Studying Discourse and Dialogue with SIDGrid*

Gina-Anne Levow
Department of Computer Science
University of Chicago
Chicago, IL 60611, USA
levow@cs.uchicago.edu

Abstract

Teaching Computational Linguistics is inherently multi-disciplinary and frequently poses challenges and provides opportunities in teaching to a student body with diverse educational backgrounds and goals. This paper describes the use of a computational environment (SIDGrid) that facilitates interdisciplinary instruction by providing support for students with little computational background as well as extending the scale of projects accessible to students with more advanced computational skills. The environment facilitates the use of hands-on exercises and is being applied to interdisciplinary instruction in Discourse and Dialogue.

1 Introduction

Teaching Computational Linguistics poses many challenges but also provides many opportunities. Students in Computational Linguistics courses come from diverse academic backgrounds, including computer science, linguistics, and psychology. The students enter with differing experience and exposure to programming, computational and mathematical models, and linguistic, psycholinguistic and sociolinguistic theories that inform the practice and study of computational linguistics. However, studying in a common class provides students with the opportunity to gain exposure to diverse perspectives on their research problems and to apply computational tools and techniques to expand the range and scope of problems they can investigate.

While there are many facets of these instructional challenges that must be addressed to support a successful course with a multi-disciplinary class and perspective, this paper focuses on the use and development of a computational environment to support laboratory exercises for students from diverse backgrounds. The framework aims to facilitate collaborative projects, reduce barriers of entry for students with little prior computational experience, and to provide access to large-scale distributed processing resources for students with greater computational expertise to expand the scope and scale of their projects and exercises.

Specifically, we exploit the Social Informatics Data Grid (SIDGrid) framework developed as part of the NSF-funded Cyberinfrastructure project, "Cyberinfrastructure for Collaborative Research in the Social and Behavioral Sciences (PI: Stevens)", to support hands-on annotation and analysis exercises in a computational linguistics course focused on discourse and dialogue. We begin by describing the SIDGrid framework for annotation, archiving, and analysis of multi-modal, multi-measure data. We then describe the course setting and the application of SIDGrid functionality to expand exercise and project possibilities. Finally, we discuss the impact of this framework for multi-disciplinary instruction in computational linguistics as well as the limitations of the current implementation of the framework.

*The work is supported by a University of Chicago Academic Technology Innovation Grant.

Proceedings of the Third Workshop on Issues in Teaching Computational Linguistics (TeachCL-08), pages 106–113,
Columbus, Ohio, USA, June 2008.

2 SIDGrid Framework

2.1 Motivation

Recent research programs in multi-modal environments, including understanding and analysis of multi-party meeting data and oral history recording projects, have created an explosion of multi-modal data sets, including video and audio recordings, transcripts and other annotations, and increased interest in annotation and analysis of such data. A number of systems have been developed to manage and support annotation of multi-modal data, including Annotation Graphs (Bird and Liberman, 2001), Exmeralda (Schmidt, 2004), NITE XML Toolkit (Carletta et al., 2003), Multitool (Allwood et al., 2001), Anvil (Kipp, 2001), and Elan (Wittenburg et al., 2006). The Social Informatics Data Grid (SIDGrid), developed under the NSF Cyberinfrastructure Program, aims to extend the capabilities of such systems by focusing on support for large-scale, extensible distributed data annotation, sharing, and analysis. The system is open-source and multi-platform and based on existing open-source software and standards. The system greatly eases the integration of annotation with analysis though user-defined functions both on the client-side for data exploration and on the TeraGrid for large-scale distributed data processing. A web-accessible repository supports data search, sharing, and distributed annotation. While the framework is general, analysis of spoken and multi-modal discourse and dialogue data is a primary application.

The details of the system are presented below. Sections 2.2, 2.3, and 2.4 describe the annotation client, the web-accessible data repository, and the portal to the TeraGrid, respectively, as shown in Figure 1 below.

2.2 The SIDGrid Client

The SIDGrid client provides an interactive multi-modal annotation interface, building on the open-source ELAN annotation tool from the Max Planck Institute[1]. A screenshot appears in Figure 2. ELAN supports display and synchronized playback of multiple video files, audio files, and arbitrarily many annotation "tiers" in its "music-score"-style graphical interface. The annotations are assumed to be time-aligned intervals with, typically, text content; the system leverages Unicode to provide multilingual support. Time series such as pitch tracks or motion capture data can be displayed synchronously. The user may interactively add, edit, and do simple search in annotations. For example, in multi-modal multi-party spoken data, annotation tiers corresponding to aligned text transcriptions, head nods, pause, gesture, and reference can be created.

