Converted 2011-09-01 11:21 from the Web using Citavi 3.1. Differences from the original page are possible.

Linguistic Annotation

This page describes tools and formats for creating and managing *linguistic annotations*. `Linguistic annotation' covers any descriptive or analytic notations applied to raw language data. The basic data may be in the form of time functions -- audio, video and/or physiological recordings -- or it may be textual. The added notations may include transcriptions of all sorts (from phonetic features to discourse structures), part-of-speech and sense tagging, syntactic analysis, "named entity" identification, co-reference annotation, and so on. The focus is on tools which have been widely used for constructing annotated linguistic databases, and on the formats commonly adopted by such tools and databases. This page began as a set of links to systems for speech annotation, and the coverage of textual annotation is still inadequate.

This page is no longer being actively maintained.

Related pages: Open Language Archives Community, Linguistic Exploration, Gesture Annotation, Italian version of this page by Piero Cosi, Speech Annotation and Corpus Tools

This page is the home of the COCOSDA technical topic domain Corpus Annotation Tools.

This page has been prepared in conjunction with our research on the logical structure of linguistic annotation, based on annotation graphs.

IRCS Workshop on Linguistic Databases [Dec 2001]

Recently added links and updated descriptions are marked with a *.

- Steven Bird, Mark Liberman, LDC

Index

Alembic	ATLAS	CA	CES	CHILDES	CLinkA	COCOSDA	CSAE
CSLU	DAISY	DAMSL	Delta	DRI	EAGLES	Emu	Festival
FSA's	GATE	Gsearch	HIAT	Hyperlex	Intex	ISIP	ISLE
LACITO	LDC	LT XML	MATE	MICASE	MPEG	MPI	Multitext
NEGRA	NITE	Observer	Partitur	Praat	SABLE	SAMPA	SGREP
SignStream	SIL	SLAM	SMDL	SNACK	SUSANNE	TalkBank	TASX*
TEI	Tipster	Transcriber	TransTool	Treebank	TSNLP	TUSNELDA	Unicode
Verbmobil	vPrism						

Key

- **F:** a systematically-documented annotation *format*
- **T:** an available *tool* for creation, display or search
- **D:** a tool is *downloadable*
- **P:** there is a citeable *paper* which documents the format/system
- **R:** other kinds of *resource*, such as books and associations
- **C:** methods and standards for transcribing *content*
- **[U/W/M]:** tools run on Unix/Windows/Macintosh

Linguistic Resources

DT [UW]

Alembic Workbench (David Day)

Alembic Workbench is an SGML-based annotation system. Apart from the usual kinds of textual annotations, the workbench enables various kinds of specialized annotations including co-reference annotations (cf. the Message Understanding Conference markup rules), various kinds of user-defined inter-tag pointers, and (shortly) general template annotation (aka relations, frames, or events). The Alembic multi-lingual NLP system provides access to taggers for a wide variety of extraction levels, and applications have now been built for several languages. The software has a sophisticated visualisation component. It runs on Sun workstations and is freely distributed.

FTD

LACITO Linguistic Data Archiving Project (Boyd Michailovsky, John B. Lowe, Michel Jacobson)

Projet Archivage, based at LACITO in Paris, aims to provide tools and formats for linguistic and anthropological field data. An interesting feature is the use of XML markup, with a DTD that supports transcriptions, phrasal and word-by-word interlinear translations, and audio references. Some XSL style sheets are provided that illustrate the potential power of XML markup to support web browsing for material of this type, giving access to text and sound.

FP

ATLAS - Architecture and Tools for Linguistic Analysis Systems (Steven Bird, David Day, John Garofolo)

ATLAS is a joint initiative of NIST, MITRE and the LDC to build a general purpose annotation architecture and a data interchange format. The starting point is the annotation graph model, with some significant generalizations. An LREC paper describes the model.

P

CA - Conversational Analysis

This page of transcriptions by Emanuel Schegloff exemplifies the style of transcription traditional among researchers working on conversational analysis.

FC

CES (Nancy Ide, Greg Priest-Dorman,Patrice Bonhomme)
The Corpus Encoding Standard (CES) is a part of the EAGLES Guidelines developed for language engineering research and applications. CES is an SGML-based, TEI-conformant specification of a minimal encoding level that corpora must achieve to be considered standardized in terms of descriptive representation (marking of structural and typographic information) as well as general architecture (so as to be maximally suited for use in a text database). It also provides encoding specifications for linguistic annotation, together with a data architecture for linguistic corpora. A section of CES on speech annotation (part 6) is under construction. Projects using CES are listed here. An XML version of CES called XCES is under development.

