1 Tutorial Description

In this tutorial, we demonstrate how elements of Generative Lexicon Theory (GL) can be used to help enrich both established and developing lexical and computational semantic resources within the CL community. This includes lexicons, ontologies, annotation schemes, and annotated corpora (WordNet, VerbNet, PropBank, FrameNet, AMR, GMB, SIMPLE, and others).

The tutorial is organized into two parts. The first part aims to acquaint the audience – computational linguists, natural language engineers, and language resource developers – with the basic assumptions and components of the theory and motivate theoretical decisions through evidence-based analysis over large linguistic datasets. We show how the theory models the interaction between lexical information and other components of grammar; in particular, how it mediates various problems in the mapping from lexical semantic representations to syntactic forms and, to a lesser extent, to pragmatic interpretation. We discuss the significant developments of the theory since the original statement in Pustejovsky (1995), including the elaboration of a general theory of semantic selection and semantic typing (Asher and Pustejovsky 2006, Pustejovsky 2011). Finally, we illustrate how the theory has drawn increasingly on the findings of corpus linguistics and distributional semantic analysis and procedures (Pustejovsky and Jezek, 2008, Pustejovsky and Rumshisky, 2008, Jezek and Quochi, 2010, Jezek and Vieu, 2014), thus creating a new dimension of evidence-based analysis and interpretation, which gives rise to an integration of empirical analysis and theoretical modeling. Some of the most difficult problems recently addressed by GL include: how to encode the dynamic interpretation of events and their participants (Pustejovsky 2013, Jezek and Pustejovsky 2015); the extension of the Telic qualia role to verbs, e.g., rationale and purpose clauses; how to
distributionally model the range and effect of coercion phenomena, incorporating a CPA methodology (Hanks and Pustejovsky 2005, Hanks 2013, Jezek et al. 2014) and a broader notion of context.

In part two, we explore how the semantic phenomena illustrated in part one are implemented and handled in existing resources, examining several case studies: VerbNet, WordNet, and AMR. We demonstrate how both the representational facilities and the compositional mechanisms native to GL can simplify and extend the theoretical infrastructure of these resources. In particular, we propose enhancements to VerbNet (Palmer 2009) and AMR (Banarescu et al. 2014) leveraging the work on dynamic event structure and argument encoding presented in part 1. We then show, following a proposal in Fellbaum (2013), how WordNet verb links can be enriched with Telic qualia values, to encode the purpose and goals associated with particular verbs. Finally, we illustrate how GL-based annotation strategies, e.g., GLML, can help in the identification and markup of metonymic selectional ambiguities, as well as Noun-Noun compounds and Adjective-Noun modification interpretations (Pustejovsky et al 2014).

2 Tutorial Outline

1. Introduction to GL (1 hour)
   (a) Basic GL concepts and GL Notational Language
   (b) Qualia Structure
   (c) Events and their Participants
   (d) Meaning Composition in GL: encoding selection, coercion, sub-selection, co-composition

2. Enriching Lexical Resources with GL (2 hours)
   (a) Case Study 1: Enriching VerbNet with Dynamic Event Structure
   (b) Case Study 2: Enriching Abstract Meaning Representation with Dynamic Argument Structure
   (c) Case Study 3: Enhancing WordNet Verb and Noun Ontology with Telic and Purpose relations
   (d) How Corpus Annotation can be enriched with GL representations and mechanisms/relations
3 Motivation and Topics of Interest

Recently, techniques and strategies for the acquisition of lexical semantic information for natural language resources have changed dramatically, influenced by the availability of ever-larger corpora, distributional methods, and newly annotated or semi-annotated corpora. In spite of these developments, however, researchers interested in creating lexical resources still face the problem of anchoring the selection of linguistic features used in the acquisition of information to a model which is theoretically well-developed, while overcoming common problems such as data sparsity and lexical ambiguity. Semantic feature do no always emerge from a purely corpus-based distributional analysis (Pustejovsky and Jezek 2008); moreover, there is often no consensus on what features to use for general acquisition tasks, and in many cases, the feature sets are constructed ad-hoc to address the goals of the specific task. Because GL has long approached these problems of polysemy, type coercion, metonymy, and co-composition from a systematic and theoretical perspective, it is worth examining how the theory can contribute to enriching and extending existing lexical resources which have emerged within the CL community.

