COLING 2012

24th International Conference on Computational Linguistics

Proceedings of COLING 2012: Posters

Program chairs:
Martin Kay and Christian Boitet

8-15 December 2012
Mumbai, India

Diamond sponsors

Tata Consultancy Services
Linguistic Data Consortium for Indian Languages (LDC-IL)

Gold Sponsors

Microsoft Research
Beijing Baidu Netcon Science Technology Co. Ltd.

Silver sponsors

IBM, India Private Limited
Crimson Interactive Pvt. Ltd.
Yahoo
Easy Transcription & Software Pvt. Ltd.

Proceedings of COLING 2012: Volume 1
Martin Kay and Christian Boitet (eds.)
Preprint edition
Published by The COLING 2012 Organizing Committee
Mumbai, 2012

Also available online in the ACL Anthology at `http://aclweb.org`

Table of Contents

K-Best Spanning Tree Dependency Parsing With Verb Valency Lexicon Reranking
Zeljko Agic 1

A Best-First Anagram Hashing Filter for Approximate String Matching with Generalized Edit Distance
Malin Ahlberg and Gerlof Bouma 13

Automatic Bilingual Phrase Extraction from Comparable Corpora
Ahmet Aker, Yang Feng and Robert Gaizauskas 23

A Formalized Reference Grammar for UNL-Based Machine Translation between English and Arabic
Sameh Alansary 33

Mapping Arabic Wikipedia into the Named Entities Taxonomy
Fahd Alotaibi and Mark Lee 43

Probabilistic Refinement Algorithms for the Generation of Referring Expressions
Romina Altamirano, Carlos Areces and Luciana Benotti 53

Measuring the Adequacy of Cross-Lingual Paraphrases in a Machine Translation Setting
Marianna Apidianaki 63

Cross-Lingual Sentiment Analysis for Indian Languages using Linked WordNets
Balamurali A.R., Aditya Joshi and Pushpak Bhattacharyya 73

The Creation of Large-Scale Annotated Corpora of Minority Languages using UniParser and the EANC platform
Timofey Arkhangeskiy, Oleg Belyaev and Arseniy Vydrin 83

Collocation Extraction using Parallel Corpus
Kavosh Asadi Atui, Heshaam Faili and Kaveh Assadi Atuie 93

Improved Spelling Error Detection and Correction for Arabic
Mohammed Attia, Pavel Pecina, Younes Samih, Khaled Shaalan and Josef van Genabith 103

Heloise —- A Reengineering of Ariane-G5 SLLPs for Application to π-languages
Vincent Berment and Christian Boitet 113

Machine Translation for Language Preservation
Steven Bird and David Chiang 125

Comparing Non-projective Strategies for Labeled Graph-Based Dependency Parsing
Anders Björkelund and Jonas Kuhn 135

Phrase Structures and Dependencies for End-to-End Coreference Resolution
Anders Björkelund and Jonas Kuhn 145

The Language of Power and its Cultural Influence
David Bracewell and Marc Tomlinson 155

Learning Opinionated Patterns for Contextual Opinion Detection
Caroline Brun 165

Does Similarity Matter? The Case of Answer Extraction from Technical Discussion Forums
Rose Catherine, Amit Singh, Rashmi Gangadharaiah, Dinesh Raghu and Karthik Visweswariah 175

Chinese Noun Phrase Coreference Resolution: Insights into the State of the Art
Chen Chen and Vincent Ng 185

Linguistic and Statistical Traits Characterising Plagiarism
Miranda Chong and Lucia Specia 195

Impact of Less Skewed Distributions on Efficiency and Effectiveness of Biomedical Relation Extraction
Md. Faisal Mahbub Chowdhury and Alberto Lavelli 205

Lattice Rescoring for Speech Recognition using Large Scale Distributed Language Models
Euisok Chung, Hyung-Bae Jeon, Jeon-Gue Park and Yun-Keun Lee 217

Morphological Analyzer for Affix Stacking Languages: A Case Study of Marathi
Raj Dabre, Archana Amberkar and Pushpak Bhattacharyya 225

Modelling the Organization and Processing of Bangla Polymorphemic Words in the Mental Lexicon: A Computational Approach
Tirthankar Dasgupta, Manjira Sinha and Anupam Basu 235

Coreference Clustering using Column Generation
Jan De Belder and Marie-Francine Moens 245

Metric Learning for Graph-Based Domain Adaptation
Paramveer Dhillon, Partha Talukdar and Koby Crammer 255

Automatic Hashtag Recommendation for Microblogs using Topic-Specific Translation Model
Zhuoye Ding, Qi Zhang and Xuanjing Huang 265

Unsupervised Feature-Rich Clustering
Vladimir Eidelman 275

Token Level Identification of Linguistic Code Switching
Heba Elfardy and Mona Diab 287

Parenthetical Classification for Information Extraction
Ismaïl El Maarouf and Jeanne Villaneau 297

A Dictionary-Based Approach to Identifying Aspects Implied by Adjectives for Opinion Mining
Geli Fei, Bing Liu, Meichun Hsu, Malu Castellanos and Riddhiman Ghosh 309

Dealing with Input Noise in Statistical Machine Translation
Lluís Formiga and José A. R. Fonollosa 319

A Comparison of Knowledge-based Algorithms for Graded Word Sense Assignment
Annemarie Friedrich, Nikos Engonopoulos, Stefan Thater and Manfred Pinkal 329

Leveraging Statistical Transliteration for Dictionary-Based English-Bengali CLIR of OCR'd Text
Utpal Garain, Arjun Das, David Doermann and Douglas Oard 339

RU-EVAL-2012: Evaluating Dependency Parsers for Russian
Anastasia Gareyshina, Maxim Ionov, Olga Lyashevskaya, Dmitry Privoznov, Elena Sokolova and Svetlana Toldova 349

Assessing Sentiment Strength in Words Prior Polarities
Lorenzo Gatti and Marco Guerini 361

Improving Dependency Parsing with Interlinear Glossed Text and Syntactic Projection
Ryan Georgi, Fei Xia and William Lewis 371

Diachronic Variation in Grammatical Relations
Aaron Gerow and Khurshid Ahmad 381

Relation Classification using Entity Sequence Kernels
Debanjan Ghosh and Smaranda Muresan 391

Translating Questions to SQL Queries with Generative Parsers Discriminatively Reranked
Alessandra Giordani and Alessandro Moschitti 401

Classifier-Based Tense Model for SMT
Zhengxian Gong, Min Zhang, Chew-lim Tan and Guodong Zhou 411