The client expands on this functionality by allowing the application of user-defined analysis programs to media, time series, and annotations associated with the current project, such as a conversation, to yield time series files or annotation tiers displayed in the client interface. Any program with a command-line or scriptable interface installed on the user's system may be added to a pull-down list for invocation. For example, to support a prosodic

[1] http://www.mpi.nl/tools/elan.html

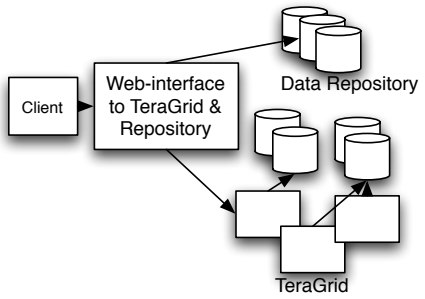

Figure 1: System Architecture

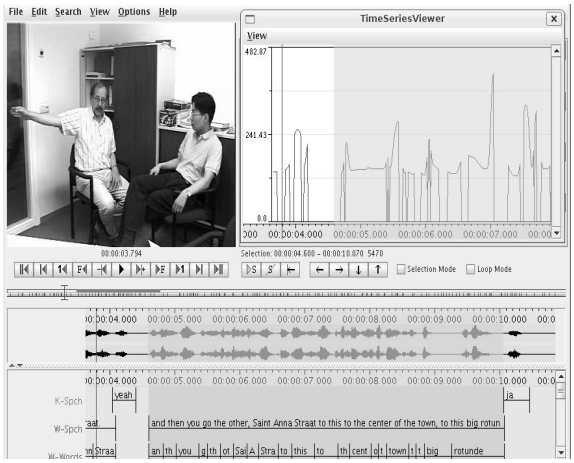

Figure 2: Screenshot of the annotation client interface, with video, time-aligned textual annotations, and time series displays.

analysis of multi-party meeting data, the user can select a Praat (Boersma, 2001) script to perform pitch or intensity tracking. Also, the client provides integrated import and export capabilities for the central repository. New and updated experiments and annotations may be uploaded directly to the archive from within the client interface. Existing experiments may be loaded from local disk or downloaded from the repository for additional annotation.

2.3 The SIDGrid Repository

The SIDGrid repository provides a web-accessible, central archive of multi-modal data, annotations, and analyses. This archive facilitates distributed annotation efforts by multiple researchers working on a common data set by allowing shared storage and access to annotations, while keeping a history of updates to the shared data, annotations, and analysis.

The browser-based interface to the archive allows the user to browse or search the on-line data collection by media type, tags, project identifier, and group or owner. Once selected, all or part of any experiment may be downloaded. In addition to lists of experiment names or thumbnail images, the web interface also provides a streaming preview of the selected media and annotations, allowing verification prior to download. (Figure 3)

All data is stored in a MySQL database. Annotation tiers are converted to an internal time-span based representation, while media and time series files are linked in unanalyzed. This format allows generation of ELAN format files for download to the client tool without regard to the original source form of the annotation file. The database structure further enables the potential for flexible search of the stored annotations both within and across multiple annotation types.

2.4 The TeraGrid Portal

The large-scale multimedia data collected for multimodal research poses significant computational challenges. Signal processing of gigabytes of media files requires processing horsepower that may strain many local sites, as do approaches such as multi-dimensional scaling for semantic analysis and topic segmentation. To enable users to more effectively exploit this data, the SIDGrid provides a portal to the TeraGrid (Pennington, 2002), the largest distributed cyberinfrastructure for open scientific research, which uses high-speed network connections to link high performance computers and large scale data stores distributed across the United States. While the TeraGrid has been exploited within the astronomy and physics communities, it has been little used by the computational linguistics community.

The SIDGrid portal to the TeraGrid allows large-scale experimentation by providing access to large-scale distributed processing clusters to enable parallel processing on very high capacity servers. The SIDGrid portal to the TeraGrid allows the user to specify a set of files in the repository and a program or programs to run on them on the Grid-based resources. Once a program is installed on the Grid, the processing can be distributed automatically to different TeraGrid nodes. Software supports arbitrarily complex workflow specifications, but the current SIDGrid interface provides simple support for high degrees of data-parallel processing, as well as a graphical display indicating the progress of the distributed program execution, as shown in Figure 4. The results are then reintegrated with the original experiments in the on-line repository. Currently installed programs support distributed acoustic analysis using Praat, statistical analysis using R, and matrix computations using Matlab and Octave.