GL has already been exploited as a theoretical background in language resources. Perhaps the most significant contribution of GL to computational lexicography took place in the framework of the EU-sponsored SIMPLEX project (Semantic Information for Multipurpose Plurilingual Lexicons), whose aim was to develop comprehensive semantic lexicons for 12 European languages. In this context, an extended version of the Qualia Structure was proposed (Lenci et al 2000). Further, qualia structure was proposed as an organizing principle for the top ontology in EuroWordNet (Vossen 2001). GL Semantic typing has also been extensively used in the construction of PDEV (Pattern Dictionary of English Verbs, Hanks and Pustejovsky 2005), where semantic distinctions among the different senses of verbs depend on the semantic type of the arguments, as predicted by the co-composition principle, as well as in the design of the Brandeis Semantic Ontology (Pustejovsky et al 2006, Havasi et al, 2009). Finally, GL’s event structure was developed into a subeventual lexical resource in Im (2013) that explores the principles of opposition structure and change in GL.

In this tutorial we make use of this background and of recent work to propose enhancements to existing resources widely used in the community. For all these reasons a tutorial illustrating how GL principles can be put into practice in linguistic analysis and lexical resource building, will benefit students and researchers interested in theoretical linguistics, computational
semantics, and language resource development.

4 Instructors

James Pustejovsky holds the TJX Feldberg Chair in Computer Science at Brandeis University, where he directs the Lab for Linguistics and Computation, and chairs both the Program in Language and Linguistics and the Computational Linguistics Graduate Program. He has conducted research in computational linguistics, AI, lexical semantics, temporal reasoning, and corpus linguistics and language annotation. He has written several books on computational semantics, computational linguistics, and corpus processing. He has authored numerous books, including *Generative Lexicon*, MIT, 1995; *Semantics and the Lexicon*, Springer, 1993; *The Problem of Polysemy*, CUP, 1996 (with B. Boguraev); *The Language of Time*, OUP, 2005 (with I. Mani and R. Gaizauskas), *Interpreting Motion: Grounded Representations for Spatial Language*, OUP, 2012 (with I. Mani), and *Natural Language Annotation for Machine Learning*, O’Reilly, 2012 (with A. Stubbs). Recent books include: *Recent Advances in Generative Lexicon Theory*, Springer, 2013; *A Guide to Generative Lexicon Theory*, OUP, Forthcoming (with Elisabetta Jezek).

Elisabetta Jezek is an Associate Professor at the University of Pavia, where she has taught Syntax and Semantics and Applied Linguistics since 2001. Her research interests and areas of expertise include lexical semantics, verb classification, theory of argument structure, event structure in syntax and semantics, lexicon/ontology interplay, word class systems, and computational lexicography. She has edited a number of major works in lexicography and published contributions focusing on the interplay between corpus analysis, research methodology, and linguistic theory. Her publications include: *Classi di Verbi tra Semantica e Sintassi*, ETS, 2003; *Lessico: Classi di Parole, Strutture, Combinazioni*, Il Mulino, 2005 (2nd ed. 2011); *The Lexicon: An Introduction*, OUP, 2015; and *A Guide to Generative Lexicon Theory*, OUP, Forthcoming (with James Pustejovsky).

References

Banarescu, Laura, Claire Bonial, Shu Cai, Madalina Georgescu, Kira Griffitt, Ulf Hermjakob, Kevin Knight, Philipp Koehn, Martha Palmer, and Nathan Schneider. 2014. *Abstract Meaning Representation (AMR) 1.2 Spec-


Jezek, Elisabetta and James Pustejovsky. 2015. “Dynamic Argument Structure”, Universit` a di Pavia and Brandeis University, manuscript.


