains open access, while protecting use for research. By default, we leave the copyright on corpora with the original holders who have agreed to allow us to make their data accessible. In addition, LDC maintains agreements with each of its users that are compatible with agreements previously negotiated with each provider. For books and video, copyright is negotiated in accord with procedures established by UDL. The UDL includes a collection of pre-1923 historical texts, which are out of copyright. The UDL has already negotiated copyright permission for every publication from the National Academy of Sciences; UDL procedures are documented in a Council on Library and Information Resource monograph (Troll Covey, 2002). The provenance (Abiteboul, Buneman, & Suciu, 1999) and authenticity of our corpora are assessed at collection time by direct contacts with the contributors.. After that point, the emphasis shifts to validation and change and version tracking through Subversion and change logging.

Validation: Both LDC and TalkBank conduct content validation at the time data are ingested. This process involves checking for transcription accuracy, file correspondence, and metadata entry. Format validation, which is run automatically by current TalkBank programs, applies at the level of the transcript, the corpus, and metadata. Because the database is structured in XML, it is easy to run tools that validate the adherence of new contributions to the standard Schema. Data must pass through a roundtrip from CHAT to XML and then back to CHAT (the ComNet transcript display format) with no validation errors. For corpora that are linked to media, each media time tag must correspond with a media file correctly stored on the streaming media servers. The ComNet XML schema can be found at http://talkbank.org/software/talkbank.xsd. All ComNet data must fit exactly into this schema.

Organization: To guarantee proper functioning of the database, all data are encoded within a set of five isomorphic trees for:
1. Raw CHAT data,
2. XML CHAT files,
3. Media matching the transcripts,
4. Streaming media matching the transcripts, and
5. Commentary pegged to both transcripts and media.

For example, the transcript for the 10th session of the Yasmin corpus is located at data-orig /romance/es/Yasmin/10.cha. The XML is at data-xml/romance/es/Yasmin/10.xml. The media is at media/romance/es/Yasmin/10.mov. The four alternative compressions of the streaming media are on a streaming server at /CHILDES/romance/es/Yasmin/10/ and the documentation in HTML format is at /CHILDES/romance/es/Yasmin/10.

Metadata Generation: The next step of curation involves the creation and maintenance of metadata for ISBN cataloging, DOI generation, and OLAC indexing. We will work with Gary Simmons, Steven Bird, Peter Wittenberg and others to integrate OLAC and IMDI (www.mpi.nl/IMDI/) metadata systems into ComNet. Metadata are important for smooth functioning of virtually all aspects of ComNet. This means that, as the project grows, we must continue to improve and extend the metadata set. As an example, let us consider the role that metadata will play in the curation and analysis of speech data. Currently, the /ae/ sound in Canadian English is shifting toward the /ao/ sound (Labov, Ash, & Boberg, 2006). In 40 years, this transition may be complete and it will then be difficult to correctly process earlier data without having metadata that indicates the time and place of recording and the dialect

background of the speaker. This underlines the urgency of placing labels on data as soon as possible. For this sound change, it is also important to record metadata regarding speaker age, gender, geography, and education (Labov, 2001). Collection of this sociolinguistic metadata supports a new trend in speech analysis that allows researchers to develop more coherent statistical models. The examination of speech in the context of the oral interviews that accompany major national surveys provides a unique opportunity to link rich metadata to detailed speech analysis. We will examine this in greater detail in the context of Thrust #5 below.

Documentation: The final step of curation involves the writing of a readable PDF description of each corpus for inclusion in the corpus description manual. This manual is structured both as an independent, readable document and as a set of individual, searchable descriptions with metadata fields.

Thrust #3: Interoperability

The goals of this thrust are to provide interoperability for (1) data, (2) metadata, and (3) programs.

