# Corpus Pattern for Semantic Processing

**Patrick Hanks**
University of Wolverhampton, UK
patrick.w.hanks@gmail.com

**Elisabetta Jezek**
University of Pavia, IT
jezek@unipv.it

**Daisuke Kawahara**
Kyoto University, JP
dk@i.kyoto-u.ac.jp

**Octavian Popescu**
IBM Research, US
o.popescu@us.ibm.com

## 1 Introduction

This tutorial presents a corpus-driven, pattern-based empirical approach to meaning representation and computation. Patterns in text are everywhere, but techniques for identifying and processing them are still rudimentary. Patterns are not merely syntactic but syntagmatic: each pattern identifies a lexico-semantic clause structure consisting of a predicator (verb or predicative adjective) together with open-ended lexical sets of collocates in different clause roles (subject, object, prepositional argument, etc.). If NLP is to make progress in identifying and processing text meaning, pattern recognition and collocational analysis will play an essential role, because:

> Many, if not most meanings, require the presence of more than one word for their normal realization. ... Patterns of co-selection among words, which are much stronger than any description has yet allowed for, have a direct connection with meaning. (J. M. Sinclair, 1998).

The tutorial presents methods for building patterns on the basis of corpus evidence, using machine learning methods. It discusses some possible applications of pattern inventories and invites discussion of others. It is intended for an audience with heterogeneous competences but with a common interest in corpus linguistics and computational models for meaning-related tasks in NLP. We report on the methodologies for building resources for semantic processing and their contribution to NLP tasks. The goal is to provide the audience with an operative understanding of the methodology used to acquire corpus patterns and of their utility in NLP applications.

## 2 Overview

Natural language sentences make use of lexical, syntactic, semantic and pragmatic information in order to fulfill their role of conveying meaning. Previous research on computing the meaning of linguistic expressions - from approaches which consider overt distributional information on words to deep semantic ones, based on first order and lambda calculus representations - has highlighted two major issues: (1) the appropriate level of formalization for meaning representation cannot be founded only on premises derived from prior experience, (2) the lack of large-scale annotated corpora which combine different levels of semantic annotation hinders the development of machine-learning applications. In particular, in the framework of big data analytics for semantically processing large corpora, these two issues must be addressed.

The regular structure of normal clauses can be used as a basis in order to learn the rules that lie behind recurrent meaningful constructs in natural language. It has been shown (Hanks&Pustejovsky 2004, Pustejovsky&Jezek 2008, Popescu&Magnini 2007, Popescu 2013, Kawahara et al. 2014)



*Proceedings of the Tutorials of the 53rd Annual Meeting of the ACL and the 7th IJCNLP*, pages 12–15,
Beijing, China, July 26-31, 2015. 

that it is possible to identify and to learn corpus patterns that encode the information that accounts for the senses of the verb and its arguments in the context. These patterns link the syntactic structure of clauses and the semantic types of argument fillers via the role that each of these play in the disambiguation of the clause as a whole. With regard to irregularities, there are quite a few clauses in a corpus where these patterns do not seem to match the text, because of the apparent incompatibility between the actual and the expected semantic types of the arguments (Jezek&Hanks 2010, Hanks 2012). However, it is possible to build statistical models that simultaneously generate both the regular and the innovative representation of a clause. Available solutions developed up to now range from supervised to totally unsupervised approaches. The patterns obtained encode the necessary information for handling the meaning of each word individually as well as that of the clause as a whole. As such they are instrumental in building better language models (Dligach&Palmer 2011). In the contexts matched by such patterns, any word is unequivocally disambiguated. The semantic types used in pattern representation play a discriminative role, therefore the patterns are sense discriminative and as such they can be used in word sense disambiguation and other meaning-related tasks (see among others Pustejovsky et al. 2004, Cumby&Roth 2003, Popescu&Magnini 2007, Pustejovsky et al. 2010, Popescu et al. 2014). Also, the meaning of a pattern as a whole is expressed as a set of basic implicatures. The implicatures are instrumental in textual entailment, semantic similarity and paraphrasing generation (Popescu et al. 2011, Nicolae&Popescu 2013, Vo et. al 2014). Depending on the proposed application, the implicatures associated with a pattern may be expressed in any of a wide variety of other ways, e.g. as a translation into another language or as a synonym set. The automatic aligning of the set of patterns of two languages via their shared semantic types is used in meaning-preserving translation tasks (Popescu&Jezek 2013).

