Robust Semantic Analysis of Multiword Expressions with FrameNet

Miriam R. L. Petruck
International Computer Science Institute
1947 Center St, #600
Berkeley, CA 94704
miriamp@icsi.berkeley.edu

Valia Kordoni
Dept. of English Studies
Humboldt-Universität zu Berlin
Germany
kordonie@anglistik.hu-berlin.de

1 Introduction

This tutorial will give participants a solid understanding of the linguistic features of multiword expressions (MWEs), focusing on the semantics of such expressions and their importance for natural language processing and language technology, with particular attention to the way that FrameNet (framenet.icsi.berkeley.edu) handles this wide spread phenomenon. Our target audience includes researchers and practitioners of language technology, not necessarily experts in MWEs or knowledgeable about FrameNet, who are interested in NLP tasks that involve or could benefit from considering MWEs as a pervasive phenomenon in human language and communication.

2 Topic Overview

NLP research has been interested in automatic processing of multiword expressions, with reports on and tasks relating to such efforts presented at workshops and conferences for at least ten years (e.g. ACL 2003, LREC 2008, COLING 2010, EACL 2014). Overcoming the challenge of automatically processing MWEs remains elusive in part because of the difficulty in recognizing, acquiring, and interpreting such forms.

Indeed the phenomenon manifests in a range of linguistic forms (as Sag et al. (2001), among many others, have documented), including: noun + noun compounds (e.g. *fish knife*, *health hazard* etc.); adjective + noun compounds (e.g. *political agenda*, *national interest*, etc.); particle verbs (*shut up*, *take out*, etc.); prepositional verbs (e.g. *look into*, *talk into*, etc.); VP idioms, such as *kick the bucket*, and *pull someone's leg*, along with less obviously idiomatic forms like *answer the door*, *mention someone's name*, etc.; expressions that have their own mini-grammars, such as names with honorifics and terms of address (e.g. *Rabbi Lord Jonathan Sacks*), kinship terms (e.g. *second cousin once removed*), and time expressions (e.g. *January 9, 2015*); support verb constructions (e.g. verbs: *take a bath*, *make a promise*, etc; and prepositions: *in doubt, under review*, etc.). Linguists address issues of polysemy, compositionality, idiomaticity, and continuity for each type included here.

While native speakers use these forms with ease, the treatment and interpretation of MWEs in computational systems requires considerable effort due to the very issues that concern linguists.

3 Content Overview

The first part of the tutorial offers a general introduction to the phenomenon of MWEs, with a discussion of the various types of MWEs (e.g. compounds, fixed phrases, idioms, etc.), their syntactic and semantic characteristics, along with a presentation of representational issues. Focusing mostly on English MWEs, the discussion provides data from languages other than English, also to show the manifestation of the phenomenon in a wide range of languages.

Part II begins with an overview of FrameNet (Ruppenhofer et al. 2010), a knowledge base that includes unique information about the mapping of meaning to form in contemporary English via the theory of Frame Semantics (Fillmore and Baker 2010). Continuing with FrameNet's treatment of a range of MWE-types, this section highlights support constructions (e.g. *say a prayer*, *make a decision*, *at risk*, *on fire*, etc.) and transparent nouns (e.g. *school of fish*, *type of drink*, *bottle of wine*, etc.), also noting the discrepancy between syntactic heads and semantic heads of such forms. Part II

Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing, pages 12–13,
Lisboa, Portugal, 17-21 September 2015.

concludes by demonstrating the advantages of using FrameNet information about MWEs.

Part III of the tutorial briefly offers a survey of computational approaches for MWE recognition, basically focusing on modeling semantic variability. In this part we also review disambiguation of MWEs in context (e.g., *bus stop*, as in *Does the bus stop here*? vs. *The bus stop is here*) and methods for the automatic detection of the degree of semantic compositionality of MWEs and their interpretation. This part serves as a brief introduction to Part IV of the tutorial, which addresses some of the challenges of using FrameNet data in NLP (semantic) tasks, especially that of coverage.

4 Tutorial Outline

Part I: General Overview of MWEs
- a. Introduction
- b. Types of MWEs
- c. Syntactic and Semantic Characteristics of MWEs
- d. Representational Issues in MWEs

Part II: MWEs in FrameNet
- a. Overview of FrameNet
- b. FrameNet's treatment of MWEs
- c. Navigating Lexicon and Grammar
- d. Exploiting FrameNet Information on MWEs

Part III: Computational Processing of MWEs
- a. Recognizing elements of MWEs: type identification
- b. Recognizing how MWE elements combine: syntactic and semantic variability
- c. Disambiguation of MWEs
- d. Compositionality and Interpretation of MWEs

Part IV: Robust Semantic Analysis of Multiword Expressions in FrameNet: main challenges

Acknowledgments

Much of the linguistic content of the tutorial, and specifically the way that FrameNet analyzes multiword expressions derives significantly from Charles J. Fillmore's insightful work on the topic (summarized in Fillmore 2006).

References

Charles J. Fillmore. 2006. Multiword Expressions: An Extremist Approach. Unpublished Power Point Presentation, International Computer Science Institute and University of California, Berkeley, CA

Charles J. Fillmore and Collin Baker. 2010. A Frames Approach to Semantic Analysis. In Bernd Heine and Heiko Narrog (Eds.). *The Oxford Handbook of Linguistic Analysis*, Oxford University Press, Oxford, UK, pages 313-340.

Valia Kordoni. To appear. *Multiword Expressions From Linguistic Analysis to Language Technology Applications,* Springer.

Carlos Ramisch, Aline Villavicencio, and Valia Kordoni. 2013. *Special Issue on Multiword Expressions*. ACM TSLP.

Josef Ruppenhofer, Michael Ellsworth, Miriam R. L. Petruck, Christopher R. Johnson, and Jan Scheffczyk. 2010. *FrameNet II: Extended Theory and Practice*. Web Publication (framenet.icsi.berkeley.edu/book)

Ivan A. Sag, Timothy Baldwin, Francis Bond, Ann Copestake, and Dan Flickinger. 2001. Multiword expressions: A pain in the neck for NLP. In *Proceedings of the 3rd International Conference on Intelligent Text Processing and Computational Linguistics* (CICLing-2002), Berlin: Springer.

Aline Villavicencio, Valia Kordoni, Yi Zhang, Marco Idiart, and Carlos Ramisch. 2007. Validation and evaluation of automatically acquired multiword expressions for grammar engineering. In *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL)*, pages 1034–1043, Prague, Czech Republic, June. ACL.

5 Instructor Bios

Miriam R. L. Petruck received her PhD in Linguistics from the University of California, Berkeley. A key member of the team developing FrameNet almost since the project's founding, her research interests include semantics, knowledge base development, grammar and lexis, lexical semantics, Frame Semantics and Construction Grammar.

Valia Kordoni received her PhD in Computational Linguistics from the University of Essex, UK. She joined the Department of English Studies, Humboldt University Berlin in 2012, where she is Research Professor of Linguistics. Her main research interests are in deep linguistic processing, semantic analysis, and multiword expressions.