Personal Statement for Gina-Anne Levow

1 Research

My research is strongly interdisciplinary, drawing on methods from computer science to investigate fundamental linguistic questions and applying findings from linguistics to develop improved techniques for automatic computational understanding of natural language. My research lies at the intersection of computational linguistics, natural language processing (NLP), and spoken language processing. I explore how information is conveyed and structured in natural language discourse, in both text and speech, to inform the development of systems that interact with users in natural language or provide information from natural language materials. In particular, my experience in developing spoken dialogue systems and information retrieval systems has motivated my choice of research topics. The challenges presented for both users and developers of spoken language and information retrieval systems include handling miscommunication (1.1, para 1), determining document similarity (1.1, para 2), detecting prominence and tone (1.2.1, para 1), modeling turn-taking and feedback (1.2.1, para 2), and evaluating dialog (1.2.1, para 3).

Conversational, accented or noisy speech remains challenging despite the dramatic improvements in automatic speech recognition. Spoken language systems not only misrecognize words, but they remain awkward to use and difficult to correct. Such systems are not yet fluent conversational partners. My work aims to improve language understanding by exploiting linguistic information beyond the surface word or sentence level. The speech signal carries not only the identity of consonants and vowels, but also prosodic information, including pitch, loudness, and duration. Human languages make crucial use of prosody to convey information, for example, distinguishing a question from an answer with a rising intonation in English ("Yes?"). However, despite the fundamental importance of this prosodic information, computational approaches to speech recognition and processing have largely viewed such variation as a source of noise to be removed.

1.1 Earlier Research

While this statement focuses on my recent research work, this section presents two areas of my earlier research which represent my research approach, for which I am known, and which have had an impact in the research community as evidenced by Google Scholar citation count.

Prosody in Discourse and Dialog Understanding: Recognizing Miscommunication Miscommunication in human-computer interaction is unavoidable. On a noisy phone line, a computer can readily misrecognize 'Dallas' for 'Dulles' in voice-based system for travel information. These communicative failures are particularly frustrating for the user, often leading to sequences of errors termed "error spirals" while the system remains oblivious to the mounting errors and user frustration. My work in the development of the prototype SpeechActs telephone-based spoken dialogue system *[Yankelovich et al., 1995]*[1] identified key challenges in speech user interface design and highlighted not only recognition problems but also the difficulty of correcting such errors. In a series of simulated interaction studies *[Oviatt et al., 1998]*, we systematically explored the acoustic-prosodic adaptations of user speech in the face of speech recognizer errors, identifying key characteristics of this exaggerated, hyperarticulate speaking style. Using data from a field trial of SpeechActs, I subsequently demonstrated that prosodic features —such as speaking rate, pausing, duration, and pitch —could be used to train a machine learning classifier, a computer system learns to apply labels (e.g., 'correction' versus 'original') based on regularities in data, to automatically recognize spoken corrections at levels approaching human performance [Levow, 1998]. My analysis also showed that these adaptations represented divergences from recognizer models for conversational speech, contributing to subsequent recognition errors [Levow, 2002]. This work was replicated and extended by other researchers for a range of languages, including German, Dutch, Swedish, and Japanese. Furthermore, I showed that similar prosodic features could go beyond determining whether a user utterance was corrective to identify

[1] Citations in italics indicate papers on which I was an author, though not first author.

which specific word was being corrected [Levow, 2004b]. These capabilities pave the way for spoken dialogue systems that are more sensitive and adaptive to miscommunications. This stream of work has had an substantial impact in the community. The original SpeechActs speech user interface paper to date has a Google Scholar citation count over 200, and the papers on hyperarticulation in spoken corrections on which I am an author have an aggregate Google citation count over 300.

Improving Access to Information My research in information retrieval has focused on multi-lingual and multimedia retrieval, primarily between English and a wide range of European and Asian languages, including French, Italian, German, Chinese, and Japanese [Levow et al., 2005]. In these tasks, a query posed in one language is used to retrieve documents —written or spoken —not only in the original language, but also in one or more other languages. As a result, the errors due to machine translation or speech recognition compound the original substantial challenge of matching the user's information need to a document. My work has aimed to develop techniques that are as language-independent as possible [Levow et al., 2001], requiring only simple linguistic resources *[Oard et al., 2000, Resnik et al.,2001]*. These techniques emphasize enriching documents or queries with additional translation alternatives and related terms from presumed relevant documents [Levow, 2003], to overcome translation ambiguity and gaps. Our approaches employed multi-scale indexing and retrieval (*[Meng et al., 2001, Lo et al., 2003, Meng et al., 2004]*) to improve effectiveness by eschewing words as the primary unit of representation in favor of sub-word and cross-word character sequences and phrasal units to overcome segmentation ambiguity and error. *[Matveeva and Levow, 2007a; Matveeva and Levow, 2007b]* developed a novel approach for matching document terms to underlying concepts to further overcome the limitations of word-for-word matching. These approaches yielded significant improvements over previous techniques. This work has proven influential in cross-language and spoken document retrieval communities yielding an aggregate Google Scholar citation count for papers in this area on which I am a co-author of over 300. On the basis of this work I was invited to give four hands-on laboratory tutorials at prestigious Summer Schools in Natural Language Processing and Machine Learning.

