

**Joint ISA-7, SRSL-3 and I2MRT Workshop on Interoperable
Semantic Annotation**

26 May 2012

ABSTRACTS

Editors:

Harry Bunt, Manuel Alcantara-Plá, Peter Wittenburg

Workshop Programme

08.30 – 08:30 Registration

08:45 - 09:00 Workshop Opening

09:00 - 10:30 **Session: Semantic representation and multimodal resources**

09:00 - 09:30 Mehdi Manshadi and James Allen: *A Universal Representation for Shallow and Deep Semantics*

09:30 - 10:00 Rodolfo Delmonte and Agata Rotondi: *Treebanks of Logical Forms: They are Useful Only if Consistent*

10:00 - 10:30 Hennie Brugman and Mark Lindeman: *A Publication Platform for Open Annotations*

10:30 - 11:00 coffee break

11:00 - 13:00 **Session: Annotation of spatial information**

11:00 - 11:30 James Pustejovsky, Jessica Moszkowics and Marc Verhagen: *The Current Status of ISO-Space*

11:30 - 12:00 Robert Gaizauskas, Emma Barker, Ching-Lan Chang, Leon Derczynski, Michael Phiri and Chengzhi Peng: *Applying ISO-Space to Healthcare Facility Design Evaluation Reports*

12:00 - 12:30 Antje Müller: *Location and Path - Annotating Sense of the German Prepositions “auf” and “über”*

12:30 - 13:00 Linda Meini, Giovanna Marotta, Leonardo Lenci and Margherita Donati: *An XML Annotation Scheme for Space in an Italian Corpus*

13:00 - 14:00 lunch break

14:00 - 16:00 **Session: Semantic Roles and their annotation**

14:00 - 14:30 *Project ISO-Semantic Roles* (Martha Palmer)

14:30 - 15:00 Claire Bonial, Weston Feely, Jena Hwang and Martha Palmer: *Empirically Validating VerbNet using SemLink*

15:00 - 16:00 *The Lexlink project* (Collin Baker, Christiane Fellbaum, Martha Palmer)

16:00 - 16:30 tea break

16:30 - 18:00 **Session: Interoperable semantic annotation in ISO projects**

16:30 - 17:00 Kiyong Lee: *Interoperable Spatial and Temporal Annotation Schemes*

17:00 - 17:30 Harry Bunt, Rashmi Prasad and Aravind Joshi: *First Steps Towards an ISO Standard for Annotating Discourse Relations*

17:30 - 18:00 *Project ISO-Basics: Principles of Semantic Annotation* (Harry Bunt)

18:00 Workshop Closing

Workshop Organizers

Harry Bunt	Tilburg University
Manuel Alcantara-Plá	Universidad Autónoma de Madrid
Peter Wittenburg	Max Planck Institute for Psycholinguistics, Nijmegen
Thierry Declerck	DFKI, Saarbrücken
Dafydd Gibbon	University of Bielefeld
Nancy Ide	Vassar College, Poughkeepsie, NY
Steven Krauwer	Universiteit Utrecht
Kiyong Lee	Korea University, Seoul
Lorenza Mondada	Université de Lyon 2
James Pustejovsky	Brandeis University, Waltham, MA
Laurent Romary	INRIA/Humboldt Universität Berlin
Oliver Schreer	Fraunhofer Institute for Telecommunications, Berlin

Workshop Programme Committee

Jan Alexandersson	DFKI, Saarbrücken
Stefan Baumann	Universität Köln
Jonas Beskow	KTH, Stockholm
Paul Buitelaar	National University of Ireland, Galway
Harry Bunt	Tilburg University
Thierry Declerck	DFKI, Saarbrücken
Raquel Fernandez Rovira	Universiteit van Amsterdam
Anette Frank	Universität Heidelberg
Dafydd Gibbon	Universität Bielefeld
Koiti Hasida	AIST, Tokyo
Nancy Ide	Vassar College, Poughkeepsie, NY
Michael Kipp	University of Applied Sciences, Augsburg
Kiyong Lee	Korea University, Seoul
Inderjeet Mani	Chiang Mai, Thailand
Jean-Clause Martin	LIMSI, Orsay
Lorenza Mondada	Université de Lyon 2
Martha Palmer	University of Colorado, Boulder
Volha Petukhova	Vicomtech, San Sebastian
Andrei Popescu-Belis	Idiap, Martigny, Switzerland
Rarhmi Prasad	University of Wisconsin, Milwaukee
James Pustejovsky	Brandeis University, Waltham, MA
Laurent Romary	INRIA/Humboldt Universität Berlin
Oliver Schreer	Fraunhofer Institute for Telecommunications, Berlin
Mark Steedman	University of Edinburgh
Mariët Theune	Universiteit Twente
Isabel Trancoso	INESC, Lisbon