Extracting and Normalizing Entity-Actions from Users' Comments
Swapna Gottipati and Jing Jiang 421

Expected Divergence Based Feature Selection for Learning to Rank
Parth Gupta and Paolo Rosso 431

LEPOR: A Robust Evaluation Metric for Machine Translation with Augmented Factors
Aaron L. F. Han, Derek F. Wong and Lidia S. Chao 441

Predicting Stance in Ideological Debate with Rich Linguistic Knowledge
Kazi Saidul Hasan and Vincent Ng 451

FeatureForge: A Novel Tool for Visually Supported Feature Engineering and Corpus Revision
Florian Heimerl, Charles Jochim, Steffen Koch and Thomas Ertl 461

Verb Temporality Analysis using Reichenbach's Tense System
André Horie, Kumiko Tanaka-Ishii and Mitsuru Ishizuka 471

A Metric for Evaluating Discourse Coherence based on Coreference Resolution
Ryu Iida and Takenobu Tokunaga 483

Comparing Word Relatedness Measures Based on Google n-grams
Aminul Islam, Evangelos Milios and Vlado Keselj 495

Two-Stage Bootstrapping for Anaphora Resolution
Balaji Jagan, T V Geetha and Ranjani Parthasarathi 507

Explorations in the Speakers' Interaction Experience and Self-Assessments
Kristiina Jokinen 517

Multimodal Signals and Holistic Interaction Structuring
Kristiina Jokinen and Graham Wilcock 527

New Insights from Coarse Word Sense Disambiguation in the Crowd
Adam Kapelner, Krishna Kaliannan, H. Andrew Schwartz, Lyle Ungar and Dean Foster 539

A Unified Sentence Space for Categorical Distributional-Compositional Semantics: Theory and Experiments
Dimitri Kartsaklis, Mehrnoosh Sadrzadeh and Stephen Pulman 549

A Knowledge-Based Approach to Syntactic Disambiguation of Biomedical Noun Compounds
Ramakanth Kavuluru and Daniel Harris 559

Classification of Inconsistent Sentiment Words using Syntactic Constructions
Wiltrud Kessler and Hinrich Schütze 569

Learning Semantics with Deep Belief Network for Cross-Language Information Retrieval
Jungi Kim, Jinseok Nam and Iryna Gurevych 579

Detection of Acoustic-Phonetic Landmarks in Mismatched Conditions using a Biomimetic Model of Human Auditory Processing
Sarah King and Mark Hasegawa-Johnson 589

Learning Verbs on the Fly
Zornitsa Kozareva 599

Decoder-based Discriminative Training of Phrase Segmentation for Statistical Machine Translation
Hyoung-Gyu Lee and Hae-Chang Rim 611

Glimpses of Ancient China from Classical Chinese Poems
John Lee and Tak-sum Wong 621

Conversion between Scripts of Punjabi: Beyond Simple Transliteration
Gurpreet Singh Lehal and Tejinder Singh Saini 633

Development of a Complete Urdu-Hindi Transliteration System
Gurpreet Singh Lehal and Tejinder Singh Saini 643

Random Walks on Context-Aware Relation Graphs for Ranking Social Tags
Han Li, Zhiyuan Liu and Maosong Sun 653

Phrase-Based Evaluation for Machine Translation
Liangyou Li, Zhengxian Gong and Guodong Zhou 663

A Beam Search Algorithm for ITG Word Alignment
Peng Li, Yang Liu and Maosong Sun 673

Active Learning for Chinese Word Segmentation
Shoushan Li, Guodong Zhou and Chu-Ren Huang 683

Fine-Grained Classification of Named Entities by Fusing Multi-Features
Wenjie Li, Jiwei Li, Ye Tian and Zhifang Sui 693

Expert Finding for Microblog Misinformation Identification
Chen Liang, Zhiyuan Liu and Maosong Sun 703

Improving Relative-Entropy Pruning using Statistical Significance
Wang Ling, Nadi Tomeh, Guang Xiang, Isabel Trancoso and Alan Black 713

Expected Error Minimization with Ultraconservative Update for SMT
Lemao Liu, Tiejun Zhao, Taro Watanabe, Hailong Cao and Conghui Zhu 723

Generalized Sentiment-Bearing Expression Features for Sentiment Analysis
Shizhu Liu, Gady Agam and David Grossman 733

Unsupervised Domain Adaptation for Joint Segmentation and POS-Tagging
Yang Liu and Yue Zhang 745

Tag Dispatch Model with Social Network Regularization for Microblog User Tag Suggestion
Zhiyuan Liu, Cunchao Tu and Maosong Sun 755

Summarization of Business-Related Tweets: A Concept-Based Approach
Annie Louis and Todd Newman 765

Towards the Automatic Detection of the Source Language of a Literary Translation.
Gerard Lynch and Carl Vogel 775

Fourth-Order Dependency Parsing
Xuezhe Ma and Hai Zhao 785

A Subjective Logic Framework for Multi-Document Summarization
Sukanya Manna, Byron J. Gao and Reed Coke 797

Manual Corpus Annotation: Giving Meaning to the Evaluation Metrics
Yann Mathet, Antoine Widlöcher, Karën Fort, Claire François, Olivier Galibert, Cyril Grouin, Juliette Kahn, Sophie Rosset and Pierre Zweigenbaum 809

Discriminative Boosting from Dictionary and Raw Text – a Novel Approach to Build a Chinese Word Segmenter
Fandong Meng, Wenbin Jiang, Hao Xiong and Qun Liu 819

Lost in Translations? Building Sentiment Lexicons using Context Based Machine Translation
Xinfan Meng, Furu Wei, Ge Xu, Longkai Zhang, Xiaohua Liu, Ming Zhou and Houfeng Wang 829

How Does the Granularity of an Annotation Scheme Influence Dependency Parsing Performance?
Simon Mille, Alicia Burga, Gabriela Ferraro and Leo Wanner 839

Does Tectogrammatics Help the Annotation of Discourse?
Jiří Mírovský, Pavlína Jínová and Lucie Poláková 853

The Effect of Learner Corpus Size in Grammatical Error Correction of ESL Writings
Tomoya Mizumoto, Yuta Hayashibe, Mamoru Komachi, Masaaki Nagata and Yuji Matsumoto 863

GRAFIX: Automated Rule-Based Post Editing System to Improve English-Persian SMT Output
Mahsa Mohaghegh, Abdolhossein Sarrafzadeh and Mehdi Mohammadi 873

