Agenda

Organizer
Mark A. Finlayson: (617) 515-0708
School of Computing and Information Sciences, Florida International University

Participants

Claire Bonial, *University of Colorado, Boulder*
Steven Cassidy, *Macquarie University*
Wendy Chapman, *University of Utah*
Markus Dickinson, *Indiana University*
Jeff Good, State *University of New York at Buffalo*
Thomas Hanke, *University of Hamburg*
Marie Hinrichs, *Tübingen University*
Erhard Hinrichs, *Tübingen University*
Nancy Ide, *Vassar College*
Michael Kipp, *Augsburg Univ. of Applied Sciences*
Brian MacWhinney, *Carnegie Mellon University*
Diana Maynard, *University of Sheffield*
Eric Nyberg, *Carnegie Mellon University*
George Petasis, *NCSR Demokritos*
James Pustejovsky, *Brandeis University*
Anna Rumshisky, *Univ. of Massachusetts, Lowell*
Gary Simons, *SIL International*
Han Sloetjes, *Max Planck Inst. for Psycholinguistics*
Brett South, *University of Utah*
Pontus Stenetorp, *University of Tokyo*
Stephanie Strassel, *Linguistic Data Consortium*
Marc Verhagen, *Brandeis University*

Sunday, March 29, 2015

Welcome reception on the Caracol Terrace, 8-10pm

Monday, March 30, 2015 Workshop will be held in the **Ocean Terrace Room**

Start	*End*	
9:00 AM	9:30 AM	Introductory Remarks by Mark Finlayson
9:30 AM	10:30 AM	**Motivations: What are the problems we are trying to solve?**
10:30 AM	11:00 PM	*Coffee Break*
11:00 PM	12:30 PM	**Motivations II**
12:30 PM	2:00 PM	*Lunch in Caracol Restaurant*
2:00 PM	3:30 PM	**Capabilities: What functionalities should we target, with what priority?**
3:30 PM	4:00 PM	*Coffee Break*
4:00 PM	5:15 PM	**Capabilities II**
5:15 PM	5:30 PM	**Discussion: Should we re-work tomorrow's agenda?**
5:30 PM	6:15 PM	*Free time*
6:15 PM		*Meet in lobby to go to dinner at Timo (17624 Collins Ave.)*

Tuesday, March 30, 2015

Start	*End*	
9:00 AM	9:30 AM	Overview of European Projects & Funding by Erhard Hinrichs
9:30 AM	10:30 AM	**Form: What form should a solution take?**
10:30 AM	11:00 PM	*Coffee Break*
11:00 PM	12:30 PM	**Incentives: How do we ensure adoption and long-term viability?**
12:30 PM	2:00 PM	*Lunch in Caracol Restaurant*
2:00 PM	3:30 PM	**Management: How should the solution be funded, built, and maintained?**
3:30 PM	4:00 PM	*Coffee Break*
4:00 PM	5:30 PM	**Discussion: Summarizing the recommendations**
5:30 PM	6:15 PM	*Free time*
6:15 PM		*Meet in lobby to go to dinner at Tony Roma's (18050 Collins Ave)*

Participant Bios

Mark Finlayson (Organizer)
Assistant Professor
School of Computing and Information Sciences
Florida International University, United States
markaf@fiu.edu http://cs.fiu.edu/~markaf

Professor Mark Finlayson received his Ph.D. in 2012 from MIT in Artificial Intelligence and Cognitive Science. His research focuses on the science of narrative, including understanding the relationship between narrative, cognition, and culture, developing new methods and techniques for investigating questions related to language and narrative, and endowing machines with the ability to understand and use narratives for a variety of applications. He has worked on linguistic annotation in service of this research, building the **Story Workbench**, an annotation tool designed to allow the simultaneous manual, automatic, or semi-automatic annotation of over 20 different layers of syntax and semantics onto text by non-technical annotators. Using the Story Workbench he has collected and annotated a number of corpora of narratives, including the **UMIREC corpus** (UCM/MIT Indications, Referring Expressions, and Co-Reference), the **N2 Corpus**, and a deeply annotated corpus of Russian folktales. With his students he has built variety of widely used NLP tools in Java, including **JWI**, **jMWE**, **jVerbnet**, and **jSemcor**.