Introduction

Three initiatives have joined forces in this workshop, which is concerned with issues in semantic annotation for language resources, especially in relation to spoken and multimodal language data, and with the interoperability and integration of resources and tools.

ISA-7 is the Seventh Workshop on Interoperable Semantic Annotation, and forms part of a series of workshops of ISO TC 37/SC 4 (Language Resources) jointly with ACL-SIGSEM (Computational Semantics). These workshops bring together experts in the annotation of semantic information as expressed in text, speech, gestures, graphics, video, images, and in multiple modalities combined. Examples of semantic annotation include the markup of events, time, space, dialogue acts, discourse relations, and semantic roles, for which the ISO organization pursues the establishment of annotation standards, in order to support the creation of interoperable semantic resources.

SRS-3 is the Third Workshop on Semantic Representation of Spoken Language in Speech and Multimodal Corpora. In these workshops researchers convene who are working on speech and multimodal resources for the semantic annotation of related corpora, and take their inspiration from the observation that the semantic gap between the content conveyed by speech and other modalities and their formal representation is a burning issue in a range of tasks such as content mining, information extraction, dialogue processing, interactive story-telling, assisted health care, and human-robot interaction.

I2MRT (Integration and Interoperability for Multimodal Resources and Tools) is an initiative to address infrastructure aspects of the creation and use of interoperable multimodal resources. Main objectives of I2MRT are to create awareness of the need to make multimodal data visible via standardized methods and accessible via registered data centers; to discuss possibilities of harmonization and standardization of multimodal annotation schemes and possible mappings between schemes; to discuss ways to make cutting-edge technologies available to multimodality researchers; and to build a community that is committed to work further on these issues.

Harry Bunt
Manuel Alcantar-Plá
Peter Wittenburg

Session Semantic Representation and Multimodal Resources

Saturday 26 May, 9:00 – 10:30

Chairperson:

A Universal Representation for Shallow and Deep Semantics

Mehdi Manshadi and James Allen

We define a graphical semantic representation that readily captures the partial semantic analyses produced by shallow processing techniques, yet is also as fully expressive as the representations used in deep analysis systems, including discourse processing. While in most existing natural language systems, robustness often comes at the expense of shallowness, our representation is designed to bridge this gap. The framework is not specific to a particular semantic theory, and may be translated into various target languages. In particular, the translation into first order or intentional logic is transparent. We show how the framework is able to capture more complex semantic phenomena, such as scopal adverbials and predicate modifiers. The graphical frameworks allows us to define mathematical notions to determine the well-formedness of a representation or the coherence of the corresponding sentence once we have the complete semantic representation of a sentence. A unique property of our semantic framework is to encode some syntactic properties of a sentence as well. We define an evaluation framework for this formalism that allows one to compute semantic recall and precision measures given gold standard representations. e

Trebanks of Logical Forms: They are Useful Only if Consistent

Rodolfo Delmonte and Agata Rotondi

Logical Forms are an exceptionally important linguistic representation for highly demanding semantically related tasks like Question/ Answering and Text Understanding, but their automatic production at runtime is highly error-prone. The use of a tool like XWNet and other similar resources would be beneficial for all the NLP community, but not only. The problem is: Logical Forms are useful as long as they are consistent, otherwise they would be useless if not harmful. Like any other resource that aims at providing a meaning representation, LFs require a big effort in manual checking order to reduce the number of errors to the minimum acceptable – less than 1% - from any digital resource. As will be shown in detail in the paper, the available resources – XWNet, WN30-lfs, ILF - suffer from lack of a careful manual checking phase, and the number of errors is too high to make the resource usable as is. We classified mistakes by their syntactic or semantic type in order to facilitate a revision of the resource that we intend to do using regular expressions. We also commented extensively on semantic issues and on the best way to represent them in Logical Forms.