Relational Structures and Models for Coreference Resolution
Truc-Vien T. Nguyen and Massimo Poesio 883

Text Summarization Model based on Redundancy-Constrained Knapsack Problem
Hitoshi Nishikawa, Tsutomu Hirao, Toshiro Makino and Yoshihiro Matsuo 893

Lexical Categories for Improved Parsing of Web Data
Lilja Øvrelid and Arne Skjærholt 903

Text-To-Speech for Languages without an Orthography
Sukhada Palkar, Alan Black and Alok Parlikar 913

Part of Speech (POS) Tagger for Kokborok
Braja Gopal Patra, Khumbar Debbarma, Dipankar Das and Sivaji Bandyopadhyay ... 923

Forced Derivations for Hierarchical Machine Translation
Stephan Peitz, Arne Mauser, Joern Wuebker and Hermann Ney 933

On Panini and the Generative Capacity of Contextualized Replacement Systems
Gerald Penn and Paul Kiparsky 943

Joint Segmentation and Tagging with Coupled Sequences Labeling
Xipeng Qiu, Feng Ji, Jiayi Zhao and Xuanjing Huang 951

Defining Syntax for Learner Language Annotation
Marwa Ragheb and Markus Dickinson 965

How Good are Typological Distances for Determining Genealogical Relationships among Languages?
Taraka Rama and Prasanth Kolachina 975

Sentence Boundary Detection: A Long Solved Problem?
Jonathon Read, Rebecca Dridan, Stephan Oepen and Lars Jørgen Solberg 985

Document and Corpus Level Inference For Unsupervised and Transductive Learning of Information Structure of Scientific Documents
Roi Reichart and Anna Korhonen 995

Light Textual Inference for Semantic Parsing
Kyle Richardson and Jonas Kuhn 1007

Korektor –– A System for Contextual Spell-Checking and Diacritics Completion
Michal Richter, Pavel Straňák and Alexandr Rosen 1019

Using Qualia Information to Identify Lexical Semantic Classes in an Unsupervised Clustering Task
Lauren Romeo, Sara Mendes and Núria Bel 1029

A Strategy of Mapping Polish WordNet onto Princeton WordNet
Ewa Rudnicka, Marek Maziarz, Maciej Piasecki and Stan Szpakowicz 1039

A Hierarchical Domain Model-Based Multi-Domain Selection Framework for Multi-Domain Dialog Systems
Seonghan Ryu, Donghyeon Lee, Injae Lee, Sangdo Han, Gary Geunbae Lee, Myungjae Kim and Kyungduk Kim 1049

A Fully Coreference-annotated Corpus of Scholarly Papers from the ACL Anthology
Ulrich Schäfer, Christian Spurk and Jörg Steffen 1059

Continuous Space Translation Models for Phrase-Based Statistical Machine Translation
Holger Schwenk 1071

Data-driven Dependency Parsing With Empty Heads
Wolfgang Seeker, Richárd Farkas, Bernd Bohnet, Helmut Schmid and Jonas Kuhn . 1081

Extension of TSVM to Multi-Class and Hierarchical Text Classification Problems with General Losses
Sathiya Keerthi Selvaraj, Sundararajan Sellamanickam and Shirish Shevade..... 1091

Calculation of Phrase Probabilities for Statistical Machine Translation by using Belief Functions
Christophe Servan and Simon Petitrenaud 1101

Sense and Reference Disambiguation in Wikipedia
Hui Shen, Razvan Bunescu and Rada Mihalcea 1111

Unsupervised Metaphor Paraphrasing using a Vector Space Model
Ekaterina Shutova, Tim van de Cruys and Anna Korhonen 1121

Memory-Efficient Katakana Compound Segmentation using Conditional Random Fields
Krauchanka Siarhei and Artsimenya Artsiom 1131

New Readability Measures for Bangla and Hindi Texts
Manjira Sinha, Sakshi Sharma, Tirthankar Dasgupta and Anupam Basu 1141

Automatic Question Generation in Multimedia-Based Learning
Yvonne Skalban, Le An Ha, Lucia Specia and Ruslan Mitkov 1151

A More Cohesive Summarizer
Christian Smith, Henrik Danielsson and Arne Jönsson 1161

Robust Learning in Random Subspaces: Equipping NLP for OOV Effects
Anders Søgaard and Anders Johannsen..... 1171

An Empirical Etudy of Non-Lexical Extensions to Delexicalized Transfer
Anders Søgaard and Julie Wulff 1181

Entropy-based Training Data Selection for Domain Adaptation
Yan Song, Prescott Klassen, Fei Xia and Chunyu Kit 1191

Corpus-based Explorations of Affective Load Differences in Arabic-Hebrew-English
Carlo Strapparava, Oliviero Stock and Ilai Alon 1201

Acquiring and Generalizing Causal Inference Rules from Deverbal Noun Constructions
Shohei Tanaka, Naoaki Okazaki and Mitsuru Ishizuka 1209

Advertising Legality Recognition
Yi-jie Tang, Cong-kai Lin and Hsin-Hsi Chen 1219

A Joint Phrasal and Dependency Model for Paraphrase Alignment
Kapil Thadani, Scott Martin and Michael White..... 1229

Sourcing the Crowd for a Few Good Ones: Event Type Detection
Caselli Tommaso and Huang Chu-Ren 1239

Combining Multiple Alignments to Improve Machine Translation
Zhaopeng Tu, Yang Liu, Yifan He, Josef van Genabith, Qun Liu and Shouxun Lin .. 1249

A New Search Approach for Interactive-Predictive Computer-Assisted Translation
Zeinab Vakil and Shahram Khadivi 1261

Automatic Extraction of Polar Adjectives for the Creation of Polarity Lexicons
Silvia Vázquez, Muntsa Padró, Núria Bel and Julio Gonzalo 1271

Optimal Scheduling of Information Extraction Algorithms
Henning Wachsmuth and Benno Stein 1281

Update Summarization Based on Co-Ranking with Constraints
Xiaojun Wan 1291

Sentence Realization with Unlexicalized Tree Linearization Grammars
Rui Wang and Yi Zhang 1301

Exploiting Discourse Relations for Sentiment Analysis
Fei Wang, Yunfang Wu and Likun Qiu 1311

Expansion Methods for Job-Candidate Matching Amidst Unreliable and Sparse Data
Jerome White, Krishna Kummamuru and Nitendra Rajput 1321

A Unified Framework for Discourse Argument Identification via Shallow Semantic Parsing
Fan Xu, Qiaoming Zhu and Guodong Zhou 1331