Claire Bonial
Research Associate
Department of Linguistics
University of Colorado at Boulder, United States
cbonial@me.com http://www.colorado.edu/ics/people/claire-bonial

Dr. Claire Bonial completed her Ph.D. in Linguistics and Cognitive Science at the University of Colorado, Boulder in 2014. Since 2007 Dr. Bonial has worked with Dr. Martha Palmer as a Research Associate in CU's Center for Language Education and Research (CLEAR) Lab. During this time Dr. Bonial has focused on the development, maintenance, and expansion of **PropBank**, **VerbNet**, **SemLink**, and the **Abstract Meaning Representation (AMR)** project. Dr. Bonial also assisted in the development of the **Jubilee** annotation tool, used for PropBank annotation, as well as **Cornerstone**, the tool used for the creating and editing the PropBank lexicon of frame files (i.e. sense inventory).

Steven Cassidy
Associate Professor
Department of Computing
Macquarie University, Australia
steve.cassidy@mq.edu.au http://web.science.mq.edu.au/~cassidy

Professor Steven Cassidy is a computer scientist (with a Ph.D .in Cognitive Science) who has worked on various areas relating to speech and language technology over the last 30 years. With Jonathan Harrington he developed the **Emu** Speech Database System to support corpus based research in speech and acoustic phonetics. Emu supports a flexible hierarchical annotation system and provides a query language and analysis environment based on the R Statistical environment. Emu is widely used to support research on small and large scale speech corpora and includes tools to support every stage of the corpus collection and analysis lifecycle. Emu is now maintained by a team of developers in Munich. Professor Cassidy was also recently involved in the development and collection of an audio visual **Corpus of Australian English** from around 1000 speakers around Australia. He built the software for data capture and a server based system for data upload and publishing. His most recent work has been on the **Alveo Virtual Laboratory** which is both a repository for language resources and a platform to support tools for exploration and analysis of language data. Alveo currently holds around 20 collections including audio, video and text resources and is working on new acquisitions of data and tools.

Wendy Chapman
Chair & Professor, Department of Biomedical Informatics
University of Utah, United States
wendy.chapman@utah.edu http://medicine.utah.edu/faculty/mddetail.php?facultyID=u0073209

Professor Wendy Chapman has a B.S. in Linguistics, and earned her Ph.D. in Medical Informatics from the University of Utah in 2000. From 2000-2010 she was a NLM postdoctoral fellow and then faculty at the University of Pittsburgh, after which she joined the Division of Biomedical Informatics at the UC San Diego in 2010. In 2013, she became the chair of the University of Utah, Department of Biomedical Informatics. Professor Chapman's research focuses on developing and disseminating resources for modeling and understanding information described in **narrative clinical reports**. She is interested not only in better algorithms for extracting information from clinical text NLP but also in generating resources for improving the NLP development process (such as shareable annotations and open source toolkits) and in developing user applications to help non-NLP experts apply NLP in informatics-based tasks like clinical research and decision support. She has led development of several openly available clinical corpora and annotated corpora, including the **Pitt NLP Corpus** (unannotated clinical reports from 13 hospitals available for NLP research), the **ShARe Corpus** (clinical reports annotated with a multi-layer syntactic and semantic schema used in the CLEF/ShARe Shared Task 2013-2014 and SemEval Challenge 2015). She has helped develop a variety of annotation schemas that have culminated in the **Schema Ontology** and **Modifier Ontology** for annotating clinical text. She has also been involved in development of several annotation tools, incluidng **e-HOST** for entities, attributes, and relations in clinical text and **Chart Review** for annotating complete patient records. She helped design a system for performing distributed annotation over private corpora (i.e., clinical reports) called **Annotation Admin**. Also, she has helped develop tools for visualizing NLP output, comparing automated annotations with manual annotations, drilling into errors using a visual interface, and providing user feedback to the NLP system based on identifying errors.