2.5 Software Availability

The client software is freely available. Access to the public portion of the repository is possible through the project website at `https://sidgrid.ci.uchicago.edu`; full access to the repository to create new experiments may also be requested there.

3 Course Setting and Activities

We explore the use of this framework in a course which focuses on a subarea of Computational Linguistics, specifically discourse and dialogue, targeted at graduate students interested in research in this area. This topic is the subject of research not only in computational speech and language processing, but also in linguistics, psychology, sociology, anthropology, and philosophy. Research in this area draws on a growing, large-scale collection of text and multi-modal interaction data that often relies on

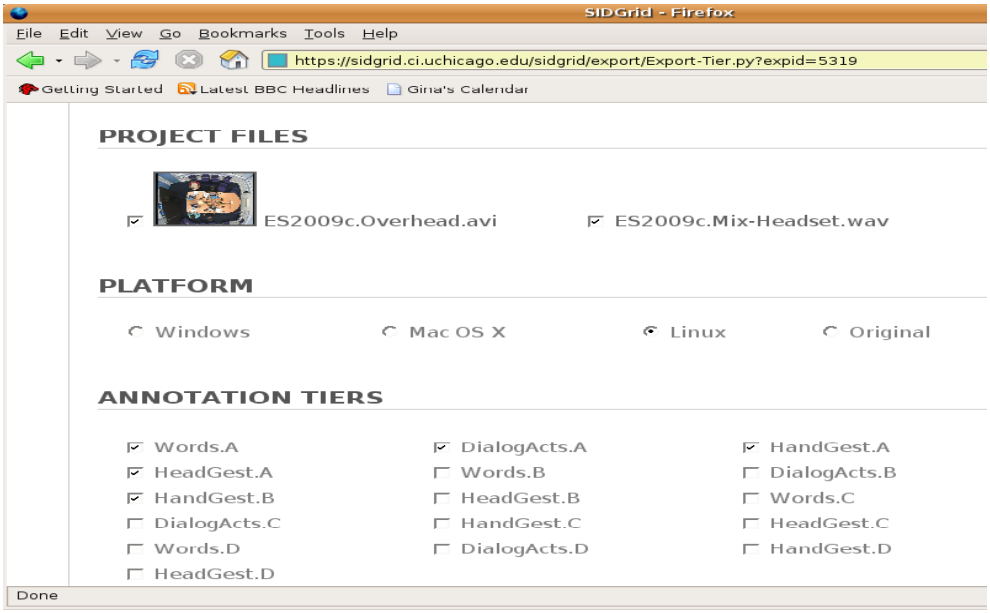

Figure 3: Screenshot of the archive download interface, with thumbnails of available video and download and analysis controls.

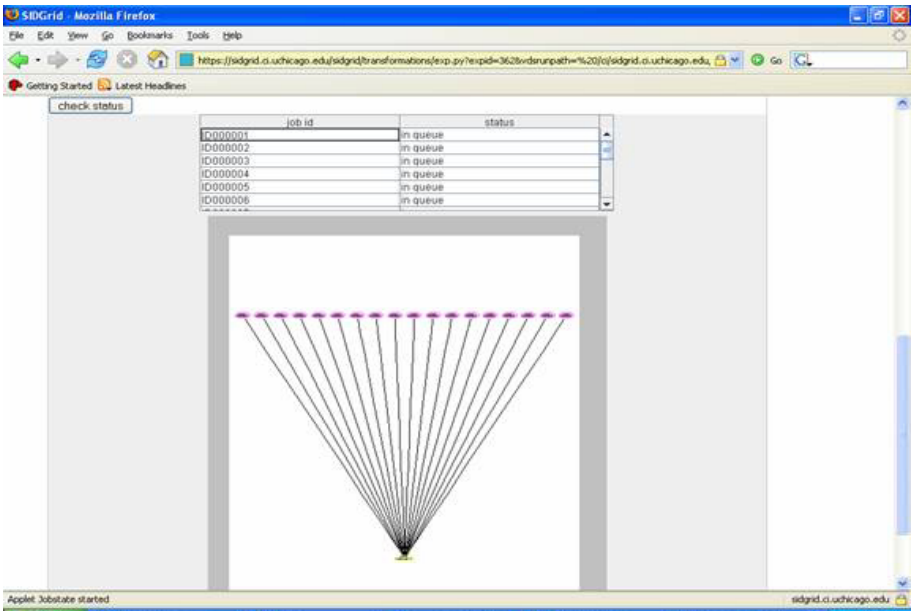

Figure 4: Progress of execution of programs on TeraGrid. Table lists file identifiers and status. Graph shows progress.