Using Deep Linguistic Features for Finding Deceptive Opinion Spam
Qiongkai Xu and Hai Zhao 1341

Latent Community Discovery with Network Regularization for Core Actors Clustering
Guangxu Xun, Yujiu Yang, Liangwei Wang and Wenhuang Liu 1351

HYENA: Hierarchical Type Classification for Entity Names
Mohamed Amir Yosef, Sandro Bauer, Johannes Hoffart, Marc Spaniol and Gerhard Weikum 1361

Identifying Temporal Relations by Sentence and Document Optimizations
Katsumasa Yoshikawa, Masayuki Asahara and Ryu Iida 1371

Affect Detection from Semantic Interpretation of Drama Improvisation
Li Zhang and Ming Jiang 1381

Analyzing the Effect of Global Learning and Beam-Search on Transition-Based Dependency Parsing
Yue Zhang and Joakim Nivre 1391

Chinese Word Sense Disambiguation based on Context Expansion
Yang Zhizhuo and Huang Heyan 1401

Using qualia information to identify lexical semantic classes in an unsupervised clustering task

Lauren ROMEO[1] *Sara MENDES*[1,2] *Núria BEL*[1]

(1) Universitat Pompeu Fabra
Roc Boronat, 138, Barcelona, Spain

(2) Centro de Linguística da Universidade de Lisboa
Avenida Professor Gama Pinto, 2, Lisboa, Portugal

{lauren.romeo,sara.mendes,nuria.bel}@upf.edu

ABSTRACT

Acquiring lexical information is a complex problem, typically approached by relying on a number of contexts to contribute information for classification. One of the first issues to address in this domain is the determination of such contexts. The work presented here proposes the use of automatically obtained FORMAL role descriptors as features used to draw nouns from the same lexical semantic class together in an unsupervised clustering task. We have dealt with three lexical semantic classes (HUMAN, LOCATION and EVENT) in English. The results obtained show that it is possible to discriminate between elements from different lexical semantic classes using only FORMAL role information, hence validating our initial hypothesis. Also, iterating our method accurately accounts for fine-grained distinctions within lexical classes, namely distinctions involving ambiguous expressions. Moreover, a filtering and bootstrapping strategy employed in extracting FORMAL role descriptors proved to minimize effects of sparse data and noise in our task.

KEYWORDS : lexical semantic classes, qualia roles, unsupervised clustering, automatic extraction of lexical information

Proceedings of COLING 2012: Posters, pages 1029–1038,
COLING 2012, Mumbai, December 2012.

1 Introduction

Acquiring lexical information is a complex problem, typically approached by relying on a number of contexts to contribute information for classification, following the Distributional Hypothesis (Harris, 1954) and the idea of distributional similarity. In this domain it is crucial to determine which distributional information is significant to characterize lexical items. In line with Pustejovsky and Ježek (2008), we will make apparent how focusing on occurrences indicative of the FORMAL role of the Generative Lexicon (GL) theory (Pustejovsky, 1995) allows for identifying lexical semantic classes.

Lexical classes are linguistic generalizations regarding characteristics of meaning that correspond to sets of properties shared by groups of words. Bybee and Hopper (2001) and Bybee (2010) state that words are organized in lexical-semantic classes defined as emergent properties of words that recurrently occur in a set of particular contexts. Though many NLP tasks rely on rich lexica annotated with lexical semantic classes, reliable lexical resources including this type of lexical information are mostly manually developed, which is unsustainable, costly and time-consuming, and makes conceiving methods to automatically acquire such information crucial. An approach for acquiring lexical semantic classes proposes to classify words according to their occurrences in contexts where other lexical items belonging to a known class also occur. Yet, this approach has some limitations, such as data sparseness and noise (see Section 2), which underline the importance of developing new strategies to improve its effectiveness. Authors such as Pustejovsky and Ježek (2008) have shown how distributional analysis and theoretical modeling interact to account for rich variation in linguistic meaning. In line with this proposal, we evaluate the significance of specific co-occurrences whose selection was motivated by aspects of GL.

This work attempts to evaluate whether information provided by qualia roles, in specific the FORMAL role, is sufficient to discriminate lexical semantic classes of English nouns. With the experiments depicted in this paper, we aim to empirically demonstrate to which extent these features draw together nouns from the same lexical semantic class in an unsupervised clustering task. In this paper, Section 2 depicts background and motivation of this work. Section 3 presents relevant information on the GL and dot-objects. Section 4 describes the methodology to automatically obtain and cluster FORMAL role descriptors of nouns. Section 5 and 6, respectively, describe and discuss results. Section 7 reflects upon lexical classes and logical polysemy and is followed by final remarks.

2 Background and Motivation

Mainstream approaches to lexical semantic class acquisition classify words according to occurrences, i.e. they use the entire set of occurrences of a word to determine class membership. Yet, this approach has some limitations. Blind-theory distributional approaches have been shown to fail to account for the wide range of linguistic behavior displayed by words in language data (see Pustejovsky and Ježek (2008)), while authors such as Bel et al. (2010) reported problems caused by sparse data, or lack of evidence, and noise, or information obtained though not aimed at. Concerning sparse data in classification tasks, nouns that appear only once or twice in a corpus, and not in sought contexts, can render ineffective any classifier or clustering algorithm by not providing sufficient information for classification. We aim to soften effects of sparse data in the context of a clustering task by using a bootstrapping technique reliant on natural language inference properties (see Section 4.1). Noise, another pervasive issue in lexical semantic class acquisition, can be due to different factors: the occurrence of very general nominal expressions (e.g. “kind of”), which do not provide distinguishing lexical information; misleading corpus features; and the use of low-level tools (see Bel et al. (2012)). We assume noise resulting from errors generated by NLP tools to be typically characterized by unique occurrences and we employ a filtering strategy to overcome its possible effects (see Section 4.1). Concerning misleading corpus features, these are often caused by ambiguity of lexical items, resulting in nouns occurring in contexts not corresponding to their assumed lexical class. This presents challenging problems in classification tasks, as most authors do not distinguish among related senses of the same word, i.e. they either consider it as part of the class

or not (Hindle, 1990; Bullinaria, 2008; Bel et al., 2012). This is particularly problematic when words allow for multiple selection, i.e. when different senses of the same lexical item can be simultaneously selected for in one sentence (see (1)). Known as logical polysemy, this type of ambiguity has been shown to have well-defined properties (see Pustejovsky (1995) and Buitelaar (1998)) and has been consistently reported as a factor in lexical semantic acquisition tasks.