FTDPRC [WM]

CHILDES (Brian MacWhinney, Steven Gillis)
The CHILDES project provides a large database of first and second language acquisition data from over 30 languages in a constant format, called CHAT. There are also programs for Windows and Macintosh that permit analysis of this database as well as alignment of text to speech and video.

T

CLinkA: Coreferential Links Annotator (Constantin Orasan)
CLinkA is a tool for annotating coreference, written in Java. It is one of several tools for working on anaphora being developed by the Computational Linguistics Group at Wolverhampton

R

COCOSDA (Nick Campbell)
The International Committee for the Co-ordination and Standardisation of Speech Databases and Assesment Techniques for Speech Input/Output, COCOSDA, has been established to encourage and promote international interaction and cooperation in the foundation areas of Spoken Language Processing.

TDC [W]

CSAE (John W. DuBois)
The Corpus of Spoken American English (CSAE) project at UCSB has developed and released several tools, including VoiceWalker, a transcription tool for audio and video, and SoundWriter, which permits aligning portions of transcripts with sound files via SMPTE time codes. They have also developed their own set of transcription conventions.

TDPRC [W]

CSLU (Kal Shobaki, Jacques De Villiers)
The CSLU Speech Toolkit, developed at the Center for Spoken Language Understanding (CSLU) has a complete set of free tools for collecting and transcribing speech. It contains an interactive speech display program (Speech View) which allows the user to align transcripts with the sound files. The toolkit also contains a complete course in spectrogram reading and acoustic phonetics, a speech recognition engine, a speech synthesizer based on the Festival architecture, a facial animation component, and an integration tool for authoring your own spoken language system. CSLU also has an ascii encoding for phonetic transcriptions (large postscript file).

FTP [UW]

DAISY Consortium (Ingar Beckman Hirschfeldt)
The DAISY Consortium is a worldwide coalition of libraries and institutions serving print disabled persons, developing the open standards, tools, and techniques for the next generation of "digital talking books" (DTB). Standard 2.0 is available. Information about the process can be found in a CSUN98 presentation.

FTRC [UW]

DAMSL / SWBD-DAMSL (James Allen, Dan Jurafsky)
DAMSL - Dialog Act Markup in Several Layers - `defines a set of primitive communicative actions that can be used to analyze dialogs'. The structure of DAMSL is the result of work by the Multiparty Discourse Group at the DRI meetings. SWBD-DAMSL was "designed as an augmentation to the [DAMSL] tag-set", for the purpose of annotating (a portion of) the Switchboard corpus. A page of links to online papers is available here. Some dialogues in the TRAINS Project have been marked up using DAMSL annotation - pointers here. There is also a dialog annotation tool called dat.

TP [UW]

Delta (Susan Hertz)
Eloquent Technologies has developed a text-to-speech toolkit which synthesizes speech by means of a multi-tiered representation of the text called a Delta.

R

DRI (Susann Luperfoy)
The Discourse Resource Initiative is an `international grass-roots effort that seeks to share corpora that have been tagged with the core features of interest to the discourse community.' DRI brings together information on documents, corpora and software associated with discourse research, including a list of links to annotation tools. DRI includes a Multiparty Discourse Group. The COCONUT project (COoperative, COordinated Natural language UTterances) has adapted the DRI scheme in order to deal with cooperative planning environments, and there is a page of pointers to online papers.

FR

EAGLES Spoken Language Working Group (Dafydd Gibbon)
The Spoken Language Working Group of EAGLES (the European Commission's Expert Advisory Group on Language Engineering Standards) has produced a *Handbook of Standards and Resources for Spoken Language Systems*, "with information and recommendations on established practice in the area of spoken language system development in multilingual environments." The EAGLES project also comprised working groups on working standards for corpora, lexicons, formalisms and evaluation. The EAGLES-II project included further work on semantic annotation and dialogue annotation.

FTDP [UW]

Emu (Steve Cassidy, Jonathan Harrington)
The Emu system offers consistent access to diverse speech databases, with facilities for easy extraction of statistics, and support for database creation as well. Emu permits complex multi-tiered and hierarchical structures, and these can be built using a combination of manual and automatic annotation. An Emu script has been written to import certain Partitur files, lending support to Steve Cassidy's contention that (in terms of annotations proper, leaving out background information about recording, speaker etc.)

the Emu framework is able to express the information content of Partitur annotations.