Data interoperability: ComNet is building on 25 years of work on data integration in the TalkBank framework. TalkBank data derive from 158 separate projects, each of which was eventually integrated into the single over-arching XML framework. During this process of integration, it was often necessary to extend the framework to represent new contrasts or distinctions marked in particular corpora. This process will continue within ComNet. Because the schema is a growing framework, TalkBank tools have been constructed to allow for repeated cyclic reformatting and validation of the whole corpus, whenever a change is made to the schema. Given the goal of constructing a single, integrated database, a major goal for ComNet will be the integration of LDC materials to the ComNet XML standard. Once this is achieved, users of LDC data will have vastly improved access to LDC materials and can use ComNet tools to process these data. This merger of the two systems will also encourage additional research groups, and even new user communities whose data formats differ, to become LDC members, thereby further strengthening ComNet sustainability.

Metadata Interoperability: Computational linguists have devoted a great deal of attention to the development of systems for annotating the ontologies of human communication. Among the efforts in this direction, we should single out OLAC, IMDI, GOLD, and WordNet. ComNet will build on each of these systems. The OLAC metadata set (Simons & Bird, 2008) is a subset of the larger IMDI set (www.mpi.nl/IMDI/tools). At a minimum, ComNet data will subscribe to OLAC. However, ComNet will also support IMDI validation, as a further option. The textual segments of ComNet will be curated using the TEI (www.tei-c.org) metadata set with input from Gibson*, Seaman*, and Unsworth*. The TEI framework extends beyond metadata to specific language tags. ComNet will also work to integrate these lower level TEI tags into the overall ComNet XML Schema. We will also work to bring the ComNet schema into accord with the developing GOLD (Farrar & Langendoen, 2003) ontology (linguistics-ontology.org). This ontology uses the SUMO upper ontology (www.ontologyportal.org) which itself is in accord with the WordNet (wordnet.princeton.edu) framework. GOLD has been most fully elaborated in regards to the features of morphosyntax that encode gender, number, person, case, evidence, evaluation, modality, tense, mood, force, size, aspect, polarity, and voice. This ontology allows us to elaborate the current XML schema to correspond to the decomposition produced automatically by the MOR program for the %mor (morphosyntax) line. In addition to use of OLAC, IMDI, WordNet, and GOLD, we will develop an additional set of metadata characterizers

important for the retrieval of materials from the web for particular research projects. For example, in the context of our work on UDL curation, we will develop a metadata set that maximizes the ability of researchers to locate appropriate machine-readable, freely accessible versions of texts over the web. Because few materials have been annotated with the necessary metadata, we will produce trainable methods for inducing these metadata characterizers from aspects of the documents and usage patterns over the web. This work will be conducted both in the context of UDL materials and for the video database described in Thrust #6.

Program interoperability: Our work on program interoperability is fairly advanced. We can now convert between every major display format in this field and the ComNet XML format. However, as ComNet moves into new areas, such as library collections and survey data, we will need to further extend interoperability by extending the ComNet XML schema and the various programming tools that rely upon it to integrate with TEI and other formats.

Thrust #4: Search and Visualization

This thrust will focus on improving methods for search, discovery, and analysis. Users will be able to access ComNet materials through a single entry port interface. We will build a powerful search engine that can take advantage of the structured, open XML databases we have created. Jamie Callan will extend the Lemur/Indri toolkit to configure a set of ComNet search programs that operate in a similar fashion both through browsers and as standalone applications. These programs will operate on all forms of ComNet data, including UDL books, the ClueWeb09 web corpus, and the many corpora of spoken language in ComNet XML. The ComNet searcher will support metadata-driven search, composition of regular expression searches, and direct playback of transcripts and media located through the search routines. The search program will implement all standard technologies for concordances, frequency counts, mean length of utterance, tagging, sequence analysis, etc. The searcher will also transmit data to standard statistical analysis in Excel, R, Matlab, and other programs for report generation and data visualization.

Callan recently created a dataset of 1 billion high PageRank web pages in the 10 major languages used on the web. This dataset, called ClueWeb09, is becoming a standard experimental dataset for research by well-equipped research labs; for example, 4 out of 7 of NIST's 2009 TREC tracks use this dataset (Entity detection; Million Query; Relevance Feedback; Web). For those labs, we provide the whole database on a set of terabyte drives. However, for the broader user community, ComNet will use the Lemur Toolkit's Indri search engine to provide interactive and batch search interfaces. Individuals will be able to upload metadata and/or text annotations, which will be added to the search engine index and then be available via the query language; this capability will promote sharing and reuse of metadata and text annotations for research and education. As compared with commercial search engine APIs such as Yahoo's BOSS or Google's API, this facility's advantages include a standard dataset, transparent text indexing, powerful query language, and transparent community-based metadata, all of which supports a broad range of reproducible research.