The newly constructed (LOCATION) *bank offers special conditions* (ORGANIZATION) *to new clients.* (1)

Approaches in this domain have usually tried to distinguish and isolate each word sense. We address this phenomenon differently, considering polysemous nouns as members of a given ambiguity class (within a wider lexical semantic class) and making apparent the relation between members of different classes by identifying shared properties beyond class limits. Given these considerations, we assume lexical units are complex objects that display rich variations of meaning in language use, placing ourselves within a theoretical framework that provides us the tools to account for this fact. Using the levels of representation and generative mechanisms in GL, we attempt to soften the effects of the aforementioned limitations in the automatic acquisition of lexical information.

3 Generative Lexicon theory

GL models the internal structure of lexical items in a computational perspective (Pustejovsky, 1995), proposing various levels of representation to semantically represent words, while allowing for the computation of meaning in context. Qualia Structure (QS) is one of these levels, consisting of 4 roles (FORMAL: what an object is; CONSTITUTIVE: what it is composed of; TELIC: its purpose; AGENTIVE: its origin), which model the predicative potential of lexical items. Here, we focus on the FORMAL role, defined as the role that distinguishes a lexical object within a larger domain (Pustejovsky, 1991).

QS also models phenomena such as polysemy of lexical items inherently complex in their meaning. These instances, *dot objects*, are the logical pairing of senses denoted by individual types in a complex type (Pustejovsky, 1995), which can pick up distinct aspects of the object, as well as properties of more than one class (Pustejovsky and Ježek, 2008), typically allowing for multiple selection (see (1)). Being able to represent lexical items as complex objects is useful in the context of our work as it provides a formal explanation for words belonging to more than one type, and essentially to more than one class.

Our experiment uses FORMAL role information as features for identifying lexical class membership. However, as there are no lexica available annotated with such information, we needed to obtain it automatically. Automatically extracting qualia roles with lexico-syntactic patterns has been receiving considerable attention for its success: Hearst (1992) identified lexico-syntactic patterns to acquire noun hyponyms, corresponding to the FORMAL role, whereas Cimiano and Wenderoth (2007) identified lexico-syntactic patterns to obtain information regarding semantic relations that correspond to each qualia role. As we needed information regarding the FORMAL role, not full lexical entries, in order for clusters to emerge, following Celli and Nissim (2009), we bypassed the representation of the entire QS, assuming semantic relations can be induced by matching lexico-syntactic patterns that convey a relation of interest.

4 Methodology

Given the unavailability of lexica annotated with FORMAL role information, and considering our basic goal of evaluating whether this information is enough to cluster together nouns of the same class, we extracted it from a corpus using lexico-syntactic patterns, following Cimiano and Wenderoth (2007), and then used it as features for a clustering task. In the experiment performed, we employed two steps: the extraction of FORMAL role descriptors from corpus data; and the clustering of this information. To obtain FORMAL role descriptors for our unsupervised clustering task, we used a part of the UkWaC Corpus (Baroni et al., 2009), consisting of 150 million tokens. We employed 60 seed nouns pertaining to three lexical semantic classes: HUMAN, LOCATION, and EVENT. The seed nouns were said to belong

to a class if they contained a sense in WordNet (Miller et al., 1990) corresponding to one of the three classes. Seed nouns were not contrasted with actual occurrences in the corpus.

4.1 Extraction of FORMAL role descriptors using lexico-syntactic patterns

Firstly, seed nouns were used in handcrafted lexico-syntactic patterns, adapted from Hearst (1992) patterns and the list proposed by Cimiano and Wenderoth (2007), to extract FORMAL role descriptors. These patterns were specified through regular expressions with PoS tags given after each token.

*x*_(or/and)_other_*y*
*x*_such_as_*y*
*x*_(is/are)_(a/an/the)_(kind(s)/type(s))_of_*y*
*x*_(is/are)_also_known_as_*y*

TABLE 1 – Clues on which patterns used to detect FORMAL role information in corpus data were built

The information obtained was stored in vectors representing co-occurrences with seed nouns in relevant contexts (patterns), where each element corresponds to occurrences of a particular seed noun (*x*) with a possible FORMAL role descriptor (*y*), following Katrenko and Adriaans (2008). Using the clues in Table 1, we obtained 185 FORMAL role descriptors for 55 of the 60 seed nouns in 353 occurrences. Considering this, and given the properties of the clustering algorithm used (see Section 4.2) a random value would be provided to nouns not sharing feature information with any other noun in our data set. To avoid random cluster assignations and provide more significant information to the system, we filtered out the features not shared between at least two seed nouns, without controlling which class the shared features belonged to, thus maintaining an unsupervised environment. Though we employed a large set of data, there were not enough shared FORMAL role descriptors for an important part of our data set, leading us to devise a strategy to increase the information available to the clustering algorithm.

a. A mammal is a [type of] animal.
b. A zebra is a [type of] mammal.
c. Therefore, a zebra is a [type of] animal. (2)

To increase the amount of FORMAL role descriptors, we employed a bootstrapping technique (Hearst, 1998) relying on monotonic patterns for natural language inference (Hoeksema, 1986; van Behthem, 1991; Valencia, 1991), illustrated in (2). This strategy is consistent with GL lexical inheritance structure (Pustejovsky, 1995; 2001), which assumes lexical items obtain their semantic representation by accessing a hierarchy of types and inheriting information according to their QS, meaning qualia elements are viewed as categories hierarchically organized. To illustrate how this applies in our case, the HUMAN noun *treasurer* obtained *officer* as a FORMAL role descriptor, whereas *officer* extracted *person* and *employee* as its own FORMAL role descriptors. Assuming this lexical organization, we consider FORMAL role descriptors extracted for *officer* to also be features of *treasurer*. Thus, we gathered additional information regarding the nouns to cluster, using originally obtained FORMAL role descriptors as "seed nouns" to extract more elements in an attempt to overcome biases due to sparse data (see Section 6), as well as to reinforce information already obtained. Employing the original patterns and original extractions as seeds, we obtained information that was added to the vectors. We conducted one iteration of the bootstrapping technique, going up one level of generalization to obtain the final distribution of information below. Newly obtained information was unified with previously extracted features, filtering out any additional noise attained. Table 2 presents the final distribution of this information.