computational tools to support annotation, archiving, and analysis. However, prior offerings of this course through the Computer Science Department had attracted primarily Computer Science graduate students, even though readings for the course spanned the range of related fields. In collaboration with researchers in co-verbal gesture in the Psychology department, we hoped to increase the attraction and accessibility of the course material and exercises to a more diverse student population. After advertising the course to a broader population through the Linguistics Department mailing list, emphasizing the use of computational tools but lack of requirements for previous programming experience, the resulting class included members of the Linguistics, Slavic Studies, Psychology, and Computer Science Departments, about half of whom had some prior programming experience, but few were expert.

3.1 Hands-on Exercises

Currently, we have only included a small number of software tools as proof-of-concept and to enable particular course exercises in discourse and dialogue. This first set of exercises explores three main problems in this area: topic segmentation, dialogue act tagging, and turn-taking.

The topic segmentation exercise investigates the impact of segment granularity and automatic speech recognition errors on topic segmentation of conversational speech. The data is drawn from the Cross-Language Speech Retrieval Track of the Cross-language Evaluation Forum (CLEF CL-SR) (Pecina et al., 2007) collection. This collection includes automatic transcriptions of interviews from an oral history project, accompanied by manual segmentation created as part of the MALACH project (Franz et al., 2003). The exercise employs the web-based portal to the TeraGrid to perform segmentation of multiple interviews in parallel on the Grid, followed by evaluation in parallel. We perform segmentation using LCSeg (Galley et al., 2003) and evaluate using the p_k and WindowDiff metrics. Students identify the best segmentation parameters for these interviews and perform error analysis to assess the effect of ASR errors.

The dialogue act tagging exercise involves both annotation and analysis components. The students are asked to download and annotate a small portion of a conversation from the AMI corpus (Carletta et al., 2005) with dialogue act tags. The AMI corpus of multiparty meetings includes recorded video, recorded audio, aligned manual transcriptions, and manually annotated head and hand gesture. Students annotate from text alone, with audio, with video, and with all modalities. Local "transformations", programs or scripts associated with the annotation client, can also provide prosodic analysis of features such as pitch and intensity. Students are asked to assess the influence of different features on their annotation process and to compare to a gold standard annotation which is later provided. The automatic analysis phase is performed on the web-based portal to assess the impact of different feature sets on automatic tagging. The tagging is done in the Feature Latent Semantic Analysis framework (Serafin and Di Eugenio, 2004), augmented with additional prosodic and multi-modal features drawn from the annotation. Since this analysis requires Singular Value Decomposition of the potentially large Feature-by-Dialogue-Act matrices, it is often impractical to execute on single personal or even departmental servers. Furthermore, feature extraction, such as pitch tracking, of the full conversation can itself strain the computational resources available to students. Grid-based processing overcomes both of these problems.

Exercises on turn-taking follow similar patterns. An initial phase requires annotation and assessment exercises by the students in the ELAN-based client tool and downloaded from the web-based repository. Subsequent phases of the exercises include application and investigation of automatic techniques using the web-based environment and computational resources of the TeraGrid. Clearly, many other exercises could be framed within this general paradigm, and we plan to extend the options available to students as our interests and available software and data sets permit.

4 Impact on Interdisciplinary Instruction

We designed these hands-on exercises to allow students to investigate important problems in discourse and dialogue through exploration of the data and application of automatic techniques to recognize these phenomena. We aimed in addition to exploit

the cyberinfrastructure framework to achieve three main goals: lower barriers of entry to use of computational tools by students with little prior programming experience, enable students with greater computational skills to expand the scale and scope of their experiments, and to support collaborative projects and a broader, interdisciplinary perspective on research in discourse and dialogue.

4.1 Enabling All Users

A key goal in employing this architecture was to enable students with little or no programming experience to exploit advanced computational tools and techniques. The integration of so-called "transformations", actually arbitrary program applications, in both the annotation client and the web-based portal to the TeraGrid, supports this goal. In both cases, drop-down menus to select programs and text- and check-boxes to specify parameters provide graphical user interfaces to what can otherwise be complex command-line specifications. In particular, the web-based portal removes requirements for local installation of software, shielding the user from problems due to complex installations, variations in platforms and operating systems, and abstruse command-line syntax. In addition, the web-based archive provides simple mechanisms to browse and download a range of data sources. The students all found the archive, download, and transformation mechanisms easy to use, regardless of prior programming experience. It is important to remember that the goal of this environment is not to replace existing software systems for Natural Language Processing, such the Natural Language Toolkit (NLTK) (Bird and Loper, 2004), but rather to provide a simpler interface to such software tools and to support their application to potentially large data sets, irrespective of the processing power of the individual user' system.