TD [U]

Festival (Paul Taylor)
The Festival Speech Architecture from CSTR at Edinburgh, which was originally designed for use in speech synthesis, has been generalized and applied to database analysis as well. Festival uses Heterogeneous Relation Graphs for representing linguistic information.

TD

FSA: Finite State Automata
There has been development of concepts and toolkits for n-ary finite automata, which might provide a useful model for expression, creation and search of multidimensional linguistic annotations. We can cite the following toolkits: LADL, Xerox, AT&T, van Noord and Grail though as far as we know, these have not been used for general manipulation of speech annotations, and may not be suitable in detail for this purpose.

FTDP [U]

GATE (Hamish Cunningham, Kevin Humphreys)
GATE has an implementation of the Tipster architecture, plus graphical tools for data visualisation, annotation, evaluation and process control. It is distributed free for research and with Information Extraction software.

T [U]

Gsearch (Frank Keller)
Gsearch is a tool for searching tagged corpora. Queries are formulated in two stages. First, the user specifies a context-free grammar which is used to parse a given corpus and to convert its tags into a standard set. Second, a search expression uses the words, terminals and non-terminals provided by the corpus and grammar. Structured output may be visualized with Ratnaparkhi's Viewtree program or Calder's Thistle system. Gsearch currently works with BNC, Brown, SUSANNE, WSJ, Frankfurter Rundschau, and NEGRA.

FTDPRC [W]

HIAT (Konrad Ehlich, Jochen Rehbein)
HIAT is a transcription system based on a score notation, developed in the 1970's by Ehlich and Rehbein, and widely used in Europe. The acronym stands for Halbinterpretative Arbeitstranskriptionen, or "semi-interpretative working transcription." Dafydd Gibbon is credited with the English name "Heuristic Interpretative Auditory Transcription," which preserves the acronym. The HIAT philosophy includes the notion of *literary transcription* (*literarische Umschrift*), which "involves systematic departures from the standard orthographic rendering of an item but in a manner that is meaningful to someone familiar with the orthographic system as a whole." Methods are provided for annotating prosody, non-verbal communication, and so on. PC and Mac software is available. An English-language description can be found in Edwards and Lampert, Talking Data.

TP [U]

Hyperlex (Steven Bird)
Steven Bird's Hyperlex system, developed in support of a field project, provides HTML-mediated access to a lexicon, speech recordings and paradigmatic catalogues for several languages. Steven plans to produce a portable version that can easily be adapted to new languages and new projects.

F [UWM]

INTEX (Max Silberztein)
INTEX is a linguistic development environment that includes large-coverage dictionaries and grammars, and parses texts of several million words in real time. INTEX includes tools to create and maintain large-coverage lexical resources, as well as morphological and syntactic grammars. INTEX can build lemmatized concordances and indices of large texts with respect to all types of Finite State patterns. INTEX is used as an information retrieval system, to analyze literary texts, to quantify language variations, to teach second languages, as a terminological extractor, etc. Large coverage linguistic resources are already available for English, French, German, Greek, Italian, Polish, Portuguese.

ISLE (Antonio Zampolli)
The ISLE project is funded for two years by the NSF and the EC, as part of a joint program Multilingual Information Access and Management. This program is `intended to further the knowledge required to build information systems that operate in multiple languages; to provide the technologies required for their application in a number of social and organizational contexts; and to demonstrate the validity of the approaches chosen.' An ISLE workshop was held at LREC-2, on Meta-Descriptions and Annotation Schemas for Multimodal/Multimedia Language Resources. US-ISLE includes a Spoken Language Group. [NSF award]

TDP [U]

ISIP (Joe Picone)
Joe Picone and others at the Institute for Signal and Information Processing (ISIP) at Mississippi State have produced nice freeware tools initially optimized for segmenting, transcribing and annotating telephone conversations: Segmenter, Transcriber.

FTDPRC

LDC (David Graff, Chris Cieri, Mark Liberman)
The Linguistic Data Consortium has developed a range of (mainly SGML-based) formats for transcripts and other types of annotation that it has published (See below for NIST's UTF format, which provides a combined framework for several of these existing formats). Some online documentation is available for individual corpora authored at different times by different groups, e.g. Switchboard at TI in 1991, Trains at Rochester in 1992-3, etc, as well as a general SGML transcription specification currently used for (orthographic) transcription of telephone conversations and broadcast news recordings. The LDC has also implemented a general data model for searching annotated text and speech corpora online, via LDC-Online.

T [UW]

LT XML ()
sggrep - an XML-aware grep tool.