A Publication Platform for Open Annotations

Hennie Brugman and Mark Lindeman

Abstract The OpenAnnotation Consortium introduced a generic model for representing annotations of resources and resource segments that complies to principles of the World Wide Web and Linked Data. This paper introduces a platform for storing, retrieving, searching, exchanging, harvesting and publishing Open Annotations on the web. It describes design considerations, functionality and architecture. Our Open Annotation server platform is set up as a distributed system with server instances that can exchange annotations in a peer-to-peer way. Each instance can persistently

publish annotations using principles of the web and thereby adds ‘annotatability’ to annotations themselves and to annotation ‘Bodies’. Additionally, the annotation platform provides efficient search and implements a Dashboard for server management tasks.

The web-oriented nature of the platform raises a number of interesting issues and opportunities that are discussed in some depth. For example, in general uploaded annotations do not have resolvable http URIs. Assigning those is not trivial. Indexing strategy, determining the boundaries of an annotation in an RDF graph and searching for annotations whose Body is somewhere else on the web are other issues that are discussed.

Session Annotation of spatial information

Saturday 26 May, 11:00 – 13:00

Chairperson:

The Current Status of ISO-Space

James Pustejovsky, Jessica L. Moszkowicz and Marc Verhagen

We report on ISO-Space version 1.4, an annotation specification for capturing spatial and spatiotemporal information in natural language that is now in its fourth incarnation. This version substantially improves upon earlier ISO-Space specifications in a few notable ways. The representation of locations is no longer overloaded such that geolocations have a more complete annotation and non-geolocations are captured with specific tags. In addition, interactions with existing annotation standards such as TimeML have been clarified. The treatment of spatial prepositions has been modified so that their annotation is more suggestive of what spatial relationships should hold between two spatial objects. Finally, spatial relationships are now captured with four distinct link tags: qualitative spatial links for topological relationships, orientation links for non-topological relations, movement links for motion, and measure links for detailing a metric relationship between two spatial objects or what the dimensions of a particular object are. The most recent version of the specification is presented with illustrative examples. We conclude with some outstanding issues that have yet to be captured in the specification.

Applying ISO-Space to Healthcare Facility Design Reports

Robert Gaizauskas, Emma Barker, Ching-Lan Chang, Leon Derczynski, Michael Phiri and Chengzhi Peng

This paper describes preliminary work on the spatial annotation of textual reports about healthcare facility design to support the long-term goal of linking report content to a three-dimensional building model.

Emerging semantic annotation standards enable formal description of multiple types of discourse information. In this instance, we investigate the application of a spatial semantic annotation standard at the building-interior level, where most prior applications have been at inter-city or street level.

Working with a small corpus of design evaluation documents, we have begun to apply the ISO-Space specification to annotate spatial information in healthcare facility design evaluation reports. These reports present an opportunity to explore semantic annotation of spatial language in a novel situation.

We describe our application scenario, report on the sorts of spatial language found in design evaluation reports, discuss issues arising when ISO-Space is applied to building-level entities, and propose possible extensions to ISO-Space to address the issues encountered.

Location and Path – Annotating Senses of the German Prepositions “auf” and “über”

Antje Müller

Many difficulties concerning so-called spatial prepositions arise from an insufficient subclassification of the prepositions' interpretations. Since there is no one-to-one mapping from possible locations to prepositions there is a substantial need to differentiate the diverse interpretations of one preposition. In this paper we present an approach for a subclassification of some spatial prepositions. We will focus on the correlation between the German route prepositions *über* and *durch* and their static local counterparts *auf* and *in*.

Route prepositions are often considered to be decomposable in a PATH function and a location. We will show that this assumption plus an adequate description of the underlying location results in a systematic classification of preposition senses. It is useful for the annotation of spatial preposition senses as well as for the analyses of the interpretations. For annotation the spatial interpretations are organized in a categorization tree. On the way through the tree different features are picked up that determine the respective interpretation. So every interpretation can be characterized as a set of features paired with the form of the preposition. This set-theoretic view of interpretations makes semantic relations between different interpretations of one and the same prepositions as well as between related interpretations of different prepositions apparent.