4.2 Enabling Large-Scale Experimentation

A second goal is to enable larger-scale experimentation by both expert and non-expert users. The use of the web-based portal to the TeraGrid provides such opportunities. The portal provides access to highly distributed parallel processing capabilities. For example, in the case of the segmentation of the oral history interviews above, the user can select several interviews, say 60, to segment by checking the associated check-boxes in the interface. The portal software will automatically identify available processing nodes and distribute the segmentation jobs for the corresponding interviews to each of the available nodes to be executed in parallel. Not only are there many processing nodes, but these nodes are of very high capacity in terms of CPU speed, number of CPUs, and available memory.

The multigigabyte data files associated with the growing number of multi-modal discourse and dialogue corpora, such as the AMI and ICSI Meeting Recorder collections, make such processing power highly desirable. For example, pitch tracking for such corpora is beyond the memory limitations of any single machine in the department, while such tasks are quickly processed on the powerful TeraGrid machines.

Expert users are also granted privileges to upload their own user-defined programs to be executed on the Grid. Finally, web services also enable execution of arbitrary read-only queries on the underlying database of annotations, media files, and time-series data through standard Structure Query Language (SQL) calls. All these capabilities enhance the scope of problems that more skilled programmers can employ in the study of discourse and dialogue phenomena.

4.3 Interdisciplinary Collaboration and Perspectives

The web-based archive in the SIDGrid framework also provides support for group distributed collaborative projects. The archive provides a Unix-style permission structure that allows data sharing within groups. The process of project creation, annotation, and experimentation maintains a version history. Uploads of new annotations create new versions; older versions are not deleted or overwritten. Experimental runs are also archived, providing an experiment history and shared access to intermediate and final results. Script and software versions are also maintained. While the version control is not nearly as sophisticated as that provided by GForge or Subversion, this simple model requires no special training and facilitates flexible, web-based distributed access and collaboration.

Finally, the interleaving of annotation and automated experimented permitted by this integrated ar-

chitecture provides the students with additional insight into different aspects of research on discourse and dialogue. Students from linguistics and psychology gain greater experience in automatic analysis and recognition of discourse phenomena, while more computationally oriented students develop a greater appreciation of the challenges of annotation and theoretical issues in analysis of dialogue data.

5 Challenges and Costs

The capabilities and opportunities for study of computational approaches to discourse and dialogue afforded within the SIDGrid framework do require some significant investment of time and effort. Incorporating new data sets and software packages requires programming expertise. The framework can, in principle, incorporate arbitrary data types: media, physiological measures, manual and automatic annotations, and even motion tracking. The data must be converted into the ELAN .eaf format to be deployed effectively by the annotation client and interpreted correctly by the archive's underlying database. Converters have been created for several established formats[2], such as Annotation Graphs (Bird and Liberman, 2001), ANVIL (Kipp, 2001), and EXMARaLDA(Schmidt, 2004), and projects are underway to improve interoperability between formats. However, new formats such as the CLEF Cross-language Speech Retrieval SGML format and NITE XML(Carletta et al., 2003) format for the AMI data used here, required the implementation of software to convert the source format to one suitable for use by SIDGrid.

Incorporating new Grid-based "transformation" programs can also range in required effort. For self-contained programs in supported frameworks - currently, Perl, Python, Praat, and Octave - adding a new program requires only a simple browser-based upload. Compiled programs, such as LCSeg here, must be compatible with the operating systems and 64-bit architecture on the Grid servers, often requiring recompilation and occasionally addition of libraries to existing Grid installations. Finally, software with licensing restrictions can only run on a local cluster rather than on the full TeraGrid. Thus, public domain programs and systems that rely on such are preferred; for example, Octave-based programs are preferred to Matlab-based ones.

Finally, one must remember that the SIDGrid framework is itself an ongoing research project. It provides many opportunities to enhance interdisciplinary instruction in Computational Linguistics, especially in areas involving multi-modal data. However, the functionality is still under active development, and current system users are beta-testers. The use of the system, both in coursework and in research, has driven improvements and expansions in service.