Markus Dickinson
Associate Professor
Department of Linguistics
Indiana University, United States
md7@indiana.edu http://cl.indiana.edu/~md7

Professor Markus Dickinson has worked extensively on two areas: 1) developing techniques to automatically detect and correct errors in different kinds of linguistic annotation; and 2) linguistically annotating corpora containing second language learner data. The former project, **DECCA** (Detection of Errors and Correction in Corpus Annotation), was an NSF-funded project and has led to a recent orthogonal project to **DAPS** (Detect Anomalous Parse Structures), with the goal of being able to build very large annotated corpora. The latter work is best exemplified by the **SALLE** (Syntactically Annotating Learner Language of English) project, an ongoing effort adding multiple layers of linguistic annotation.

Jeff Good
Associate Professor
Department of Linguistics
State University of New York at Buffalo, United States
jcgood@buffalo.edu http://buffalo.edu/~jcgood

Professor Jeff Good's current research interests include the linguistic typology of linear relations, the comparative morphosyntax of Niger-Congo, the documentation and description of Bantoid languages of the Lower Fungom region of Northwest Cameroon, and the role of emerging digital methods in the documentation of endangered and other low-resource languages. In this last area, he has been especially interested in the development of standards for encoding and annotating lexical and grammatical data and in tools to facilitate linguistic fieldwork. He has served as co-PI on the Lexicon Enhancement via the **GOLD Ontology project**, where he directed the conversion of thousands of wordlists stored in a legacy format into an XML format designed for interoperability, and as PI of the pilot **Pangloss project**, which explored the possibility of building an annotation tool within a word processing system. He has additionally served as PI on a number of projects involving the documentation of **endangered languages** of Cameroon, and his current work in this area, also funded by NSF, has a collaborative component with a specialist in databases to build tools to support the management of data

collected in the field. In 2014, he organized the **ComputEL workshop** to explore how computational linguists and and endangered language linguists could more effectively collaborate http://buffalo.edu/~jcgood/ComputEL.html). In his work in linguistic typology, he has explored how graph-based descriptions of linguistic constructions can be rigorously compared with each other and developed prototype tools to support this. Within the linguistics community, he often serves as an informal liaison between the **language documentation** community and those developing digital standards for language resources.

Thomas Hanke
Researcher
Institute for German Sign Language and Communication of the Deaf
University of Hamburg, Germany
thomas.hanke@sign-lang.uni-hamburg.de http://dgs-korpus.de

Dr. Thomas Hanke works on language resources for **sign languages** and corresponding tools. He developed **syncWRITER** (1990), a first approach to annotate digital video. He also worked on **iLex**, a full-fledged team annotation environment for sign languages integration image processing. He is currently managing a longterm research project working towards a reasonably-sized **corpus of German Sign Language**, funded by the German Academies of Sciences program.

Erhard Hinrichs
Professor
General and Computational Linguistics
Tübingen University, Germany
erhard.hinrichs@uni-tuebingen.de http://www.sfs.uni-tuebingen.de/~eh

Professor Erhard Hinrichs is director of the Computational Linguistics research group at the University of Tübingen, Germany. He obtained a Ph.D. in Linguistics from The Ohio State University. His previous positions include Research Fellow at the Beckman Institute for Advanced Science and Technology; Assistant Professor at UIUC; and Research Scientist at BBN. His research interests include the computational modeling of language comprehension (particularly of syntax and semantics) and of language variation with special emphasis on the use of machine learning approaches to dialectology. Dr. Hinrichs has extensive experience in project coordination and leadership, e.g., as scientific coordinator of the **D-SPIN** project, member of the executive board of the **ESFRI** project **CLARIN**, and co-director and member of the **CLARIN-ERIC** Board of Directors. He has done key work on the annotation web-based linguistic annotation tool **WebLicht Tübingen Treebanks (TüBa)** of Spoken (German, English, Japanese) and Written (German) Language, with the following annotation layers: tokenization, lemmatization, morphology, part-of-speech, syntax (constituency and dependency), word senses, anaphora, named entity classification, and discourse relations.