1.2 Recent Research

1.2.1 Modeling Spoken Discourse and Dialog

Prosody in Discourse and Dialog Understanding: Recognizing Tone Pitch also is employed to convey discourse or lexical meaning at the word level. In the case of targeted spoken corrections, the specific word being corrected is uttered with greater prominence, often described as a pitch accent. In the case of tone languages, such as Mandarin Chinese, the meaning of each syllable is determined by its pitch height and contour. My work in this area began before my arrival at UW and has continued in my time here. As PI on the NSF-funded project "Learning Tone" (NSF IIS), I have developed techniques to recognize lexical tone and pitch accent in languages from Mandarin Chinese to isiZulu (a Bantu family language) to English [Levow, 2005a; Levow, 2006; Levow, 2008; Levow, 2009]. Recognition of tone and pitch accent is particularly challenging because the actual pronunciation can be dramatically altered due to the surrounding speech context. To overcome this challenge, my approach employs a common model of local syllabic and phrasal context across this wide variety of languages. Significant improvements in accuracy are achieved through this contextual modeling [Levow, 2005a] and rival the best published results on this task.

A second barrier to exploitation of the information in these prosodic features is the need for large quantities of manually labeled data to train standard machine learning classifiers; such data is costly to collect both in terms of time and money. My work overcomes this barrier through the novel use of specialized machine learning techniques, that require little or no manually labeled data for training but nevertheless achieve accuracies for tone and pitch accent recognition that would require tens of thousands of examples with standard models [Levow, 2006]. In collaboration with my doctoral students, I have also further improved techniques for tone recognition, by using novel acoustic *[Surendran and Levow, 2006, Wang and Levow, 2006, Wang and Levow, 2008]* and positional information *[Wang and Levow, 2011*]. I have also investigated the use of prosody in assessment of second language learners. I demonstrated that this contextual framework supported training of pitch accent recognition models on native speech for prediction of learner pitch accent with no significant reduction in accuracy relative to models trained directly on learner speech [Levow, 2009].

Furthermore, I supervised a UW Master's thesis integrating prosodic, pronunciation, and word evidence to improve automatic assessment of language learner proficiency [Podgornik, 2011].

Prosody in Discourse and Dialog Understanding: Modeling the Flow of Conversation Natural conversations are complex processes, involving the exchange not only of information but also of the turns, responses, and signals of attention that allow the smooth flow between interlocutors. Such conversational cues are often subtle and culture-specific and may exploit multiple modalities such as gaze, head nod, posture, or gesture. As PI and investigator the NSF-funded "Dyadic Rapport Within and Across Cultures: Multimodal Assessment of Human-Human and Human-Computer Interaction" (BCS HSD), I led a multi-institutional, interdisciplinary team comprising researchers in Computer Science, Psychology, and Cognitive Science to characterize and recognize signals to conversational rapport in three language/cultural groups: American English, Iraqi Arabic, and Mexican Spanish. Through a corpus of unrehearsed story-telling narratives in these languages that we collected and annotated, I identified basic contrasts in the rates of verbal and non-verbal feedback for these groups [Levow et al., 2010; Levow and Duncan, 2012]. Furthermore, we were able to exploit prosodic cues to recognize contexts where listener feedback was appropriate and identify differences in cues across languages *[Wang and Levow, 2010; Wang and Levow, 2012]*. Finally, we found that different types of interactional feedback - head nods versus vocal backchannels versus speaking turns - were elicited with significantly different cues, suggesting that these different forms of feedback might be tied to different communicative roles [Levow and Duncan, 2012].

