

# Building a Hebrew FrameNet Lexical Resource from Parallel Movie Subtitles

Ben Eyal

October 10, 2018

# Outline

- 1 Introduction
- 2 Motivation
- 3 Previous Work
- 4 Starting Points
- 5 Research Question
- 6 Methods
- 7 Results and Evaluation
- 8 Conclusion

# Outline

- 1 Introduction
- 2 Motivation
- 3 Previous Work
- 4 Starting Points
- 5 Research Question
- 6 Methods
- 7 Results and Evaluation
- 8 Conclusion

## The “What”

- What is semantic role labeling (SRL)?
- Annotating sentences with labels that answer questions such as “Who did what to whom, when, and where?”
- Answers to these questions are called “roles”
- Also known as “shallow semantic parsing”, since unlike part-of-speech tags, SRL deals with the meaning of the text

## The “Why”

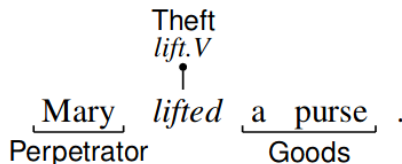
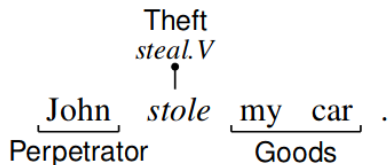
From the explanation above, one can understand the uses of SRL in more difficult tasks such as question answering and information extraction.

## The “How”

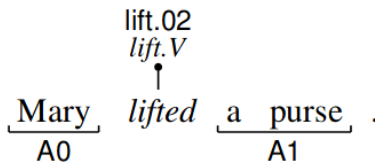
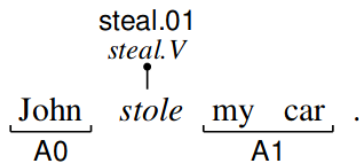
- Two major SRL annotation schemes are FrameNet and PropBank
- Work surrounding these schemes has contributed fully-annotated corpora in various languages, three CoNLL shared tasks, and many automatic systems
- In this work we set out to contribute a new Hebrew corpus and lexical resources to further the advancement of the Hebrew FrameNet Project

# Prerequisites

Before diving in, we need to understand the difference between FrameNet-style annotations (a) and PropBank-style annotations (b):



(a)



(b)

# Outline

- 1 Introduction
- 2 Motivation**
- 3 Previous Work
- 4 Starting Points
- 5 Research Question
- 6 Methods
- 7 Results and Evaluation
- 8 Conclusion



## Example

Consider the following sentence:

“Sally fried an egg in butter.”

Of course, we can ask syntactical questions such as “What is the verb?”, “What is the head of the sentence?”, “How many nouns are there?”, etc. These questions are boring. SRL and FrameNet tells us more. It tells us the lexical unit (LU) “fried” *evokes* the *frame* “APPLY\_HEAT”.

## Example

This frame has five “core frame elements” (FEs):

- Container
- Cook
- Food
- Heating instrument
- Temperature setting

So now we can answer questions like “Who’s cooking?”, “How does Sally like her egg?”, and of course, the age-old question: “What’s cookin’?”. We can also know things that are not explicitly in the sentence, e.g., that Sally is hungry, that she is not lactose-intolerant, and that she likes eggs.

## Multilingual SRL

So we understand that SRL in general, and FrameNet annotations in particular, are useful. The problem is that FrameNet is in English, and we would like to do SRL on a variety of languages for uses such as question answering, etc.

We would also like to explore whether FrameNet's frames are stable across languages.

For example:

סאלי שיגנה ביצה בחמאה

We see that the frame `APPLY_HEAT` still applies: שיגנה evokes the frame, סאלי is the cook, etc.

## Working in Hebrew

Unlike English and French, Hebrew barely has any textual resources, which means we need an English-Hebrew parallel corpus.