6 Conclusions and Future Directions

We have explored the use of the SIDGrid framework for annotation, archiving, and analysis of multimodal data to enhance hands-on activities in the study of discourse and dialogue in a highly interdisciplinary course setting. Our preliminary efforts have demonstrated the potential for the framework to lower barriers of entry for students with less programming experience to apply computational techniques while enabling large-scale investigation of discourse and dialogue phenomena by more expert users. Annotation, analysis, and automatic recognition exercises relating to topic segmentation, dialogue act tagging, and turn-taking give students a broader perspective on research and issues in discourse and dialogue. These exercises also allow students to contribute to class discussion and collaborative projects drawing on their diverse disciplinary backgrounds. We plan to extend our current suite of hands-on exercises to cover other aspects of discourse and dialogue, both in terms of data sets and software, including well-known toolkits such as NLTK. We hope that this expanded framework will encourage additional interdisciplinary collaborative projects among students.

Acknowledgments

We would like to thank Susan Duncan and David McNeill for their participation in this project as well as the University of Chicago Academic Technology Innovation program. We would also like to thank Sonjia Waxmonsky for her assistance in implementing the course exercises, and the entire SIDGRID team for providing the necessary system infrastruc-

[2] www.multimodal-annotation.org

ture. We are particularly appreciative of the response to our bug reports and functionality requests by Tom Uram and Sarah Kenny.

References

Jens Allwood, Leif Groenqvist, Elisabeth Ahlsen, and Magnus Gunnarsson. 2001. Annotations and tools for an activity based spoken language corpus. In *Proceedings of the Second SIGdial Workshop on Discourse and Dialogue*, pages 1–10.

S. Bird and M. Liberman. 2001. A formal framework for linguistic annotation. *Speech Communication*, 33(1,2):23–60.

Steven Bird and Edward Loper. 2004. Nltk: The natural language toolkit. In *Proceedings of the ACL demonstration session*, pages 214–217.

P. Boersma. 2001. Praat, a system for doing phonetics by computer. *Glot International*, 5(9–10):341–345.

J. Carletta, S. Evert, U. Heid, J. Kilgour, J. Robertson, and H. Voormann. 2003. The NITE XML Toolkit: flexible annotation for multi-modal language data. *Behavior Research Methods, Instruments, and Computers, special issue on Measuring Behavior*, 35(3):353–363.

Jean Carletta, Simone Ashby, Sebastien Bourban, Mike Flynn, Mael Guillemot, Thomas Hain, Jaroslav Kadlec, Vasilis Karaiskos, Wessel Kraaij, Melissa Kronenthal, Guillaume Lathoud, Mike Lincoln, Agnes Lisowska, Iain A. McCowan, Wilfried Post, Dennis Reidsma, and Pierre Wellner. 2005. The AMI meetings corpus. In *Proceedings of the Measuring Behavior 2005 symposium on Annotating and measuring Meeting Behavior*.

M. Franz, B. Ramabhadran, T. Ward, and M. Picheny. 2003. Automated transcription and topic segmentation of large spoken archives. In *Proceedings of EUROSPEECH*.

Michel Galley, Kathleen McKeown, Eric Fosler-Lussier, and Hongyan Jing. 2003. Discourse segmentation of multi-party conversation. In *Proceedings of ACL'03*.

M. Kipp. 2001. Anvil- a generic annotation tool for multimodal dialogue. In *Proceedings of the 7th European Conference on Speech Communication and Technology (Eurospeech)*, pages 1367–1370.

Pavel Pecina, Petra Hoffmannova, Gareth J. F. Jones, Ying Zhang, and Douglas W. Oard. 2007. Overview of the clef-2007 cross language speech retrieval track. In *Working Notes for CLEF 2007*.

Rob Pennington. 2002. Terascale clusters and the TeraGrid. In *Proceedings for HPC Asia*, pages 407–413. Invited talk.

T. Schmidt. 2004. Transcribing and annotating spoken language with EXMARaLDA. In *Proceedings of the LREC-Workshop on XML-based richly annotated corpora*.

Riccardo Serafin and Barbara Di Eugenio. 2004. Flsa: Extending latent semantic analysis with features for dialogue act classification. In *Proceedings of the 42nd Meeting of the Association for Computational Linguistics (ACL'04), Main Volume*, pages 692–699, Barcelona, Spain, July.

P. Wittenburg, H. Brugman, A. Russel, A. Klassmann, and H. Sloetjes. 2006. Elan: a professional framework for multimodality research. In *Proceedings of Language Resources and Evaluation Conference (LREC) 2006*.