An XML Annotation Scheme for Space in an Italian Corpus

Linda Meini, Giovanna Marotta, Leonardo Lenci, and Margherita Donati

The new resource we present consists of a corpus of oral spatial descriptions performed by congenital blind and sighted Italian subjects. The collection of the data is part of a wider project on semantic representations in the language of the blind, carried out at the Department of Linguistics, University of Pisa. The long-term goal of the project is to use the evidence collected on congenital blind subjects to get at a better understanding of the relationship between linguistic and perceptive information. The corpus is currently being enhanced with different layers of annotation, focusing on spatial information. The annotation allows us to highlight the effect of the specific lexical and grammatical features of Italian on the encoding of space (e.g. with respect to the way spatial relations are encoded in motion verbs). Our resource is not only one of the few annotated corpora of spoken Italian, but it is also the first one that focuses on spatial categories.

Session: Semantic Role Annotation and Definition

Saturday 26 May, 14:00 – 16:00

Chairperson:

Empirically Validating VerbNet Using SemLink

Claire Bonial, Weston Feely, Jena D. Hwang and Martha Palmer

This research describes efforts to empirically validate a lexical resource, VerbNet, using the PropBank annotations found in the SemLink corpus. As a test case, we examine the frequency with which verbs in SemLink appear in the Caused-Motion syntactic frame: NP-V-NP-PP (e.g., *textit{She poured water into the bowl}*). To do this, we find the frequency with which a given verb

is used in this construction, we then determine each verb's VerbNet class membership, and compare the overall frequency of the Caused-Motion construction in the verb class to how the verbs' behavior is currently represented in VerbNet. We find evidence that VerbNet's current classification fails to capture generalizations about the likelihood of a class' compatibility with the Caused-Motion construction. Specifically, classes where Caused-Motion is currently represented in VerbNet as a characteristic syntactic frame were found to have a lower frequency of realization in that frame than other classes where Caused-Motion is not represented. We therefore suggest augmenting VerbNet's classification with information on the probability that a class will participate in a certain syntactic frame, and given the challenges of this research, offer potential improvements for increasing the interoperability of VerbNet.

Session: ISO projects on semantic annotation

Date / Time [Saturday 26 May, 16:30 – 18:00]

Chairperson:

Interoperable Spatial and Temporal Annotation Schemes

Kiyong Lee

ISO-TimeML (2012) was just published as an international standard for the annotation of temporal and event-related information in language. Almost at the same time, Pustejovsky and Moszkowicz (2012) produced a revised version of ISO-Space specifications as a spatial annotation scheme. The purpose of this paper is to argue for the need of making these two annotation schemes interoperable to allow a unified treatment of annotating spatial and temporal information in language. This task is mainly motivated by many occurrences of spatio-temporal signals (e.g., at, in, after) in text that trigger both spatial and temporal relations between various types of basic elements annotated to text offsets or segments, called markables. We argue that these two semantic annotation schemes can be made interoperable by merging some of their specifications, especially concerning the use of spatial or temporal signals and those relations triggered by these signals and, furthermore, that this merging results in designing an integrated spatio-temporal annotation and interpretation scheme

First Steps Towards an ISO Standard for Annotating Discourse Relations

Harry Bunt, Rashmi Prasad and Aravind Joshi

This paper describes initial studies in the context of a new effort within ISO to design an international standard for the annotation of discourse with relations that account for its coherence, in particular so-called 'discourse relations'. This effort takes the Penn Discourse Treebank (PDTB) as its starting point, and applies a methodology for defining semantic annotation languages which distinguishes an abstract syntax, defining annotation structures as set-theoretical constructs, a concrete syntax, that defines a reference XML-based format for representing annotation structures, and a formal semantics. A first attempt is described to formulate an abstract syntax and a concrete syntax for the annotation scheme underlying the PDTB. The abstract syntax clearly shows an overall structure for a general-purpose standard for annotating discourse relations, while the resulting concrete syntax is much more readable and semantically transparent than the original format. Moreover, some additional elements are introduced which have an optional status, making the proposed representation format compatible not only with the PDTB but also with other approaches.