Date: Tuesday, October 18

Time: 11.00

Room: 02 New Wing, Faculty of Philosophy, Aristotle University

Multiword Expressions: from the English perspective towards a multilingual analysis

Professor Valia Kordoni, Department of English, Huboldt University Berlin

Abstract

In this talk, I mainly deal with acquisition of linguistic knowledge as a means of enhancing robustness of lexicalised grammars. The case study I focus on in the best part of this talk is Multiword Expressions (henceforward MWEs). Specifically, in the first part of the talk I am taking a closer look at the linguistic properties of MWEs, in particular, their lexical, syntactic, as well as semantic characteristics. The term Multiword Expressions has been used to describe expressions for which the syntactic or semantic properties of the whole expression cannot be derived from its parts (cf., Sag et al., 2002), including a large number of related but distinct phenomena, such as phrasal verbs (e.g., "come along"), nominal compounds (e.g., "frying pan"), institutionalised phrases (e.g., "breadand butter"), and many others. Jackendoff (1997) estimates the number of MWEs in a speaker's lexicon to be comparable to the number of single words.

However, due to their heterogeneous characteristics, MWEs present a tough challenge for both linguistic and computational work (cf., Sag et al., 2002). For instance, some MWEs are fixed, and do not present internal variation, such as "ad hoc", while others allow different degrees of internal variability and modification, such as "spill beans" ("spill several /musical /mountains of beans"). With the observations about the linguistic properties of MWEs at hand, I turn in the second part of the talk to methods for the semi-automated acquisition of these properties for robust grammar development. To this effect, I first investigate the hypothesis that MWEs can be detected by the distinct statistical properties of their component words, regardless of their type, comparing various statistical measures, a procedure which leads to extremely interesting conclusions. I then investigate the influence of the size and quality of different corpora, using the BNC and the Web search engines

Google and Yahoo. I conclude that, in terms of language usage, web generated corpora are fairly similar to more carefully built corpora, like the BNC, indicating that the lack of control and balance of these corpora are probably compensated by their size.

Then, I show a qualitative evaluation of the results of adding extracted MWEs to existing linguistic resources. To this effect, I first discuss two main approaches commonly employed for the treatment of MWEs: the words-with-spaces approach which models an MWE as a single lexical entry and it can adequately capture fixed MWEs like “by and large”, and compositional approaches which treat MWEs by general and compositional methods of linguistic analysis, being able to capture more syntactically flexible MWEs, like “rock boat”, which cannot be satisfactorily captured by a words-with-spaces approach, since this would require lexical entries to be added for all the possible variations of an MWE (e.g., “rock/rocks/rocking this/that/his…boat”). On this basis, I argue that the process of the automatic addition of extracted MWEs to existing linguistic resources improves qualitatively, if a more compositional approach to grammar/lexicon extension is adopted.

Short bio:

Valia Kordoni joined the faculty of the Department of English and American Studies of the Humboldt-Universität zu Berlin (Germany) in November 2012. Before that (2000-2012) she was a senior lecturer at the Department of Computational Linguistics and Phonetics of Saarland University, and a senior researcher at the Language Technology Lab of DFKI. Her research interests include Computational Semantics, Machine Translation, Syntax-Semantics-Pragmatics Interface, Context and Discourse Modelling, as well as Machine Learning for the automated acquisition of lexical knowledge, especially concerning multiword units and their impacts in Grammar Engineering.

She is the author of many refereed journal and conference publications and she has served as guest editor of special issues of journals like "Computational Linguistics" (Special Issue on Prepositions in Applications), "Linguistic Issues in Language Technology" (LiLT's Special Issue on the Interaction between Linguistics and Computational Linguistics), as well as the very recently published ACM Transactions on Speech and Language Processing (TSLP) - Special issue on multiword expressions: From theory to practice and use, part 1 and part 2.

She has co-organised conferences and workshops dedicated to research on MWEs, recently

including the ACL 2011 workshop on "Multiword Expressions: from Parsing and Generation to the real world" in Portland, Oregon, the NAACL HLT 2013 "9th Workshop on Multiword Expressions (MWE 2013)" in Atlanta, Georgia, and the EACL 2014 "10th Workshop on Multiword Expressions (MWE 2014)" in Gothenburg, Sweden, among others. She is the Local Chair of ACL 2016 - The 54th Annual Meeting of the Association for Computational Linguistics which will take place at the Humboldt-Universität zu Berlin from August 7 to August 12, 2016.

She has very recently taught a ACL 2013 tutorial on "Robust Automated Natural Language Processing with Multiword Expressions and Collocations". She is also the author of “Multiword Expressions - From Linguistic Analysis to Language Technology Applications” (to appear, Springer). She is a member of the ICT COST Action IC1207 Parsing and multi-word expressions. Towards linguistic precision and computational efficiency in natural language processing (PARSEME). Since 2011, she has intensely worked on establishing the ever growing MWE research community as part of the big SIGLEX community and has succeeded in doing so with the establishment in January 2013 of the SIGLEX-MWE Section. Since September 2013 she has been elected by the community as the SIGLEX Multiword Expression (MWE) Section Officer, and as board member of SIGLEX.

She is the local chair and the local sponsorship chair of ACL 2016.