Tabular Display. Cieri and Strassel will extend the DASLTran tool developed for Talkbank so that it supports search and visualization of hierarchical and tabular data, for example, session and subject specific metadata. This new tool will also import and export format used within the quantitative sociolinguistic research community such as NCSLAAP (ncslaap.lib.ncsu.edu), Excel, Praat (www.fon.hum.uva.nl/praat) GoldVarb, and R (www.r-project.org).

Multilingual Access: We will provide multilingual localization for our search tools and other user facilities. This localization will extend both to major languages and to the smaller language communities involved in Thrust #7. We will configure our tools to make use of state-of-the-art MT facilities provided by Google at http://www.google.com/language_tools?hl=en. Searches can be structured either in the language of the target material or the user's language. Retrieved text can then be translated into the user's language, if requested. We will also store high-quality translations provided by users. Within the framework of UDL materials we will work to specify books that are translations and match these up to their originals, when they are available.

Automatic Analyses: ComNet will provide a vast database for the application and development of automatic language analysis. Specifically, programs capable of reading ComNet XML can automatically produce part of speech tagging, grammatical dependency analysis, and a variety of content analyses. The data are also linked to programs for in-depth phonological analysis through Phon and Praat. These analyses use the methods of computational linguistics to construct analyses that are useful for scientists throughout the behavioral sciences. Within ComNet, we plan to extend these analytic tools to include new systems for emotion analysis based on Pennebaker's* LIWC (Pennebaker, Francis, & Booth, 2001) and Wiebe* (1994, 2000). We will also work with CASL and others to develop data-mining methods for examining ComNet transcripts and video for specific activities or deceptions.

Support for Medical Discovery: Liberman will lead an effort to develop a system that extends previous work to enables content experts in medicine to build information extraction engines that operate on biomedical text. In the NSF sponsored "Mining the Bibliome", Liberman and his group created corpora of abstracts of biomedical text annotated syntactically and for biomedical entities, for example carcinogens, and the various ways they are named, mentioned, abbreviated and referenced. This data was then used to build and evaluate systems that automatically identified and classified such entities. The team that built these systems consisted of linguists, computer scientists, programmers and content experts. Under ComNet, Liberman will develop tools and best practices that will allow content experts to build extraction systems without the intervention of linguists and programmers. The system will include a wrapper for the Mallet program that isolates users from the need to write code and guides them through a series of decisions that lead to high performing extraction systems.

Visualization: To supplement our new tools for data analysis, we will build additional tools for data visualization. These tools will be built on top of the 3KSoft XML tools that are designed to help build quickly new methods of visualizing heterogeneous data. Let us mention two as illustrations: BungeeView and Commenter. BungeeView (bungeeview.com) allows the user to discover patterns in metadata across documents and videos. Commenter implements a system for creating collaborative commentary (MacWhinney et al., 2004) much like the Pea's* DIVER system (http://diver.stanford.edu/contact.html). Commenter provides both a browser frontend and a server database backend to link comments to the media and to each other. A prototype version of this system is linked to the prototype browser at talkbank.org/browser. The system will allow users to categorize and access comments by type, data, user group, and claim status.

Thrust #5: Speech Technology

Our work in speech technology will examine three areas: social linkage, dialog systems, and software for language communities.

Social Linkage. When collecting speech data, researchers often obtain only minimal information about the speakers – gender, age, and place of residence. However, more detailed work by Labov (2001) and others shows that the more we know about our informants, the more powerful our speech models (Eberhardt, forthcoming). If we collect information about the places speakers have lived, their jobs, what they do in their leisure time, etc. at the time of recording, then we may, using machine-learning techniques, automatically find groups whose speech is similar. The subset of speech data from such a group can be used to train a speech recognizer to obtain finer acoustic and language models. Subset selection will use techniques inspired by the work of Groves & Heeringa (2006).