Marie Hinrichs
Research Scientist
Tübingen University, Germany
marie.hinrichs@uni-tuebingen.de

Marie Hinrichs has a B.S. in Computer Science from the Ohio State University. She is currently working on the **CLARIN-D** project in the Department of Computational Linguistics at the University of Tübingen, Germany. She has a leading role in the development of **WebLicht**, an environment for the construction and execution of NLP processing chains. She has also recently become involved in the technical aspects of maintenance and release of the **TüBa-D/Z**, a treebank derived from German newspaper articles. In addition, she is involved in developing a generic execution environment for executing workflows close to the data in the European infrastructure project **EUDAT**.

Nancy Ide
Chair & Professor
Department of Computer Science
Vassar College, United States
ide@cs.vassar.edu http://www.cs.vassar.edu/~ide

Professor Nancy Ide has worked in the field of computational linguistics for over 30 years and made significant contributions to research in word sense disambiguation, computational lexicography, discourse analysis, and the use of semantic web technologies for language data. She has been involved in the development of annotation standards throughout her career, first as the founder of the **Text Encoding Initiative (TEI)**, the first major standard for representing electronic language data. She later developed the **XML Corpus Encoding Standard (XCES)** and, most recently, the **ISO LAF/GrAF** representation format for linguistically annotated data. She is the convener of ISO TC 37 SC4 WG1 on Basic Mechanisms for Language Resource Management, and has participated in the development of several ISO standards for language data. Professor Ide has managed the development of several major linguistically-annotated corpora, including the EU-funded **MULTEXT** and **MULTEXT-EAST** corpora, and, more recently, the **Open American National Corpus (OANC)** and the **Manually Annotated Sub-Corpus (MASC)**. She has been a pioneer in efforts toward open data and resources, publishing the OANC and MASC as the first linguistically-annotated corpora freely available for any use. She is Co-Editor-in-Chief of the journal Language Resources and Evaluation and Editor of the Springer book series Text, Speech, and Language Technology. She has been the Principal Investigator (PI) or co-PI on multiple major US National Science Foundation and EU-funded projects; currently, she is co-PI of the **LAPPS Grid** project, in which context she is developing standards for web service exchange of linguistically-annotated data.

Michael Kipp
Professor
Department of Computer Science
Augsburg University of Applied Sciences, Germany
michael.kipp@hs-augsburg.de http://michaelkipp.de/j15/index.php?lang=en

Professor Michael Kipp is full professor at Augsburg University of Applied Sciences, Germany. Before he was head of the Embodied Agents research group at the Cluster of Excellence "Multimodal Computing and Interaction" at Saarland University and a senior researcher at the German Research Center of AI. He has co-authored more than 70 peer-reviewed publications in the areas of human-computer interaction, virtual characters, multimodality research and video annotation. He created the **ANVIL** video annotation for his research on the automatic synthesis of coverbal gestures and has since then further developed it with the help of his students, especially in the direction of motion capture visualization.

Brian MacWhinney
Professor
Department of Psychology
Carnegie Mellon University, United States
macw@cmu.edu http://talkbank.org

Brian MacWhinney is Professor of Psychology, Computational Linguistics, and Modern Languages at Carnegie Mellon University. He has developed a model of first and second language processing and acquisition based on competition between item-based patterns. In 1984, he and Catherine Snow co-founded the **CHILDES** (Child Language Data Exchange System) Project for the computational study of child language transcript data. He is now extending this system to six additional research areas in the form of the **TalkBank** Project. MacWhinney's recent work includes studies of online learning of second language vocabulary and grammar, neural network modeling of lexical development, fMRI studies of children with focal brain lesions, and ERP studies of between-language competition. He is also exploring the role of grammatical constructions in the marking of perspective shifting and the construction of mental models in scientific reasoning. In the area of annotation, my work has focused on the enrichment of the CHILDES and TalkBank corpora with phonological, morphological, and syntactic coding. We have also developed coding systems for gesture, speech acts, sign language, and other areas, as specified in the manual for the **CHAT** transcription and coding system. We have also developed methods for converting between CHAT format and 8 other transcript formats, based on a detailed XML Schema for CHAT that includes the format used by

the **Phon** program for detailed phonological analysis. The CHILDES database includes corpora from 30 languages and we have developed morphological and syntactic taggers for 8 of these languages. I am particularly interested in developing methods for increased linkage of our CHAT format to other annotation tools.