The relatively recent research on corpus data has shown that intermediate text representations (ITRs), built in a bottom-up manner from corpus examples towards a complex representation of clauses, play an important role in dealing with the meaning disambiguation problem. ITRs offer an important degree of freedom in finding the right cut between various levels of semantic information. Large-scale corpus-driven lexical analysis leads to two apparently contradictory conclusions. On the one hand, the regularities of word use (valencies, collocations) are more regular than what most pre-corpus linguists would have predicted. On the other hand, the irregularities are more irregular. In particular, verb usage in language displays a continuous blend between regular constructs with clearly distinct senses and new and innovative usages. The Theory of Norms and Exploitations (Hanks 2013) maintains that language exhibits mainly a rule-governed behavior, but argues that there is not just one monolithic system of rules. Instead, there are two interactive sets of rules: 1) Norms: a set of rules for using words normally and idiomatically: these are the rules of grammar; they account for 70%-90% of all utterances - depending on the type of the verb, the topic, and the domain. However, they do not account for linguistic creativity, nor for changes in word meaning; 2) Exploitation rules, which account for creativity and innovative usage (about 10%-30% of corpus examples). Exploitation rules also account for phenomena such as meaning shift. Pattern Dictionaries are resources based on Corpus Pattern Analysis (CPA). They contains examples for each category for a large number of English and Italian verbs and are available at http://pdev.org.uk/ (Hanks 2004), and at http://tpas.fbk.eu/resource (Jezek et al. 2014).

The corpus-pattern methodology is designed to offer a viable solution to meaning representation. The techniques we present are widely applicable in NLP and they deal efficiently with data sparseness and open domain expression of semantic relationships.

The tutorial is divided into three main parts, which are strongly interconnected: (A) Building Corpus Patterns via the Theory of Norms and Exploitations, (B) Inducing Semantic Types and Semantic Task Oriented Ontologies, and (C) Machine Learning and Applications of Corpus Patterns.

## 3 Outline

### 3.1 Corpus, Language Usage and Computable Semantic Properties of Verb Phrases section

Basic Computational Semantic Concepts

Theory of Norm and Exploitation of Language Usage

Corpus Pattern Analysis in Sketch Engine

Sense Discriminative Patterns

### 3.2 Semantic Types and Ontologies

Argument Structures

Frames and Semantic Types

Inducing Semantic Types

Discriminative Patterns

### 3.3 Statistical Models for Corpus Pattern Recognition and Extraction. NLP Applications

Finite State Markov Chains

Naive Bayesian and Gaussian Random Fields for Conditional Probabilities over Semantic Types

Latent Dirichlet Analysis for Unsupervised Pattern Extraction

Probably Approximately Correct and Statistical Query Model

Joint Source Channel Model for Recognition of Norm and Exploitation

Textual Entailment, Paraphrase Generation and Textual Similarity with Corpus Patterns

## 4 Tutors

**Patrick Hanks** is Professor in Lexicography at the Research Institute of Information and Language Processing at the University of Wolverhampton. He is also a visiting professor at the Bristol Centre for Linguistics (University of the West of England). He studied English Language and Literature at Oxford and was awarded a PhD in Informatics at the Masaryk University in Brno, Czech Republic. In the 1980s he was the managing editor of Cobuild, an innovative corpus-based dictionary compiled at the University of Birmingham. In 1989-90 he co-authored with Ken Church and others a series of papers on statistical approaches to lexical analysis. For ten years (1990–2000) he was chief editor of Current English Dictionaries at Oxford University Press. He is the author of *Lexical Analysis: Norms and Exploitations* (MIT Press, 2013), which presents a new theory of word meaning and language in use. He is a consultant on lexicographical methodology and definition to several institutions throughout Europe, including Oxford University Press, and is a frequent invited plenary speaker at international conferences on lexicography, corpus linguistics, figurative language, onomastics, and phraseology.