Evaluating Spoken Dialog Systems Even as understanding of fundamental mechanisms of human dialog has improved, understanding of what constitutes good human-computer dialog remains frustratingly limited. Effective evaluation of human-computer dialog systems, such as telephone-based systems for air travel information, has historically been performed through questionaires completed by small groups of subjects in controlled, and artificial, conditions. However, as spoken dialog systems are more widely deployed, evaluation becomes both more challenging and more important. It is important to be able to assess large numbers of naturalistic dialogs from real users accomplishing real tasks and then to identify and model those aspects of the dialog systems which contribute most to their success (or failure). Since real users are unlikely to complete a user survey and expert annotation of large corpora of calls is prohibitively costly, together with Dr. Helen Meng and Dr. Irwin King at the Chinese University of Hong Kong, I developed a crowdsourcing methodology for dialog system evaluation whereby judgments of large numbers of non-expert raters could be gathered with high levels of agreement for recorded interactions of real users of deployed systems *[Yang et al., 2010; Yang et al, 2013]*. Based on these judgments, we were able to develop improved predictive models of dialog system quality, *[Yang et al, 2010; Li et al, 2010, Zhu et al, 2010, Yang et al, 2012]*. Since my arrival at UW, this work led to four SLTC-2010 papers, a journal article, an edited volume on crowdsourcing for speech processing with a contributed chapter, and a standing-room only special session at Interspeech 2011, the flagship conference of the International Speech Communication Association.

New Directions for Prosody in Dialog: Attitude in Spoken Language I am pursuing new directions in the use of prosody to improve understanding paralinguistic functions in dialog through developing collaborations at the University of Washington. These two efforts relate to the expression of speaker attitude in potentially contentious discussions and how contentious discussions and stance-taking are signaled through linguistic means, particularly by changes in prosody. With Prof. Mari Ostendorf (Electrical Engineering), I am working on a DARPA-funded project investigating the relationship between atypical prosody, such as emphasizing typically minimized words like "the", and expressions of attitude or disagreement in spoken dialog. With Prof. Ostendorf and Prof. Richard Wright (Linguistics), I have begun exploration of automatic recognition of stance-taking in negotiation dialogs and legal hearings, focusing on the use of novel acoustic cues, to augment typical word-based models. This research, funded by a newly awarded NSF grant, moves forward in novel domains and new collaborations, while building on key elements of my prior research: using prosody to enhance disourse and dialog understand by recognizing speaker state, harnessing speaking style as a source of information, and identifying prominence to improve spoken language understanding.

1.3 Accomplishments

My research has made contributions in natural language processing, spoken language processing, and information retrieval. These lines of research and others have led to the publication of more than 60 peer-reviewed papers in journals, conferences, and workshop proceedings, with more than 15 publications or refereed presentations since my arrival at UW. I have been Principal Investigator on two prior NSF-funded projects, “Learning Tone” (CISE IIS) and “Dyadic Rapport Within and Across Cultures: Multimodal Assessment of Human-Human and Human-Computer Interaction” (BCS HSD), with almost all of the publications and refereed presentations in the latter project coming after joining UW. I was also PI on a University of Chicago funded Academic Technology Innovation grant, “Multi-modal Discourse Investigation with SIDGRID.”

I am now PI on a new NSF-funded project, “ATAROS: Automatic Tagging and Recognition of Stance” (CISE IIS) and co-PI on a Hong Kong Research Grants Council-funded project, ”The Use of Phonologically-Motivated Distinctive Features for Computational Acoustic Characterization of Dysarthric Speech.”

1.4 Ongoing Research and Future Directions

I hope to develop the next generation of spoken language systems that would be characterized less by creating a sufficiently restrictive interaction that it is easy to predict user behavior, than by exploiting improved understanding of spoken interaction to enable fluent, flexible conversational systems. These systems can employ improved automatic prosodic labelling to allow unit selection for enhanced speech synthesis. The output of such speech synthesis systems would then be more appropriate to the discourse and dialogue context. These systems could recognize and respond correctly to user topic and turn-taking behavior, anticipating turn changes and identifying interruptions. By identifying topic structure and emphasized information in spoken documents, systems will improve flexible access to information. These systems could also effectively integrate information about discourse, turn-taking, and interaction available in other modalities, such as gaze, nod, and gesture as opportunities for multi-modal conversational systems become more available. To achieve fully robust natural conversation, systems must not only obtain high word recognition accuracy but also need to approach human levels of discourse and dialogue control and understanding and even become sensitive to indications of uncertainty or certainty, frustration, or anger. Only through understanding and integrating prosody will we be able to achieve these advances in language processing and understanding. My research aims to make such systems possible.