Diana Maynard
Research Fellow
Department of Computer Science
University of Sheffield, United Kingdom
d.maynard@sheffield.ac.uk http://gate.ac.uk

Dr. Diana Maynard has been a Research Fellow at the University of Sheffield, UK since February 2000, after receiving the Ph.D. in Natural Language Processing from Manchester Metropolitan University. Her main interests are in information extraction, opinion mining, social media analysis, terminology and semantic web technologies. She is the chief developer of Sheffield University's open-source multilingual Information Extraction tools, and currently leads the work on Information Extraction and Opinion Mining on the EU DecarboNet project. Dr. Maynard is a senior member of the current **GATE** team of 12 researchers, and has been involved in developing the GATE architecture and toolkit since its inception (in its current format) in 2000 and has been heavily involved in the design of both manual and automatic annotation tools, in particular from the user point of view. She developed many of the linguistic processing resources in GATE, in particular the core Information Extraction system **ANNIE**, and was responsible for the development of many of the multilingual tools in GATE. She led the GATE team's work on the NIST ACE evaluations, and on the TIDES Surprise Language Evaluation, both of which met with great success and led to the development of tools for a variety of new languages in GATE (Arabic, Chinese, Cebuano and Hindi) in addition to the English components. Currently, Dr. Maynard is best known for her work on opinion mining and sentiment analysis (and particularly for some recent work on sarcasm detection) and more generally for her work on social media analysis (for example, adapting core IE tools to deal with Twitter and other noisy data).

Eric Nyberg
Professor
School of Computer Science
Carnegie Mellon University, United States
ehn@cs.cmu.edu http://www.cs.cmu.edu/~ehn

Professor Eric Nyberg was a member of the **Unstructured Information Management Architecture (UIMA)** steering committee and has many years of experience with annotation systems for information extraction, semantic retrieval, and question answering, including work on annotation type systems in the IARPA AQUAINT, DARPA GALE and DARPA MRP programs.

George Petasis
Researcher
Institute of Informatics and Telecommunications
NCSR Demokritos, Greece
petasis@iit.demokritos.gr http://www.ellogon.org/petasis/

Dr. Georgios Petasis holds a Ph.D. in Computer Science from University of Athens on machine learning for natural language processing. His research interests lie in the areas of natural language processing, knowledge representation and machine learning, including information extraction, ontology learning, linguistic resources, grammatical inference, speech synthesis and natural language processing infrastructures. He is the author of the **Ellogon** natural language engineering platform. He is a member of the program committees of several international conferences and he has been involved in more than 15 European and national research projects. As a visiting professor at University of Patras he has taught both undergraduate and postgraduate courses. His work has been published in more than 50 international journal, conferences and books. He is the treasurer and a member of the board of the Greek Artificial Intelligence Society (EETN). Finally he is co-founder of “Intellitech”, a Greek company specializing in natural language processing.

James Pustejovsky
Professor & Chair
Department of Computer Science
Brandeis University, United States
jamesp@cs.brandeis.edu http://jamespusto.com