# Outline

- 1 Introduction
- 2 Motivation
- 3 Previous Work**
- 4 Starting Points
- 5 Research Question
- 6 Methods
- 7 Results and Evaluation
- 8 Conclusion

## Semantic Role Labeling

- 1 **2002** Gildea and Jurafsky treated the problem as a sequence-tagging problem rather than a parsing problem, meaning each word is considered separately and has its own probability distribution over tags. For example, in the sentence

[*B-Judge* She] [*O* **blames**] [*B-Evaluee* the] [*I-Evaluee* Government]  
[*B-Reason* for] [*I-Reason* failing] [*I-Reason* to] [*I-Reason* do]  
[*I-Reason* enough] [*I-Reason* to] [*I-Reason* help].

we are given that the target word is **blames**, and it evokes some frame. Now, given the frame, the most probable tag for “She” is “Judge”, and it is the beginning of the role, so it’s tagged “B” for **beginning** (the other tags are “I” for **inside**, and “O” for **outside**).

## Semantic Role Labeling

- ② **2004** Xue and Palmer proposed other features to the above method, and split the problem in three: pruning, argument identification, and argument classification.
- ③ **2005** Punyakanok et al. added a fourth step to Xue and Palmer which is inference, where a list of eight constraints need to be satisfied, using an ILP.

## Semantic Role Labeling

- ④ **2015** Täckström et al. breaks a 10-year silence in the field by reformulating the SRL task as a dynamic program. It is four times faster than off-the-shelf ILP solvers.
- ⑤ **2015** Zhou and Xu set the stage for “the next big thing” by being the first to introduce deep neural networks to the SRL task, and returning the task to its roots as a sequence-tagging problem. The network is eight BiLSTMs deep, with a CRF as the top-most layer, outputting the tag distribution.
- ⑥ **2017** He et al. improves Zhou and Xu by adding “highways” between the layers, and use a softmax layer instead of a CRF.



## Hebrew SRL

Avi Hayoun started working on a Hebrew FrameNet (HebFN), building an annotation tool, frame repository, and a collection of lexical units. As of October 2015, HebFN contains ~3,000 LUs across 167 frames.

## Hebrew SRL

### The Abandonment frame

#### ^ Abandonment

An **Agent** leaves behind a **Theme** effectively rendering it no longer within their control or of the normal security as one's property.











**Carolyn** **abandoned** **her car** and jumped on a red double decker bus.








Perhaps **he** **left** **the key** in the ignition






















**Abandonment** **of a child** is considered to be a serious crime in many jurisdictions.

#### Frame elements ▾

#### Inter-frame relations ▾

English LUs	
Click to see translations	
<input type="checkbox"/> abandon.v	 
<input type="checkbox"/> abandonment.n	 
<input type="checkbox"/> forget.v	 
<input checked="" type="checkbox"/> leave.v	 
<input type="checkbox"/> abandoned.a	 

Translations of leave.v	
Associate with frame or show example sentences	
	השאיר.v
	התר.ח
	אשור.ח
	רשות.ח
	העדרות.ח
	חפשה.ח
	פרדה.ח

Hebrew LUs	
<input type="button" value="Add a custom LU"/>	
   12	הופקר.ח
   26	הפקיר.ו
   17	זנח.ו
   18	נטוש.א
   26	נטישה.ח
   20	נטש.ו
   17	עזב.ו

# Outline

- 1 Introduction
- 2 Motivation
- 3 Previous Work
- 4 Starting Points**
- 5 Research Question
- 6 Methods
- 7 Results and Evaluation
- 8 Conclusion

# Starting Points

## The Plan

In order to develop a Hebrew FrameNet, we need an English-Hebrew parallel corpus. Once we have the corpus, we project the English FrameNet annotations to its aligned Hebrew sentence.

## Prerequisites

- Collect pairs of aligned sentences
- Compute lexical alignment across the corpus
- Perform syntactic analysis with a similar annotation scheme across English and Hebrew sentences
- Perform SRL analysis of the English side
- Project annotations from English to Hebrew

## Aligned Corpus Collection: OpenSubtitles

- 2.6 billion sentences
- 60 different languages
- 152,939 movies and TV episodes from the OpenSubtitles database
- 1,689 bitexts
- 23,727,452 aligned sentence pairs English-Hebrew in varying degrees of quality

## Hebrew Syntactic Analysis with YAP

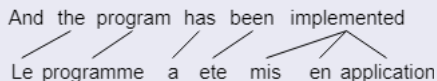
Just like with English, we want to find information such as part-of-speech tags and dependency parse trees, but unlike English, there are two steps we must perform first: segmentation and morphological disambiguation. Let us see an example of why these are needed:

Segmentation	Meaning
בצלם	name of human rights association (Betselem)
בצלם	while taking a picture
בצלם	their onion
בצלם	under their shadow
בצלם	in a photographer
בצלם	in the photographer

# Starting Points

## Word Alignment with `fast_align`

- Out of three word aligners, we chose `fast_align` to produce word alignments.
- `fast_align`'s model is a reparameterization of the IBM Model 2 translation model. What the IBM translation models do is use expectation maximization (EM) to maximize the probability of a target word given a source word and an alignment. For example, this is a good alignment:



- 24 hours to train and align 383M tokens (194M English, 189M Hebrew).
- 181M English-Hebrew word pairs were identified, with one English word mapping, on average, to 1.25 Hebrew words.

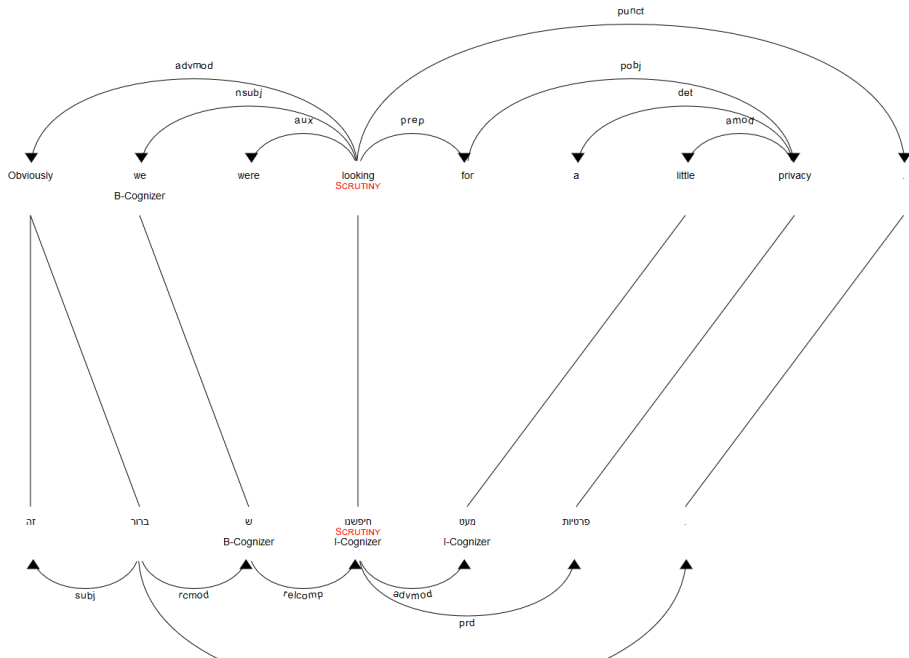
## English SRL Analysis with SEMAFOR

- Based on manual inspection, SEMAFOR triumphed over other automatic SRL systems
- SEMAFOR works by splitting the task in three
  - 1 **Target Identification.** Heuristically find frame-evoking words and/or phrases (“targets”) in sentences.
  - 2 **Frame Identification.** Find a probability distribution over frames for each target, and label the target with the most probable frame.
  - 3 **Argument Identification.** Find the spans in each sentence that fit the selected frame, and ensure that arguments do not overlap.



## Projection Strategy

**Direct Semantic Transfer (DST)** Transfer semantic relationships between a source sentence and a target sentence iff the constituents in the relationship are aligned. For example:



# Outline

- 1 Introduction
- 2 Motivation
- 3 Previous Work
- 4 Starting Points
- 5 Research Question**
- 6 Methods
- 7 Results and Evaluation
- 8 Conclusion

## Questions

Our work addressed the following questions:

- Given an aligned corpus of English-Hebrew sentences and English SRL annotations, can we obtain a reliable projection mechanism to Hebrew sentences?

## Questions

- Can we design a mechanism to automatically identify reliable pairs of sentences given the set of annotations available on the English and on the Hebrew sides?