1.5 Publishing in Computational Linguistics

In the field of Computational Linguistics, as in Computer Science as a whole, research is disseminated largely through peer-reviewed conference and workshop publications in addition to journal articles. Peer-review for these conferences is similar in many ways to journal review in other fields. Full papers are submitted and detailed review is carried out, typically by three reviewers. Review is rigorous, and authors are expected to revise papers based on the reviews prior to final publication. Acceptance at these venues is also highly competitive, with typical acceptance rates ranging from 25% to 55% or so. Mode of presentation (poster or oral) is not determined based on quality.

Furthermore, joint authorship is very common. Research is carried out in different settings: by single investigators, jointly with a researcher and students under their supervision, and often in collaborative teams of researchers. Publications reflect that structure. I have carried out research and published it in all of these settings. In general, when work was carried out jointly by a team and when student research was conducted within a framework I had created, the resulting publications are jointly authored, with me as one of the co-authors. When a graduate student worked in an area of their creation with little guidance from me, the resulting publications would bear their name alone.

2 Service

My service within the University of Washington to date has been focused within the Department of Linguistics and particularly within the Computational Linguistics Professional Master’s Program (CLMS). I have served on the program's admissions committee (2010, 2011). With Fei Xia, I have also organized (2010-) a series of "Graduation Planning" meetings. These monthly meetings aim to prepare and guide students in the CLMS program's required project —either an internship or thesis, and they cover topics ranging from choosing a thesis topic to preparing an effective industry resume. As a professional program, internship placement serves as a crucial bridge to full-time professional employment. We also supervise 10-12 students per year throughout the internship process, from applications through execution and final project evaluation. To date, I have had thirteen such advisees who have completed their projects, with another five currently underway. Finally, to build bridges with the local and regional language processing communities for both the program and our students, I have served as a co-organizer of the quarterly UW/Microsoft Computational Linguistics Symposium series (2011-2013), as co-chair (with Luke Zettlemoyer in Computer Science) of the Northwest Natural Language Processing Regional Workshop (Spring 2012), and as founding Board Member of the Seattle chapter of the Applied Voice Input-Output Society (2012 - current), in events at Microsoft and UW.

In the broader research community, I perform regular reviewing for a range of journals (3-4 per year) and major conferences as well as serving on conference program committees (3-4 per year). I served as a member of the Editorial Board for *Computational Linguistics* (2003-2005), the flagship publication of the eponymous Association, and Associate Editor of *ACM Transactions on Asian Language Processing (TALIP)* (2007-2009). I am currently Secretary/Webmaster for the Association for Computational Linguistics' Special Interest Group on Chinese Language Processing (SIGHAN) (2006-current).

3 Teaching

I view teaching as crucially providing students with tools – linguistic, algorithmic, analytic, and problem-solving – that will be of use to them throughout their academic and professional careers. A challenge of teaching is to present the material in a way that is not only clear and well-organized, but also well-motivated and relevant to the students. Making a conscious and continuous effort to show students why they are learning *this* material *now* helps to engage students and to provide an organizing principle for courses.

My teaching at the University of Washington has been within the context of the CLMS program. I have taught all four of the courses in the computational core curriculum as well as specialized topics courses in areas such as "Discourse and Dialog" and "Spoken Dialog Systems". All of these courses employ the program's hybrid presentation format, with lectures presented simultaneously to in-class and remote students through specialized meeting room software. Working to foster engagement and participation in my students by connecting course material to real-world applications has significantly affected the development of my course design and presentation style. This perspective is particularly crucial in the context of the CLMS program since it is a professional program and students expect to be able to translate what they learn in our courses directly to work in their chosen field. At the same time, the courses aim to teach the fundamentals of the field, on which students will be able to build in a dynamic, rapidly changing area.

One of my regularly taught classes, Natural Language Processing Systems and Applications, exemplifies this approach. This course functions as the CLMS program's capstone course, providing the opportunity for students to apply the tools from prior classes to a more real-world problem. Recently, students developed "Question-Answering Systems", to provide short textual answers to factual questions posed in English, based on text documents. I link their task to high-profile, public successes of NLP such as Google and Watson's recent victory at Jeopardy!. I demonstrate that, although the curriculum sometimes creates a dichotomy between shallow word-based approaches and deeper syntactic and semantic techniques, both can successfully solve the problem. Furthermore, students learn to connect research results to these real-world problems, as research papers form the class readings and source of methods. They also learn to critique and assess these results in the face of real-world constraints of time, processing power, and new, complex, and often messy data. Finally, in addition to exercising their computational linguistic skills, the students also gain experience for future work in real-world settings, such as working in teams, managing software projects, completing deliverables, and presenting written and oral reports. The increase in discussion during group project presentations, the friendly rivalry among teams for the best results, and substantial improvements in system effectiveness as the course progresses demonstrate the success of this instructional strategy.