Professor James in the TJX Feldberg Professor of Computer Science at Brandeis University, and chair of both the Linguistics Program and the Computational Linguistics MA Program. He first started working seriously on annotation in 2002, when he lead the creation and development of **TimeML** and **TimeBank**, in the context of a six-month iARPA (ARDA) workshop. This was then incorporated into **ISO-TimeML**, which has been adopted as an ISO standard. His involvement with ISO began in 2006, where he is the sub-chair responsible for **SC 4** within **TC 37**. In 2008, he started the working group on spatial annotation in language, which has been adopted recently as the ISO standard, **ISOspace**. TimeML has been used as the reference annotation specification, and TimeBank the reference gold standard for all of the SemEval shared task challenges, TempEval and their affiliates. Likewise, this year at SemEval 2015, SpaceBank and ISOspace were adopted as the corpus and associated annotation standard for Task 8, SpaceEval. In 2012, Professor Pustejovsky and Amber Stubbs released an O'Reilly book on "Natural Language Annotation for Machine Learning", which is intended as a guide to the ins and outs of MATTER, the development cycle for modeling, annotating, training, and testing with ML algorithms. He is, along with Nancy Ide, finishing up a comprehensive "**Handbook of Natural Language Annotation**", to be published later this year by Springer. Other annotation projects he has been involved with include: factuality and veridicality (**FactBank**), with Roser Sauri; semantics of images (ImageML), with Julia Bosque Gil; and temporal annotation for the clinical domain (**THYME**), with Guergana Savova and Martha Palmer.

Anna Rumshisky
Assistant Professor
Department of Computer Science
University of Massachusetts at Lowell, United States
arum@cs.uml.edu http://www.cs.uml.edu/~arum/

Professor Anna Rumshisky received her Ph.D. in Computer Science from Brandeis University in 2009, followed by postdoctoral training at the MIT Computer Science and Artificial Intelligence Lab. Her research primarily concerns natural language processing applications in clinical informatics, computational lexical semantics, temporal reasoning, and digital humanities and social science. She has been directly involved in several large-scale annotation initiatives, including **Corpus Pattern Analysis** and **TimeML**. She has co-organized 2012 **i2b2** Workshop on Challenges in NLP for Clinical Data, overseeing the development and release of the first large-scale corpus of temporally annotated narrative provider notes. She co-developed **Generative Lexicon Markup Language (GLML)** for the annotation of compositional operations in text and co-organized SemEval-2010 Task 7 on Argument Selection and Coercion, based on GLML. She has developed methods for decomposition of complex annotation tasks into pairwise similarity judgments tasks for Amazon Mechanical Turk. She presented this methodology in a 2014 tutorial on Deep Semantic Annotation with Shallow Methods given at the International Conference on Lexical Resources and Evaluation.

Gary Simons
Chief Research Officer
SIL International, United States
gary_simons@sil.org http://www.sil.org/~simonsg

Gary F. Simons is currently the Chief Research Officer for SIL International in Dallas, Texas and Executive Editor of the Ethnologue (http://www.ethnologue.com). He has contributed to the development of cyberinfrastructure for linguistics as co-founder of the **Open Language Archives Community** (http://www.language-archives.org), co-developer of the ISO 639-3 standard of three-letter identifiers for the known languages of the world, and co-developer of linguistic markup for the **Text Encoding Initiative (TEI)**. He was formerly Director of Academic Computing for SIL International in which role he oversaw the development of linguistic annotation tools like **IT (Interlinear Text Processor)**, **CELLAR (Computing Environment for Linguistic, Literary, and Anthropological Computing)**, **LinguaLinks**, and the beginnings of **FLEx (FieldWorks Language Explorer)**. He holds a Ph.D. in general linguistics (with minor emphases in computer science and classics) from Cornell University.

Han Sloetjes
Software Developer
Max Planck Institute for Psycholinguistics, The Netherlands
han.sloetjes@mpi.nl http://www.mpi.nl/people/sloetjes-han

Han Sloetjes has been working as a software developer at the Max Planck Institute for Psycholinguistics since 2003. In 2004 I joined the group of developers working on the multimedia annotation tool **ELAN**, a tool that emerged sometime at the end of the nineties. Between 2006 and 2007 he became the main developer responsible for maintaining and extending ELAN, at which point he also became the main developer providing support and training.