**Elisabetta Jezek** has been teaching Syntax and Semantics and Applied Linguistics at the University of Pavia since 2001. Her research interests and areas of expertise are lexical semantics, verb classification, theory of Argument Structure, event structure in syntax and semantics, corpus annotation, computational Lexicography.

**Daisuke Kawahara** is an Associate Professor at Kyoto University. He is an expert in the areas of parsing, knowledge acquisition and information analysis. He teaches gradu-

ate classes in natural language processing. His current work is focused on automatic induction of semantic frames and semantic parsing, verb polysemic classes, verb sense disambiguation, and automatic induction of semantic frames.

**Octavian Popescu** is a researcher at IBM T. J. Watson Research Center, working on computational semantics with focus on corpus patterns for question answering, textual entailment and paraphrasing. He taught various NLP graduate courses in computational semantics at Trento University (IT), Colorado University at Boulder (US) and University of Bucharest (RO).

## References


C. Cumby and D. Roth "On Kernel Methods for Relational Learning", in Proceedings of ICML 2003, Washington 2003

D. Dligach and M. Palmer: "Good Seed Makes a Good Crop: Accelerating Active Learning Using Language Modeling", in Proceedings of ACL, Oregon, 2011

P. Hanks, "Corpus Pattern Analysis". In Williams G. and S. Vessier (eds) Proceedings of the XI Euralex International Congress, Lorient, Université de Bretagne-Sud, 2004

P. Hanks and J. Pustejovsky. "Common Sense About Word Meaning:Sense in Context", in Proceedings of the TSD, Volume 3206, 2004.

P. Hanks "How People use words to make Meanings. Semantic Types meet Valencies". In A. Bulton and J. Thomas (eds.) Input, Process and Product: Developments in Teaching and Language Corpora. Masaryk University Press, 2012

P. Hanks "Lexical Analysis: Norms and Exploitations.". MITPress 2013

E. Jezek and P. Hanks, "What lexical sets tell us about conceptual categories", In Lexis, E-Journal in English Lexicology, 4, 7-22, 2010.

E. Jezek, B. Magnini, A. Feltracco, A. Bianchini, O. Popescu "T-PAS; A resource of Typed Predicate Argument Structures for linguistic analysis and semantic processing", in Proceedings of LREC, Reykjavik 2014

D. Kawahara, D. Pederson, O. Popescu, M. Palmer 2014. "Inducing Example-based Semantic Frames from a Massive Amount of Verb Uses", in Proceedings of the EACL, Gothenburg, 2014

V. Niculae and O. Popescu, "Determining is-a relationships for Textual Entailment", in Prceedings of JSSP , Trento, 2013

O. Popescu, B. Magnini "Sense Discriminative Patterns for Word Sense Disambiguation", in Proceedings of Semantic Content Acquisition and Representation, NODALIDA, Tartu, 2007.

O. Popescu, E. Cabrio, B. Magnini Journal Proceedings of the IJCAI Workshop Learning by Reasoning and its Applications in Intelligent Question-Answering, Barcelona 2011

O. Popescu , E. Jezek. "Pattern Based Translation", in Proceedings of Tralogy-II, Paris 2013

O. Popescu. "Learning Corpus Pattern with Finite State Automata", in Proceedings of IWSC, Berlin, 2013.

O. Popescu, P. Hanks, M. Palmer, "Mapping CPA onto Ontonotes Senses", in Proceedings of LREC, Reykjavik, 2014

J. Pustejovsky, P. Hanks, and A. Rumshisky. "Sense in Context", in Proceedings of COLING 2004, Geneva, 2004

J. Pustejovsky, E. Jezek "Semantic Coercion in Language: Beyond Distributional Analysis", Italian Journal of Linguistics 20, 1, 181-214, 2008.

J. M.Sinclair "The Lexical Item", in E. WEigand (ed.) Contrastive Lexical Semantics. Benjamins, 1998

N. Vo, O. Popescu, T. Caselli, "FBK-TR: SVM for Semantic Relatedness and Corpus Patterns for RTE", in Proceedings SemEval, Dublin, 2014