LDC will configure a new corpus to bridge the gap between sociolinguistics and speech technologists working in language and automatic dialect identification (Cieri et al., 2008; Schwartz et al., 2007). The corpus will contain the audio and transcripts of approximately 1000 telephone calls selected from the already published Fisher English corpora. The calls will be stratified for age, sex, region and socioeconomic class and, as proof of concept, partially annotated for dialect features using the DASLTrans tool described in Thrust #5. For the cross-cultural comparisons that will be carried out by Malcah Yaeger-Dror and others, we also plan to combine data from interviews conducted in different cities with the AAVE and DLV (dominant local vernacular]) speakers matched for age/gender/etc as well as style. We will use the triaged Mixer and Story Corps/Griot data, which are already transcribed and aligned, to give us a small bootstrap representation of social groups that we can use for studies of both local vernaculars (such as Pittsburghese) and social dialects (such as AAVE).

The CallFriend corpus at the LDC has been partially transcribed into CHAT. To date there are balanced numbers of English, Spanish and Japanese corpora available for comparative study of discourse phenomena (Yaeger-Dror, Deckert, & Hall-Lew, 2003). Comparison of Japanese, Spanish, and English CallFriend and other corpora available from LDC and talk bank, have shown the degree to which pragmatic routines are not cross culturally translatable. For example, comparative studies of disagreement strategies in these three languages based on conversations in the corpus, have shown that the prosody of disagreement differs radically in these 3 cultures (Yaeger-Dror, Takano, Hall-Lew, & T., in press). With improved access to the CallFriend corpus, predisagreement structures in Japanese and English are being studied as well (Yaeger, Takano, Rinnert 2009). Such studies impact not only on cross-cultural pragmatic and conversational analytical comparisons but also the teaching of foreign languages (Kasper & Blum-Kulka, 1993) and intercultural communication (Yamada, 1992).

Dialog systems: The Let's Go dialog system answers the phone for the Port Authority of Allegheny County every night when human operators are not working. The system has collected 75,000 spoken dialogues, and it continues to collect data at a rate of about 1500 dialogues per month. As the system shifts to full-time operation, this number will increase tenfold. The recorded digital conversations have been automatically labeled with estimates of dialogue success and number of turns per dialogue. To support more fine-grained analyses of the type conducted by Raux & Eskenazi (2007) and Gravano et al. (2007), we need to convert this database to full ComNet format and index it using the Lemur toolkit. For example, if someone wants to conduct research on turntaking and needs to get data on dialogues that contain turns that were more than 10 seconds long and compare their success rates to dialogues with turns that were more than 20 seconds long, then the index should enable them to bring up that subset of waveforms and logfiles. We will create a web-based interface that allows the user to call up subsets of data and, for the automatic labels that we already have, or new ones that the user requests, produces a comparison using a variety of statistical tools.

Software for Language Communities: The third focus of this thrust will be on the development of speech technology to support the needs of language communities. For example, CMU has produced a system that allowed Edna MacLean*, our primary contact with the Iñupiaq community in Barrow, Alaska, to train an Iñupiaq text-to-speech system by reading a few passages to produce digitized audio files. We will also build linkages between dictionaries and speech to promote language learning and maintenance in communities. Similarly, we will develop localized computer interfaces that respond in the language of the speaker and which can translate using word-based text-rollover between the user's language and English or the other dominant language of the community.

Thrust #6: Multimodal Analysis

ComNet will provide unique opportunities for the development of new methods for analysis of the multimodal aspects of human communication, including the use of sign language. Many segments of the database will provide high quality transcriptions linked to linguistic, video, and content annotations. Using this rich annotational foundation, researchers will be able to align gestures and postures with changes in scenes, gaps in conversation, and many other structural features of conversational interactions and narratives. The multimodal analysis community currently relies on four major annotation tools: Anvil, ELAN, EXMaRLDA, and The Observer. (Earlier we looked at an example of a TalkBank transcript from a problem-based learning class in medical school displayed in both CLAN and ELAN.) Working with the developers of these systems, we have produced programs that manage the conversion of TalkBank data into each of the relevant formats for further detailed work.