FT

MATE (Laila Dybkjaer)
The multi-partner EC-funded MATE project (Telematics Project LE4-8370) aims to develop an SGML-based standard for annotating spoken dialogue corpora, and tools to "make the processes of knowledge acquisition and extraction more efficient." Deliverable D1.1 surveys a wide variety of annotation schemes, and includes a useful page of links. MATE implements Thompson and McKelvie's notion of standoff markup, which is related to the CES and TEI proposals.

TDC [W]

MICASE (Rita Simpson)
The Michigan Corpus of Academic Spoken English (MICASE) is an ongoing project to record and transcribe a diverse range of academic speech, based at the English Language Institute of the University of Michigan. To support transcription work, a freeware Windows 95 program has been developed, called SoundScriber.

FPR

MPEG (Leonardo Chiariglione)
MPEG-4, established as an ISO/IEC standard in early 1999, provides standards for streaming interactive multimedia. More specifically, it aims to support production, distribution and content access for digital television, interactive graphics, and interactive multimedia. It provides representations for basic "media objects," which might be recorded or synthetic; it describes the composition of compound media objects, or "scenes"; it provides for multiplexing and synchronizing of such data for network transport; and it offers standards for interaction with the audiovisual scenes generated at the receiver. MPEG-4's media objects include text, graphics, talking synthetic heads and associated text, synthetic sounds, still images, video and audio elements. It supports complex combination of these elements into time-varying scenes, and also provides for streaming of the underlying data, and interaction with the receiver. Recently, MPEG-4 has been partly integrated with Apple's Quicktime file format, using QuickTime as "the starting point for the development of a unified digital media storage format for the MPEG-4 specification." How much of MPEG-4's ambitious program has been fully defined or implemented remains unclear.

MPEG-7 is `the content representation standard for multimedia information search, filtering, management and processing, (to be approved July 2001).' Draper discusses the need for multiple non-commensurable structures for video annotation in his paper MPEG-7 and IR.

FT [UWM]

Linguistic Applications at MPI (Peter Wittenburg)
The MPI Language and Cognition Group have produced a variety of tools for working with annotated speech and video data. EUDICO (European Distributed Corpora Project) is a universal workbench for corpus linguistics, written in Java and operating with a variety of formats, including CHAT, Shoebox, Tipster, and various relational database formats. CAVA (Computer Assisted Video Analysis) is a suite of programs intended for scientists in the humanities, including a Transcription Editor for digital transcription of analog video on PCs, and a Macintosh program called MediaTagger, for creating and searching multi-tier annotation of digital video in QuickTime format. A new multi-platform tool for managing CHAT-like transcriptions and aligned speech data, the Spoken Childes Tool, is available. Many of these tools are available free to academic sites.

F

Multitext Project (Gord Cormack, Forbes Burkowski, Charlie Clarke)
The MultiText Project is concerned with developing techniques for the indexing and retrieval of massive text corpora. Queries can refer to document structure, even though this is annotated in different ways in different corpora.

FTPC [U]

NEGRA Corpus (Thorsten Brants)
The NEGRA corpus consists of approximately 10,000 sentences of German newspaper text. The corpus is a type of treebank, but with a novel annotation scheme for discontinuous constituents. An example tree showing the visual format, the annotation format, and the Penn Treebank equivalent, is available here. Annotate is a sophisticated tool which supports human-machine collaboration on the construction of syntactic trees.

NITE: Natural Interactivity Tools Engineering (Laila Dybkjaer)
NITE aims to build a "standard-setting workbench for multi-level, cross-level and cross-modality annotation, retrieval and practical exploitation of multi-party human-human and human-machine natural interactive dialogue data."

T [W]

The Observer (Lucas Noldus)
`The Observer' is a commercial tool for classifying and logging events. In the video version of the tool, one can create time-aligned annotations of video recordings, using the Event Recorder. The temporal patterning of the observations can be displayed using the time-event plot, and various summary statistics can be generated. The software was developed by Noldus IT and runs on MS Windows platforms.

FT

Partitur (Florian Schiel, Christoph Draxler)
The Bavarian Archive of Speech Signals has created the Partitur format based on their experience with a variety of speech databases. The aim has been to create ``an open (that is extensible), robust format to represent results from many different research labs in a common source."

TD [UWM]

Praat (Paul Boersma)
The Praat system offers a variety of nice tools for interacting with speech data, including tools for transcribing and annotating on multiple tiers.

