The Hebrew FrameNet Project

Avi Hayoun and **Michael Elhadad**
Dept. of Computer Science
Ben-Gurion University
Beer Sheva, Israel
{hayounav,elhadad}@cs.bgu.ac.il

Abstract

We present the Hebrew FrameNet project, describe the development and annotation processes and enumerate the challenges we faced along the way.

1 Introduction

Based on the linguistic theory of *Frame Semantics* proposed by Fillmore (Fillmore, 1982), the FrameNet project (Fillmore and Baker, 2010; Ruppenhofer et al., 2010) is a human-annotated linguistic resource with rich semantic content. FrameNet defines a collection of *semantic frames*, each containing a set of frame evoking predicates called *Lexical Units*, definitions of roles of participants in the event described by the frame and additional information describing relationships between roles and frames.

FrameNet databases in languages other than English have been created in multiple languages, including German, Spanish, Japanese, Swedish. Inspired by Swedish FrameNet++ (Friberg Heppin and Voionmaa, 2012) and the ideas put forth by Petruck (Petruck, 2005; Petruck, 2009), we have started the development of a Hebrew FrameNet, a semi-automatic translation of the English FrameNet.

As part of the development and adaptation process, we were confronted with issues specific to the Hebrew language, which we discuss. We first present the process we have adopted to develop the Hebrew FrameNet resource, the supporting tools we developed and provide information on the linguistic issues.

2 Development Process

We started with the English FrameNet data (version 1.5.1) as a basis on which to build. We currently use the same semantic frames included in the original FrameNet data, including definitions, participant roles inter-role and inter-frame relationships.

To accelerate the process of adding Hebrew lexical units (LUs) to frames, our system suggests translations of the English LUs. However, annotators are free to add new LUs on their own. We collected translation pairs for most of lexical units occurring in FrameNet from online lexicographic resources and introduced lookup procedures as part of the online annotation tool we developed for the project.

The next step in the annotation process is the selection of exemplar sentences for each LU for each frame. To assist in this step, we prepared a large corpus of about 2M sentences collected from a variety of Modern Hebrew sources (newspaper, blogs, wikipedia) and pre-processed these sentences with full morphological analysis, and automatic syntactic parsing (both constituent and dependency parsing). We gave access to this annotated corpus from the FrameNet annotation tool through a full text search where specific lexical items can be searched and matched irrespective of their morphological inflection. In addition, annotators can refine the search query by adding specification of part of speech, morphological features (number, gender, person etc), and syntactic context (items related to other items, *e.g.*, a word appearing as the subject of a verb). We apply the syntactic diversification algorithm of (Borin et al., 2012) to the search result, so that the top N sentences presented to the annotator exhibit a wide range of syntactic constructs. Annotators can quickly create a range of syntactic examples for a single semantic concept.

The final stage in the annotation process of a single Frame consists of annotating occurrences of frame elements with roles within the selected example sentences. Spans of text are selected by annotators to fill the various roles encoded in the semantic frame definition.

2.1 Project Status

As of May 2015, the Hebrew FrameNet project contains $3,006$ LUs across 167 frames, with an average of 18 LUs per frame. Additionally, there are 423 annotated exemplar sentences pending final review across 66 LUs, with an average of 6.41 sentences per LU. Before starting a more intense annotation campaign, we are now reviewing the linguistic issues faced during the initial annotation trial and assessing the potential to speed up annotation with semi-supervised expansion.

3 Hebrew-Specific Issues

We identified the following issues specific to Hebrew in our initial annotation effort.

3.1 Multi-word Lexical Units

While the English FrameNet project contains several multi-word LUs (MWLUs), such as *give up.v* and *turn in.v*, they are annotated as contiguous units. In Hebrew, we found so many morphological and syntactic variants of MWLUs that we decided to enable annotation of discontinous units. For example, מזל רע is a contiguous unit, but יצר קשר is the LU in both of the following sentences illustrating the wide range of discontinuous constructions we observed:

החייל יצר קשר עם מפקדו.
הוא יצר איתי קשר אתמול.

We solved this issue by adding a "contiguous" flag to the multi-word LUs in our dataset, which indicates their expected behavior.

3.2 Role-Bearing Phrases Embedded in Morphology

In some cases, a role-bearing phrase is embedded in another word in the text, due to the rich morphology which exists in Hebrew. For example, consider the following exemplar sentence from the **Abandonment** frame: זנח את אשתו. If the possessive ו in the word אשתו were a separate token, the correct annotation would be

[Agent ו] [Theme אשת] זנח את

Since this is not the case, we needed a way to encode the fact that the Agent is embedded in another phrase. We followed Petruck's recommendation and borrowing from the Spanish FrameNet project, we annotate such roles as "externally instantiated", meaning no single phrase or token in the sentence can represent the role.

4 Semi-Supervised Dataset Expansion

Fürstenau and Lapata (Fürstenau and Lapata, 2012) presented a semi-supervised method for automatically expanding the FrameNet corpus, using structural alignment of dependency-parse trees. They reported a success rate of about 33% of the projected annotations as completely correct. We implemented this algorithm in the Hebrew FrameNet tool to assist annotators; gold annotations are used to seed automatic annotations of candidate sentences from Wikipedia. We computed a Word2vec word embedding on the Hebrew corpus taking into account syntactic relations in the dependency trees as opposed to n-grams following (Levy and Goldberg, 2014). Candidate sentences are extracted using a lexical similarity measure, based on the the computed word embeddings.

The projected annotations are manually reviewed before being accepted into the Hebrew FrameNet corpus. We will report on the accuracy of this process.

References

Lars Borin, Markus Forsberg, Karin Friberg Heppin, Richard Johansson, and Annika Kjellandsson. 2012. Search result diversification methods to assist lexicographers. In *Proceedings of the Sixth Linguistic Annotation Workshop*, pages 113–117. Association for Computational Linguistics.

Charles J Fillmore and Collin Baker. 2010. A frames approach to semantic analysis. *The Oxford handbook of linguistic analysis*, pages 313–339.

Charles J. Fillmore, 1982. *Frame semantics*, pages 111–137. Hanshin Publishing Co., Seoul, South Korea.

Karin Friberg Heppin and Kaarlo Voionmaa. 2012. Practical aspects of transferring the english berkeley framenet to other languages. In *SLTC 2012 The Fourth Swedish Language Technology Conference Lund, October 24-26, 2012 Proceedings of the Conference*, pages 28–29.

Hagen Fürstenau and Mirella Lapata. 2012. Semi-supervised semantic role labeling via structural alignment. *Comput. Linguist.*, 38(1):135–171, March.

Omer Levy and Yoav Goldberg. 2014. Dependencybased word embeddings. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics*, volume 2, pages 302–308.

Miriam R.L. Petruck. 2005. Towards hebrew framenet. *Kernerman Dictionary News*, 13:12.

Miriam Petruck, 2009. *Typological Considerations in constructing a Hebrew FrameNet*, pages 183–208. Mouton de Gruyter, Berlin.

Josef Ruppenhofer, Michael Ellsworth, Miriam R. L. Petruck, Christopher R. Johnson, and Jan Scheffczyk. 2010. FrameNet II: Extended Theory and Practice. Technical report, International Computer Science Institute in Berkeley, 09.