Video Analysis Data: The video analysis community has long attempted to bridge the gap from low-level feature extraction to semantic understanding and retrieval of the communicated content. To solve this fundamental problem, we will create a large shared video database as a focused target for further analysis and evaluation. This shared database will include media, transcripts, screen text data, web text metadata, corpus metadata, shot segmentation, image features (Gabor texture, Grid Color Moment, and Edge Direction Histogram), local feature descriptors, motion features (kinetic energy, optical flow, MPEG motion vectors), audio features (FFT, SFFT, and MFCC), and characterizations of the data with the LSCOM concept ontology. Over 60 TRECVID participants have done this type of sharing of automatically extracted metadata for the non-public TRECVID collections. We will work with this community to create similar metadata for our open-access content.

Video Annotation Toolkit: To further bootstrap the process of annotation of the video test bank, we will provide a complete video annotation toolkit. This resource will allow researchers throughout the video community to annotate their own data, expand the concept ontology, and explore higher-level search services. We have already fielded several very effective systems, allowing annotators to efficiently label representative key frames in video, as well as longer video clips. We will make a robust version of this tool available to others early on in the project.

Video Analysis Toolkit: A number of researchers have expressed interest in applying our tools to their own data sets. Responding to this need, we will provide make key components of the Informedia library system available as open source, including shot detection, speech recognition, alignment, etc. This toolkit will also include modules for finding shots, labeling motions, and classifying content automatically. Using this suite, researchers can quickly customize tools, refine concept ontologies, and re-train classifiers for diverse applications. This allows further shared development of the software for analysis or services such as summarization, without having to expend many additional person-years in development.

Web-based Annotation: We will also make use of the cyberinfrastructure opportunity provided by the ESP web game developed by CMU researcher Luis van Ahn (van Ahn, Kedia, & Blum, 2006), to allow collaborative annotation of video on the web. For this task we again expect the undergraduate students as well as the high school students participating in the summer programs at CMU to contribute ideas and implement code. Since this generation is very much in tune with the characteristics of social networks of Web 2.0, we expect enthusiastic involvement in this project. This work builds on the notion of "Human Computation", whereby manual work is efficiently spread out over large numbers of people on the Internet, hereby providing innovative solutions for collaborative annotation of web video. Validation and verification of this annotation effort is done through duplication, which means annotations are only accepted after multiple people independently create the same label. This has been highly effective for image annotation and we will extend this annotation principle to our video collections.

Video Annotation Evaluation: Finally, to test the efficacy of new annotations against the video test bank, we will provide a benchmark set of tasks for video analysis evaluation.

Thrust #7: Language Communities

We will construct general methods for harvesting data from targeted language communities. We will work with four types of communities:
1. non-endangered U.S. minority languages (Spanish*, ASL*, Hawaiian, Navajo),
2. endangered U.S. minority languages (Iñupiaq*, Ojibwe*),
3. non-endangered non-U.S. languages (Mapudungan*, Welsh*, Aymara, Nahuatl*),
4. endangered non-U.S. minority languages (Atayal, Cree).

The importance of work in language preservation for endangered languages is widely recognized. Without intervention, more than half of the world's 7000 spoken languages are not expected to survive this century (Crystal, 2000), and many others, even those with millions of speakers, are struggling to maintain a state of stable bilingualism within a surrounding dominant language community. Although we will focus our work on languages spoken within the United States, the tools we construct in this thrust will help in the preservation of languages for language communities of all these types.

Our first language communities already have their own websites and most group members can access these sites. We will work with the developers of these sites to make them interoperable with ComNet tools and formats, while still making sure that communities maintain full control over their sites. The site will have resources that will allow the target communities to construct social networking systems, cultural documentation, and links to community activities – all in the local language. The social networking software will be developed using open source versions of Facebook, role-playing games, and virtual worlds. These materials will rely on methods for linking transcripts to media, so that speakers can engage in direct dialogs over the web, while still producing written records of their conversations. Our language communities will also use facilities like the ComNet Commenter system to engage in blogging and commentary in the local language with all data being stored on the servers. Apart from these social uses of the web, we will also emphasize tools that assist in language learning and maintenance. These include on-line dictionaries and methods for grammatical analysis, constructed using ComNet and GOLD frameworks. We will also deploy local versions of the various language learning software methods developed by MacWhinney and colleagues at http://talkbank.org/pslc.