Class	Elements	Occurrences
HUMAN	61 elements	841 occurrences
LOCATION	43 elements	225 occurrences
EVENT	36 elements	216 occurrences

TABLE 2 – Distribution of FORMAL role descriptors extracted (after filtering and bootstrapping) per class of seed noun

4.1.1 Error Analysis

Basing our clustering experiment on automatically extracted FORMAL role descriptors, the accuracy of the information obtained was a concern. To assess the accuracy of the information obtained, the FORMAL role descriptors extracted were revised manually. Extractions were considered erroneous if they provided information not in accordance with the class that the seed nouns pertained to. Table 3 presents the results of this analysis. Erroneous extractions were due to faults of the extraction mechanism (i.e. problems handling phenomena such as PP attachment), PoS tagging errors, lexical ambiguity or erroneous statements in text (Katrenko and Adriaans, 2008), as well as errors due to logical polysemy (see Section 6). Note that although errors were identified, they were not filtered for the clustering task, i.e. all information (erroneous or not) was included (on the impact of errors in results see Section 6).

Class	**% of accurate FORMAL role descriptors extracted**
HUMAN	87.60%
LOCATION	63.54%
EVENT	75.96%

TABLE 3 – Percentage (%) of accurate FORMAL role descriptors obtained per class

4.2 Clustering nouns using FORMAL role information

The second step of our experiment consisted in clustering nouns using the FORMAL role descriptors extracted. Given the nature of our data, we selected the sIB clustering algorithm (see Slonim et al. (2002) for a formal definition) for the manner it manages large data sets. This algorithm calculates similarity between two vectors using the *Jensen-Shannon* divergence, which measures similarity between probability distributions, rather than the Euclidean distance, which can bias the results when the number of attributes representing the factors is unequal (Davidson, 2002). This was our case as our feature spaces depend on the number of FORMAL role descriptors each seed noun occurred with in the corpus. To empirically demonstrate to which extent FORMAL role descriptors draw together nouns from the same class, we designed an experiment using the sIB algorithm in WEKA (Witten and Frank, 2005) to cluster seed nouns into lexical semantic classes, based only on the FORMAL role information obtained.

5 Results

As mentioned, our goal was to cluster together nouns from the same lexical semantic class using only FORMAL role descriptors. As the evaluation of unsupervised distributional clustering algorithms is usually done by comparing results to manually constructed resources (see Rumshsiky et al. (2007), among others), we employed our list of pre-classified seed-words to determine if nouns of the same class clustered together. Tables 4 and 5 present clustering results. The distribution of nouns across each cluster is given by the percentage of nouns pertaining to each lexical class included in it. The total number of seed nouns in each cluster is also given.

Cluster 0	Cluster 1	Cluster 2	**Class**
0.9285	0	**0.5714**	HUMAN
0.0769	0.3913	0.1429	LOCATION
0	**0.6087**	0.2857	EVENT
14	**23**	7	TOTAL NUMBER OF SEED NOUNS PER CLUSTER

TABLE 4 – Distribution of nouns in a 3-way clustering solution

We experimented with a 3-way and a 4-way clustering solution. In the first, the number of clusters was defined by the number of known classes, and resulted in the clustering of HUMAN nouns (Cluster 0). LOCATION and EVENT nouns grouped together in Cluster 1, the remaining cluster being composed of nouns from all classes with very few features available (less than three), i.e. insufficient information for classification. Considering this, we employed a 4-way solution to see whether LOCATION and EVENT nouns could be discriminated. This solution distinguished between the three classes (Cluster 0, 1 and 3 in Table 5) with a fourth cluster containing the "sparse data" nouns also affecting the 3-way solution.

Cluster 0	Cluster 1	Cluster 2	Cluster 3	**Class**
0	0	**0.5714**	**0.9286**	HUMAN
0	**0.9**	0.1429	0.0769	LOCATION
1	0.1	0.2857	0	EVENT
13	10	7	14	TOTAL NUMBER OF SEED NOUNS PER CLUSTER

TABLE 5 – Distribution of nouns in a 4-way clustering solution

The results show that even after filtering and bootstrapping the features extracted, sparse data still affected the results. However, nouns whose most salient common trait was the lack of sufficient information were consistently grouped together. Thus, the clustering is able to both discriminate between lexical semantic classes and act as a filter to detect those nouns for which there is not sufficient information using only FORMAL role information extracted from corpus data.

6 Discussion

As shown, the clustering algorithm discriminated between the three classes considered, using only the FORMAL role descriptors extracted from corpora data as features. Leaving aside the nouns for which there was not enough information available (12.7% of our data set), EVENT, HUMAN and LOCATION nouns were discriminated in the 4-way clustering solution (Clusters 0, 1 and 3 in Table 5, respectively). In this section we analyze misclassified nouns, to understand the reasons behind their misclassification, aiming to evaluate to which extent they correspond to recurring phenomena in language, which can possibly be accounted for by additional strategies.

Although their impact is not significant, noisy extractions (see Section 4.1.1) play a role in misclassification. In the 4-way clustering results, for instance, an EVENT noun is included in the cluster dominated by LOCATION nouns due to errors in extraction, specifically the incorrect identification as a FORMAL role descriptor of the noun in a PP modifying the head noun of an NP, which should be the one extracted. This type of noise is mostly generated by the use of low-level NLP tools. Overall, however, the existence of some noise in the data did not significantly affect the clustering, as demonstrated by the accuracy of the results presented in the previous section.

Concurrently, although general patterns can be identified in language use, one of the main characteristics of language data is its heterogeneity, which means that elements of a given lexical class do not necessarily share all their features or show perfectly matching linguistic behavior. Moreover, considering lexical items are complex objects with different semantic dimensions, they may share properties with elements of more than one lexical class. This type of phenomenon is behind some of the misclassifications in our data, such as the inclusion of *factory*, whose expected lexical class was LOCATION, in the HUMAN nouns cluster. This misclassification seems to be related to the fact that a part of HUMAN class members tended to obtain FORMAL role descriptors typical of HUMAN nouns, as well as of ORGANIZATION nouns, making apparent that nouns do not always occur in the sense considered in our pre-classified list of seed nouns.