Brett South
Research Scientist
Department of Biomedical Informatics
University of Utah, United States
brett.south@hsc.utah.edu

Dr. Brett South has extensive experience leading annotation projects and manual human review efforts that support development of clinical NLP systems. He has 18 years of academic and professional experience in various research, leadership and operational roles. He is currently a Senior Scientist and Post-doc working under the primary mentorship of Professor Wendy Chapman in the Department of Biomedical Informatics, University of Utah. Previously he was a Senior NLP Research Engineer for the **Nuance Clinical Language Understanding Group** where he helped lead a group of 80 clinical language analysts tasked with large-scale semantic annotation of clinical corpora to support development of a computer-assisted coding module. Previously he was with the Division of Epidemiology and VA Salt Lake City IDEAS Center as a Senior Research Scientist serving in several investigative and co-investigative roles on the VA CHIR/VINCI collaborations. Prior to completing his Ph.D. in Biomedical Informatics from University of Utah, Mr. South obtained his Master's degree in Health System's Management from University of Maryland, Baltimore. His research interests include: clinical NLP, integrating efficiencies with manual human review tasks via improvements in tools, workflow modifications, or distributed review, human cognition, and data analysis.

Pontus Stenetorp
Researcher
University of Tokyo, Japan
pontus@stenetorp.se http://pontus.stenetorp.se

Dr. Pontus Stenetorp is a post-doctoral researcher in Natural Language Processing. He is mainly interested in Natural Language Processing and Machine Learning, more specifically, parsing, information extraction, annotation tooling, deep learning, and representation learning. He is co-creator of the annotation tool **brat**.

Stephanie Strassel
Senior Associate Director
Linguistic Data Consortium
University of Pennsylvania, United States
strassel@ldc.upenn.edu https://www.ldc.upenn.edu

Stephanie Strassel oversees the Linguistic Data Consotium's Annotation and Collection Groups and is responsible for directing all aspects of data collection and creation, human subject research and linguistic annotation for a diverse set of externally sponsored human language technology research programs and evaluation campaigns. She has acted as PI or Co-PI on multiple efforts including most recently DARPA DEFT, BOLT, RATS, MADCAT, GALE and Machine Reading; IARPA ALADDIN; NIST OpenMT, LRE, SRE, OpenHaRT, TAC KBP, Rich Transcription, VAST and HAVIC. Previous projects include **TIDES**, **EARS**, **ACE**, **Phanotics** and **TRECVid**. In her work on sponsored projects Ms. Strassel has overseen all aspects of corpus creation including linguistic annotation for a diverse set of technologies including machine translation, speech recognition, question answering, handwriting recognition, document classification, topic detection, information extraction, knowledge base population, natural language understanding, speaker identification, dialect recognition, multimodal event detection and related areas.

To date Ms. Strassel has co-authored over 100 corpora published in LDC's catalog, with several dozen additional corpora pending publication. Ms. Strassel has experience with virtually every type of linguistic annotation and has led LDC efforts to define dozens of new annotation and data collection protocols, many of which are widely used in resource creation elsewhere. She has experience with resource creation and annotation for over 40 linguistic varieties including several low resource languages. At LDC Ms. Strassel directs a staff of 13 full-time and 75+ part-time linguists, managers, researchers and programmers and maintains a large network of independent contractors for a variety of languages. She provides leadership for LDC senior staff, developing and disseminating infrastructure for activities that support multiple functional areas. She frequently serves as an external expert on language resource creation activities via review panels and oversight committees.

Marc Verhagen
Senior Research Scientist
Department of Computer Science
Brandeis University, United States
marc@cs.brandeis.edu http://www.cs.brandeis.edu/~marc

Dr. Marc Verhagen is a Senior Research Scientist at Brandeis University, prior to which he was co-founder and master toolbuilder at **LingoMotors**. He holds a Ph.D. in computer science, an M.A. in computational Linguistics and a B.A. in Geography. Recently, he has worked on temporal and spatial processing, relation extraction from PubMed abstracts, technology extraction from technical texts, and interoperable web services for Natural Language Processing. Verhagen is the main developer of the **Tarsqi Toolkit** for temporal processing and the creator of various simple annotation tools as well as the web-based **Brandeis Annotation Tool (BAT)**. He participated in the creation of the **TimeML** and **ISO-Space** annotation languages. In the context of organizing the **TempEval-1** shared task, he oversaw annotation of the task data, creating ad hoc annotation tools in the process.