From the materials constructed by these communities, we will develop corpora for each language formatted in ComNet XML. These corpora will then be used for the development of

orthographic and linguistic standardization, spelling checkers, on-line dictionaries, speech recognition and synthesis, information retrieval, telephone dialogue systems, and machine translation.

Proposed partner communities: Within each community, we have a lead contact who will work closely with us. For Ojibwe (aka Anishinaabemowin), we will work with Margaret Noori* who has already developed extensive social networking and game facilities for the Ojibwe community. These resources will serve as an excellent prototype and existence proof for our work in extending similar resources to other communities. Along similar lines, the Welsh community, Delyth Prys* has developed web facilities that emphasize language instruction and Welsh localization. Here, again, we will use these materials as a guide for our extensions of similar facilities to other language communities. We will also work with Jonathan Amith* who has collaborated with LDC to develop state-of-the-art linguistic software (dictionary, morphology, tagger, and corpus) for Nahuatl. We will use Amith's work as a model for our efforts in this direction. For ASL, our contact is Brandon Scates* who, as a hearing child of deaf adults (CODA), has spent much of his life working on activities in the ASL community. For the Hispanic community, we can rely on Rodolfo Vega and many other Hispanic linguists and sociologists. For Iñupiaq, we will work with Edna MacLean*, who has developed web-based systems and computational software. For Mapudungun (Huilliche), we will work with Pilar Alvarez* of Universidad de Los Lagos in Osorno, Chile. Once we have consolidated our work with these six communities, we will move on to work with Greenlandic*, Aymara*, Tzeltal, and Quechua. Throughout, we will coordinate our work with the DOBES* and E-MELD* Projects.

ComNet will organize user-group workshops in Pittsburgh that bring together organizers of web sites from each of our target communities with the goal of sharing experiences, methods, and software across language communities and sites.

Outreach Activities

Education and Training: ComNet educational outreach will focus on three areas: the internship program, week-long training workshops, and materials development.

Internship program. Maxine Eskenazi has directed a highly successful internship program for the NSF Pittsburgh Science of Learning Center (PSLC). This program has succeeded in attracting both majority and URM students to CMU for a two-month summer program involving research in faculty laboratories. We will emulate and extend this program for ComNet, coordinating with the successful IGERT program at the University of Pennsylvania.

Research and Training Workshops. Each summer, we will conduct a series of four one-week workshops. Participants in these workshops will include both graduate students and faculty. Three of these workshops will deal with topics relevant to the seven ComNet thrusts with the goal of utilizing and refining ComNet tools for specific analytic questions. We expect that each workshop will involve about twenty participants. The fourth workshop will focus on the training of new participants in the use of ComNet tools and exploration of the database. We will also link this training workshop to the European summer courses in transcription analysis (Mondada*).

Materials development. TalkBank data are now being used in science education in five domains (K/12, undergraduate, graduate, professional, and public). The data are used in Science Museums, classes in linguistics and psychology, public discussions of language policy, and professional work in speech and hearing. As we have done for the CHILDES data, we will

configure instructional materials that can help guide teachers in their use of ComNet data for learning about social processes.

Community Input: A core principle of ComNet is that long-term sustainability is grounded on research group community participation and input. The specific groups involved in ComNet are described in Appendix A5. Thrust #1 on Partnering (directed by MacWhinney) will conduct ongoing evaluation of input from the relevant communities. Currently, each of our primary research group communities uses a discussion mailing list or forum at googlegroups.com. We will continue this format, extending it to additional research communities and language communities.   We are particularly interested in the application of ComNet methods for the cultivation and preservation of language community resources. To this end, we will configure a Language Communities Advisory Board, as described in further detail in Appendix A5.

International Participation: TalkBank has forged a wide array of international collaborations and coordinated repositories with funded TalkBank projects in Taiwan, Israel, Denmark, Japan, England, Germany, France, and Hong Kong. ComNet will be even more internationalized than TalkBank, as we are now forming additional links with European groups such as CLARIN, CLAPI, DOBES, IDS, INSERM, and KTI. Our basic mission here is to collaborate and cooperate with all groups dealing with data on human communication in all countries, and to urge these groups to share and use ComNet data. Details regarding these links are given in Appendix A5 on Participation.