7 Lexical classes and logical polysemy

As aforementioned, some HUMAN nouns in our list of seed nouns obtain FORMAL role descriptors typical of ORGANIZATION nouns. This is a type of polysemy that occurred in our data only with plural HUMAN nouns, alluding to the work of Copestake (1995) and Caudal (1998), according to whom some HUMAN nouns show a specific type of polysemy when heading definite plural NPs: the polysemy between the individual HUMAN sense and the collection of HUMANs sense, which in turn is polysemous between the HUMANGROUP and ORGANIZATION senses. In (3) we see how the definite plural NP *the doctors* can select for the two senses typically denoted by collective nouns, while having also the possibility to denote individual entities, which is not possible with collectives (see (4a)) that cannot occur in contexts that force a distinct individual entity reading.

a. *The doctors lay in the sun.* (several individual HUMAN entities)
b. *The doctors protested in front of the hospital.* (HUMANGROUP)
c. *The administration negotiated with the doctors.* (ORGANIZATION) (3)

a. # *The staff lay in the sun.* (several individual HUMAN entities)
b. *The employees lay in the sun.* (several individual HUMAN entities)
c. *The staff protested in front of the hospital.* (HUMANGROUP)
d. *The administration negotiated with the staff.* (ORGANIZATION) (4)

As both collectives and definite plural NPs denote collections, Caudal (1998) states that it is desirable to account for the polysemy of such items morpho-syntactically. This analysis is further strengthened by the observation that, unlike pairs such as *employee* and *staff*, for nouns like *doctor* there is no lexicalization for "group of doctors" in English, the same being true for collective nouns like *audience* or *committee*, whose individual members are not lexicalized. Given such lexical gaps, morpho-syntax is the strategy available. However, though logically polysemous, plural definite NPs like *the doctors* do not allow for multiple selection as is typical of complex types: once the individual HUMAN sense has been selected for there is no access to the HUMANGROUP·ORGANIZATION sense, as suggested by (5) (see Buitelaar (1998) and Rumshisky et al. (2007)).

The administration negotiated with the doctors, which later lay in the sun. (several individual HUMAN entities) (5)

Pustejovsky (1995:155) claims these patterns of linguistic behavior are due to the information in the QS. In the case of expressions like *the doctors*, the dot element denoting the individual HUMAN entity and the complex type HUMANGROUP·ORGANIZATION correspond to different qualia roles, as represented in (6). Hence, the different senses of the expression cannot be selected at the same time.

$$\left[\begin{array}{l}\textbf{the doctors}\\ \text{ARGSTR} = \left[\begin{array}{l}\text{ARG1} = \textbf{x: human}\\ \text{ARG2} = \textbf{y: humangroup} \cdot \textbf{organization}\end{array}\right]\\ \text{QUALIA} = \left[\begin{array}{l}\text{FORMAL} = \textbf{x}\\ \text{CONST} = \textbf{is\_part\_of(x,y)}\end{array}\right]\end{array}\right] \quad (6)$$

Going back to the case of *factory*, which was clustered with HUMAN nouns (see Section 6), we will see how the polysemy described above partially applies to this noun. Among the descriptors obtained for *factory* we found, alongside descriptors typical of LOCATION nouns, nouns such as *sector, organization* and *profession*, also extracted for HUMAN nouns showing the HUMANGROUP·ORGANIZATION logical polysemy, indicating that nouns like *factory* are also complex objects, as illustrated below by (7):

a. *The factory on the corner of Main Street is big and brown.* (LOCATION)
b. *The factory summoned a protest against the new government sanctions.* (ORGANIZATION)
c. *There was a protest organized* (ORGANIZATION) *by the factory that burned down* (LOCATION) *last week.* (7)

In our data, *factory* shared features both with definite plural NPs headed by HUMAN nouns like *teacher* and *employee* and LOCATION nouns such as *kitchen* and *resort*. The linguistic behavior of *factory* can, therefore, be assumed to reflect the logical polysemy of ORGANIZATION·LOCATION·HUMANGROUP dot types identified by Rumshisky et al. (2007), and represented as follows:

$$\left[\begin{array}{l}\textbf{factory}\\ \text{ARGSTR} = \left[\begin{array}{l}\text{ARG1} = \textbf{x: location}\\ \text{ARG2} = \textbf{y: organization}\\ \text{ARG3} = \textbf{z: human}\end{array}\right]\\ \text{QUALIA} = \left[\text{FORMAL} = \textbf{x} \cdot \textbf{y} \cdot \textbf{z}\right]\end{array}\right] \quad (8)$$

For our work, the most relevant aspect of the behavior displayed by nouns like *factory* is that it makes apparent how our strategy to extract FORMAL role descriptors reflects the ambiguity of nouns to be

clustered, which is often difficult to handle in NLP, particularly in classification tasks. The clustering solutions we obtained (see Section 5) grouped together HUMAN nouns, both those that display the ambiguity discussed in this section and those that do not, the same being true for LOCATION nouns. And yet, polysemous nouns display features that clearly point towards the existence of finer-grained distinctions, i.e. sub-classes within lexical semantic classes. This way, particularly given that these fine-grained distinctions are mirrored in FORMAL role descriptors, we assume it should also be possible to automatically recognize groups of nouns within the same ambiguity class, i.e. dot objects.

Hence, we expected the clustering algorithm to identify polysemous lexical items and distinguish them from other members of the same class. To validate this hypothesis we performed an additional iteration of the clustering using the same features and algorithm over previously identified clusters. The iteration was run individually over Clusters 1 and 3 (LOCATION and HUMAN noun clusters, respectively) from our 4-way clustering solution, as both clusters contained logically polysemous nouns. We obtained a 2-way clustering solution for each class, aiming to discriminate nouns strictly containing the LOCATION sense and those reflecting the polysemy described above for *factory*, on one hand, and nouns in the HUMAN·HUMANGROUP·ORGANIZATION ambiguity class from those strictly denoting human individuals on the other. Cluster 1 split into 2 clusters distinguishing between polysemous LOCATION nouns and those that are not, whereas for Cluster 3 the clustering algorithm arrived at a near perfect distinction of dot object nouns and non-ambiguous HUMAN nouns. The noun *factory* clustered with polysemous HUMAN nouns, once more confirming its semantic proximity with nouns of the HUMAN·HUMANGROUP·ORGANIZATION type. Hence, a second iteration of the same clustering algorithm over the same feature vectors was able to identify finer-grained distinctions within lexical classes, automatically recognizing groups of nouns in the same ambiguity class. In doing this, we validate our analysis regarding the role of logical polysemy and dot object types in the clustering solutions obtained, and further strengthen our original hypothesis.