FP

SABLE (Andrew Hunt, Richard Sproat, Paul Taylor)
The SABLE standard for annotation of linguistic properties of speech synthesis input necessarily shares a lot of characteristics with systems for linguistic annotation of naturally produced speech.

C

SAMPA (John Wells)
SAMPA (Speech Assessment Methods Phonetic Alphabet) is a machine-readable ASCII transliteration of the International Phonetic Alphabet. Originally developed by phoneticians to code six European languages, it is currently being extended to cover many more languages. SAMPROSA is an extension for transcribing prosodic information, and XSAMPA is an extension which covers every symbol on the IPA chart, in principle allowing transcription of all the world's languages. See also the CSLU Worldbet.

TDP [UW]

SGREP (Jani Jaakkola, Pekka Kilpeläinen)
SGREP (structured grep) is a tool for searching and indexing text, SGML,XML and HTML files and filtering text streams using structural criteria. The data model of sgrep is based on regions, which are nonempty substrings of text. Regions are typically occurrences of constant strings, SGML-tags, or meaningful text elements, which are recognizable through some delimiting strings

	or the builtin SGML, XML and HTML parser. Regions can be arbitrarily long, arbitrarily overlapping, and arbitrarily nested. There is also a paper which would be useful for anyone wishing to use SGREP.
TDP [M]	**SignStream (Carol Neidle, Dawn MacLaughlin)** The SignStream project aims to develop a database tool for transcription and analysis of video-based language data (particularly, signed language data). SignStream allows a user to enter data into any number of user-definable fields, where each datum is associated with a start and end frame of a video. Although a SignStream database is stored in a non-readable, binary format, the program includes a text export feature. However, there is no import feature. The program is currently being distributed at cost to researchers, educators, and students.
TDPF [WM]	**SIL (Larry Hayashi, Gary Simons, Terry Gibbs)** The Summer Institute of Linguistics (SIL) has enormous experience in providing tools and data formats for use in primary linguistic description. LinguaLinks is "an electronic productivity support system for language workers," based on the CELLAR object-oriented "Computing Environment for Linguistic, Literary and Anthropological Research.", including linguistics tools. Other SIL software tools are Speech Analyser and Speech Manager, Windows programs for labelling speech files and for searching a database of labelled speech files; Shoebox, for interlinear text annotation and its predecessor, IT. SIL also has an SGML based annotation format named PTEXT ("parsed text") described in this paper.
TDP [W]	**SLAM (Piero Cosi)** The `Segmentation and Labelling Automatic Module' (SLAM) is a tool for semi-automatic segmentation and labelling of speech signals. The tool was developed at the Institute of Phonetics and Dialectology, and runs under MS Windows.
P	**SMDL (Stephen R. Mounce)** The annotation of instrumental and vocal music has some interesting similarities with the annotation of speech. See the Music Encoding Standards page for general pointers. There is a proposed standard, the Standard Music Description Language (SMDL), and the proposal is available as well.
TDP [UWM]	**Snack (Kåre Sjölander)** Snack is a general toolkit to handle acoustic data with an emphasis on speech. It features real-time visualization, supports many file-formats, and is extensible. A speech signal viewer and phonetic label editor is included in the package. There is also a Snack "plug-in" for web browsers. Wavesurfer is a tool based on Snack.
CP	**SUSANNE (Geoffrey Sampson)** The SUSANNE annotation scheme provides a detailed encoding of the logical and surface grammar of English. The SUSANNE corpus/treebank, which is freely available, contains a subset of the Brown corpus that has been marked up according to this scheme. The CHRISTINE project sets out to expand the SUSANNE analytic scheme and corpus to cover spoken English, and particularly spontaneous, informal spoken English.
R	**TalkBank (Brian MacWhinney, Steven Bird, Mark Liberman, Peter Buneman)** TalkBank is a new research project whose goal is to foster fundamental research in the study of human and animal communication by providing standards and tools for creating, searching, and publishing primary linguistic materials on the Internet. The project grows out of the experience with CHILDES and LDC corpora, and aims to support the methodologies and notations of a broad range of disciplines including: classroom discourse, conversation analysis, discourse, language acquisition, gesture, signed languages, ethology, anthropology, field linguistics, speech analysis. TalkBank is using the ATLAS model.
TD [UWM]	**TASX: Time Aligned Signal data eXchange format (Jan-Torsten Milde)** TASX provides a general framework for creating and managing corpora, including XML-based annotation of the multimodal data, transformation of non XML-annotations, and web-based analysis and dissemination of the data. TASX-annotator is a user-friendly program for multilevel annotation and transcription of (multi-channel) video and audio data. TASX-annotator runs under WinXX (98, XP, 2000), Linux, Solaris and MacOS, and is distributed under an open source license.
F	**TEI (Lou Burnard)** The Text Encoding Initiative (TEI) published its first detailed recommendations for the encoding and transcription of all manner of written and spoken materials, using an extensible SGML framework, in 1994. This was widely used by many academic research projects and digitization initiatives (examples) and was also very influential in many corpus creation projects (e.g. BNC, Parole, Multex, Silfide) as well as in defining EAGLES and other standards. A membership consortium was formed in 2000 to continue maintenance and development of the TEI, and the most recent recommendations (P5) are expressed using a modular XML framework, which supports most kinds of linguistic annotation.
F	**Tipster (Ralph Grishman, Robert Gaizauskas, Hamish Cunningham, Remi Zajac)** This concerns annotation of text rather than speech, but has several interesting properties.
TDP [UWM]	**Transcriber (Claude Barras, Edouard Geoffrois)** Transcriber is free software for transcribing and annotating digital audio, aimed initially at transcription of broadcast news data. Its user interface is written in Tcl/Tk. It uses the same transcription formats as the LDC's Broadcast News data, and has also been adapted for XML I/O. It was developed by Claude Barras and Edouard Geoffrois, at DGA in Paris, in collaboration with LDC. A new version of Transcriber will be based on the annotation graph model, and the plans are described in an LREC paper.
TD [UW]	**TransTool (Jens Allwood, Elisabeth Ahlsén, Joakim Nivre)** The Swedish Spoken Language Corpus, developed at the Department of Linguistics, Göteborg University, has several interesting tools: Transtool, to aid in transcription; Synchtool, for synchronizing transcriptions with audio and video files; TRASA, a tool for automatic analysis of the corpus; and TRACTOR, a tool to support coding.