4.   Structure

Management. MacWhinney will serve as P.I. with Carbonell and Cieri as co-Directors. The Executive Committee (EC) will be responsible for all major decisions. It will be composed of the leaders of the seven thrusts, as described in Appendix 2. The EC will continually review work on the thrusts. A full-time Project Manager (TBA) will support the work of the Executive Committee, and a Lead Programmer will be responsible for integration of the cyberinfrastructure. Separate advisory boards will focus on (1) curation, (2) minority community outreach, (3) cyberinfrastructure, and (4) computational methods.

Users: ComNet researcher users will come from the NSF fields described in Appendix 5. Current TalkBank and LDC researcher communities include 12,000 scientists in six of these areas. We expect that 24,000 researchers will eventually rely on ComNet resources. Beyond this, we expect a wide base of users in the public schools, government, minority communities, and business.

Cyberinfrastructure: Our cyberinfrastructure relies on the expertise of the Schools of Computer Science at CMU and Penn, along with new grid/cloud computing infrastructure. We will have mirrors at LDC, Antwerp, Chukyo, and UDL sites in China, India, and Egypt. To visualize this linkage, consider that ComNet will integrate data and methods from the efforts depicted by their logos here, as well as from many others:

We will also include representatives of other funded DataNet Projects on our Cyberinfrastructure Advisory Board.

Expertise: The ComNet research team includes experts in computer science, database structures, cyberinfrastructure, information science, language technologies, neuroscience, speech analysis, computational linguistics, machine translation, human-computer interaction, and video analysis.

Diversity: The senior personnel include three women (Lori Levin, Gloriana St. Clair, and Maxine Eskenazi), two Hispanics (Carbonell, Vega), and one person with disabilities (Gloriana St. Clair – mobility impaired). ComNet has an extensive commitment to the development of resources for minority communities in the U.S, including American Sign Language (ASL), Spanish, and Native American languages. In the context of our four-week long summer schools, we will involve minority students in the scientific study of language, often in ways that can directly benefit their communities.

5. The Current Opportunity

Many of the components of ComNet are likely to receive NSF funding over the next decade. Work on spoken dialog systems, video summarization, aphasiology, animal communication, and metadata extraction will be continued with or without ComNet. Given this, why is it so important to devote scarce NSF cyberinfrastructure funds to this project? The reason is that only through ComNet can we obtain a massive, integrated, open-access database. We could pursue work on child language or legal argumentation without ComNet, but our research projects would be data-starved. Researchers would continue to be frustrated by the barriers of Licensing and Format Babel. ComNet will correct this situation by articulating clear standards for data collection, transcription, and analysis in each of its relevant domains. Without shared protocols and standards, we cannot produce data that leads smoothly to an accumulation of knowledge. Without ComNet, we will be trying to produce 21st century science with 20th century tools.

Without ComNet, the massive investments made by NSF and other agencies in the collection of raw data will fail to contribute fully to the cumulative progress of science. In areas such as endangered languages or studies of classroom learning of math and science, the amount of wasted, unshared data is staggering. The success of TalkBank projects in child language, aphasia, second language, and legal argumentation have shown that it is possible to stem the tide of this data loss. If NSF had had sufficient funds six years ago for the support of ComNet, we could have begun to move ahead and the fields affected by ComNet would already be producing higher quality science. It is crucial that we correct this situation.

ComNet occupies a unique position within the future of cyberinfrastructure. The data conveyed through communication streams are fundamental to all the Human Sciences, and the computational tools for studying communication are international. However, nations and regions will want to place specific focus on their local languages and cultures. Eventually, there will be cyberlinked Human Science Data Centers in China, Japan, Australia, Chile, the Gulf, and elsewhere. For this process of internationalization to develop smoothly, we must begin now with a systematic consolidation of ComNet in the American context and the integration of ComNet into the wider cyberinfrastructure for the Human Sciences. We should wait no longer; we owe it to science and society to move forward aggressively to open up this new frontier.