Final remarks

In this paper, we proposed using automatically obtained FORMAL role descriptors as features to draw together nouns from the same lexical semantic class in an unsupervised clustering task. As there were no available lexica annotated with such information, we obtained it automatically and carried out clustering experiments. In line with the results, our initial hypothesis was supported: in an unsupervised clustering task using FORMAL role descriptors automatically extracted from corpora data as features, we showed it was possible to discriminate between elements of different lexical semantic classes. The filtering and bootstrapping strategy employed proved to minimize effects of sparse data and noise in our task. As shown in the 4-way clustering solution (see Table 5), the clustering exercise, as we designed it, also discriminated the nouns for which there was not sufficient information for a decision to be made on their membership to a cluster corresponding to one of the classes considered. Finally, we explained misclassifications through logical polysemy and showed how the method outlined in this paper allows for making finer-grained distinctions within lexical classes, recognizing lexical items in the same ambiguity class.

The results depicted in this paper demonstrate the validity of our hypothesis, while simultaneously showing that it is possible to incorporate the polysemous behavior of nouns in classification tasks (Hindle, 1990; Bullinaria, 2008) by using an approach that minimizes the effects of sparse data and noise (Bel et al., 2010; 2012). Considering these promising results, in future work we will address the possibility of extending our experiments to other qualia roles, as well as to other lexical semantic classes. At a more applied level, a further step consists in evaluating the feasibility of this approach to automatically extract lexical semantic classes in the automatic acquisition of rich language resources.

Acknowledgments

This work was funded by the EU 7FP project 248064 PANACEA and the UPF-IULA PhD grant program, with the support of DURSI, and by FCT post-doctoral fellowship SFRH/BPD/79900/2011.

References

Baroni, M., Bernardini, S., Ferraresi, A. and Zanchetta, E. (2009). The WaCky Wide Web: A Collection of Very Large Linguistically Processed Web-Crawled Corpora. *Language Resources and Evaluation*, *43*(3), 209-226.

Bel, N., Coll, M. and Resnik, G. (2010). Automatic detection of non-deverbal event nouns for quick lexicon production. In *Proceedings of the 23rd International Conference on Computational Linguistics, (COLING 2010)*, Beijing, China (pp. 46-52).

Bel, N., Romeo, L. and Padró, M. (2012). Automatic Lexical Semantic Classification of Nouns. In *Proceedings of the Language Resources and Evaluation Conference (LREC 2012)*, Istanbul, Turkey.

Buitelaar, P. (1998). *CoreLex: Systematic Polysemy and Underspecification.* Doctoral dissertation, Brandeis University.

Bullinaria, J.A. (2008). Semantic Categorization Using Simple Word Co-occurrence Statistics. In M. Baroni, S. Evert and A. Lenci (Eds.), *Proceedings of the ESSLLI Workshop on Distributional Lexical Semantics*, 1-8. Hamburg, Germany.

Bybee, J. L. and Hopper, P. (2001). *Frequency and the emergence of language structure*. Amsterdam: John Benjamins.

Bybee, J. (2010). *Language, usage and cognition*. Cambridge: Cambridge University Press.

Caudal, P. (1998). Using complex lexical types to model the polysemy of collective nouns within the Generative Lexicon. In *Proceedings of the Ninth International Workshop on Database and Expert Systems Applications*, Viena, Austria (pp.154-159).

Celli, F., Nissim, M., (2009) Automatic Identification of semantic relation in Italian complex nominals, *In Proceedings of the 8th International Conference on Computational Semantics (IWCS-8)*, Tilburg, Netherlands.

Cimiano, P. and Wenderoth, J. (2007). Automatic Acquisition of Ranked Qualia Structures from the Web. In *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics*, Prague, Czech Republic (pp.888-895).

Copestake, A. (1995). The representation of group denoting nouns in a lexical knowledge base. In P. Saint Dizier and E. Viegas (Eds.) *Computation Lexical Semantics* (pp. 207-230). Cambridge: Cambridge University Press.

Davidson, I. (2002). *Understanding K-means non-hierarchical clustering.* (Tech. Rep. 02-2). Albany: State University of New York.

Harris, Z. (1954). *Structural Linguistics.* Chicago: Chicago University Press.

Hearst, M. (1992). Automatic acquisition of hyponyms from large text data. In *Proceedings of the 14th International Conference on Computational Linguistics (COLING 92)*, Nantes, France (pp. 539-545).

Hearst, M. (1998). Automated Discovery of Word-Net relations. In C. Fellbaum (Ed.), *An Electronic Lexical Database and Some of Its Applications* (pp. 131-153). Cambridge: The MIT Press.

Hindle, D. (1990). Noun classification from predicate-argument structures. In *Proceedings of the 28th Annual Meeting of the Association for Computational Linguistics*, Pittsburgh. Pennsylvania (pp. 268-275).

Hoeksema J. (1986). Monotonicity Phenomena in Natural Language. *Linguistic Analysis, 16*, 25-40.

Katrenko, S. and Adriaans, P. (2008). *Qualia Structures and their Impact on the Concrete Noun Categorization Task*. In *Proceedings of the "Bridging the gap between semantic theory and computational simulations" workshop (ESSLLI 2008)*, Hamburg, Germany.

Miller, G.A., Beckwith, R., Fellbaum, C., Gross, D. and Miller, K.J. (1990). Introduction to WordNet: An online lexical database. *International Journal of Lexicography, 3*(4), 235-44.

Pustejovsky, J. (1991). The Generative Lexicon. *Computational Linguistics. 17*(4), 409–41.

Pustejovsky, J. (1995). *Generative Lexicon*. Cambridge: The MIT Press.

Pustejovsky, J. (2001). Type Construction and the Logic of Concepts. In P. Bouillon and F. Busa (Eds.), *The Language of Word Meaning* (pp. 91-123). Cambridge: Cambridge University Press.

Pustejovsky, J. and Ježek, E. (2008). Semantic coercion in language. beyond distributional analysis. *Italian Journal of Linguistics, 20*(1), 175-208.

Rumshisky, A., Grinberg, V. and Pustejovsky, J. (2007). Detecting Selectional Behavior of Complex Types in Text. In *4th International Workshop on Generative Lexicon*, Paris, France.

Slonim, N., Friedman, N. and Tishby, N. (2002). Unsupervised document classification using sequential information maximization. In *Proceedings of the 25th International ACM SIGIR Conference on Research and Development in Information Retrieval*, Tampere, Finland (pp.129-136).

Valencia, V. (1991). *Studies on Natural Logic and Categorial Grammar*. Doctoral dissertation, University of Amsterdam.

van Benthem, J. (1991). *Language in Action: Categories Lambdas and Dynamic Logic*. North Holland: Elsevier Science Publishers.

Witten, I.H. and Frank, E. (2005). *Data Mining: Practical machine learning tools and techniques* (2nd ed.). San Francisco: Morgan Kaufmann.