C	**Treebank (Mitch Marcus, Ann Taylor)** The Penn Treebank Project has produced semantic and syntactic annotations of naturally-occuring text for the Wall Street Journal, Brown, ATIS and Switchboard Corpora. The annotations produced by the Treebank project were published by LDC. Treebank has two query languages: tgrep (at LDC-Online) and CorpusSearch. The principle advantage of tgrep is its speed, and of CorpusSearch is its ability to pipeline queries together. Chris Brew has recently developed an extensible visualisation tool to aid treebank exploration, called TreeStyle. See also the NEGRA Corpus. Douglas Rohde has developed a more powerful version of tgrep called tgrep2. Treebanks for other languages are in development, including: German, Turkish, Polish, Czech, Portuguese, Bulgarian, Chinese, ...
FT	**TSNLP (Klaus Netter, Doug Arnold, Stephan Oepen)** `Test Suites for Natural Language Processing' is a European consortium providing NLP test suite technology and test suite fragments for German, French and English. For a description of the TSNLP annotation scheme, see their WP2.2. See section 5 for examples. The annotation scheme includes sentential information (such as a well-formedness score) and also word and sub-string annotation which is stored in tabular form (for analytical information including syntactic constituents and error descriptions). There is a web interface to the test suites. The database schema is described in the user manual, volume 2, section 4. **TUSNELDA: Tübingen Collection of Reusable, Empirical, Linguistic Data Structures (Andreas Wagner)** SFB-441, a new research program at the University of Tübingen, is collecting and developing data and annotation resources called TUSNELDA. A corpus annotation standard for TUSNELDA is under development.
RC	**Unicode (info@unicode.org)** The Unicode Consortium brings together software corporations and researchers at the leading edge of standardizing international character encoding. The outcome of this cooperation is The Unicode Standard, which provides the foundation for internationalization and localization of software. Unicode has a conference series and an FAQ. There are character charts, including one for IPA extensions.
FC	**Verbmobil (Wolfgang Wahlster, Reinhard Karger)** Verbmobil is a large German speech-to-speech translation project for the domains of appointment negotiation, travel planning and hotel reservation. The Verbmobil annotation project includes orthography, segmental annotation (with BAS Partitur) prosody (German ToBI), morphological and POS tagging, semantic annotation and dialog act annotation. Dafydd Gibbon has developed a model for sharing lexical databases in Verbmobil, called HyprLex.
T [M]	**vPrism (vPrism@lessonlab.com)** vPrism is commercial macintosh software for time-aligned annotation and coding of video, intended for use in educational and behavioral research.

Last update: 5 December 2001
Steven Bird, sb@ldc.upenn.edu