## Questions

- We develop a manual control dataset to measure the performance of various unsupervised projection methods. How can we decide what to annotate manually and how much data is necessary to qualify the dataset produced automatically?

## Questions

Our approach is empirical – we collect datasets, measure empirical distributions and report on the complexity of the task through various statistical metrics.

# Outline

- 1 Introduction
- 2 Motivation
- 3 Previous Work
- 4 Starting Points
- 5 Research Question
- 6 Methods**
- 7 Results and Evaluation
- 8 Conclusion



## Data Collection and Filtering

- The OpenSubtitles dataset provides 23.7 million English-Hebrew sentence pairs. Due to the nature of the dataset, there are pairs (sometimes entire files) that are very noisy.
- Common sources of noise were OCR artefacts (s|\\|e| instead of “smell”), music cues for the hearing-impaired, and “group banners”, e.g., “These subtitles brought to you by group ...”
- Automatic and manual denoising leave us 23 million sentence pairs.

## Hebrew Preprocessing

As we have seen with the **בצלם** example, Hebrew preprocessing requires:

- ➊ Morphological analysis segments the words and generates their (possibly many) meanings
- ➋ Morphological disambiguation picks the most likely candidates from the previous step
- ➌ Dependency parsing gives a dependency parse tree for each disambiguated sentence

## English Preprocessing

English preprocessing is a breeze by comparison. SEMAFOR (the SRL system we use) uses MaltParser for its preprocessing, which includes part-of-speech tagging and dependency parsing, and that is all that is needed.

## Semantic Role Labeling

Semantic role labeling is done using SEMAFOR and FrameNet v1.5. SRL systems working with v1.7 exist, but produce results inferior to SEMAFOR's.

## Filtering

- ① We have 23 million sentence pairs. Statistically, most of them are not that good. Bad, even.
- ② To filter the trash from the treasure, we built a tool that visualizes the alignment, dependency trees for both languages, and SRL for English.
- ③ Each pair is tagged with one of the following:
  - Error in sentence alignment
  - Error in word alignment
  - Poor translation
  - Poor syntactic parsing
  - Poor frame parsing
  - Good

## Filtering

- ④ We manually tag 124 pairs:

Quality	Amount
Error in sentence alignment	11
Error in word alignment	43
Poor translation	11
Poor syntactic parsing	0
Poor frame parsing	3
Good	56

## Filtering

- ⑤ We build a classifier around the manual annotations. Features used:
  - English/Hebrew sentence length and ratio between them
  - Number of frames
  - Number of 1-1/1-n alignments
  - English/Hebrew sentence parse tree depth
- ⑥ The classifier outputs the probability of a sentence being marked as “Good”.
- ⑦ We automatically annotate as “Good” alignments which the classifier gave a probability of more than 80%.
- ⑧ We rejoice as it seems the classifier did its job in tagging as “Good” sufficiently long and well-aligned sentences.



## Projection

- ① Find the head word of an annotated span
- ② Get the Hebrew word aligned to that head word
- ③ Find its subtree
- ④ Annotate according to the English span annotation



# Outline

- 1 Introduction
- 2 Motivation
- 3 Previous Work
- 4 Starting Points
- 5 Research Question
- 6 Methods
- 7 Results and Evaluation**
- 8 Conclusion

## Results

- **Word Pairs.** 115 million English-Hebrew word pairs which give us an empirical English-Hebrew dictionary
- **Exemplar Sentences.** 11,205 fully-annotated candidates to serve as exemplar sentences of their respective frame
- **Frames.** Among the 11,205 sentences there are 678 frames, about one third of FrameNet v1.5 frame inventory
- **Lexical Units.** Assuming one word per LU, we manage to add 5,258 LUs to Hebrew FrameNet.

# Outline

- 1 Introduction
- 2 Motivation
- 3 Previous Work
- 4 Starting Points
- 5 Research Question
- 6 Methods
- 7 Results and Evaluation
- 8 Conclusion**

## Pipeline

- Hebrew + noisy dataset = bad time
- Improvements to different parts of the pipeline will benefit the entire pipeline
- The sheer amount of data, albeit noisy, can add to the Hebrew FrameNet project:
  - 678 frames (out of ~1,000 frames in the FrameNet inventory)
  - ~11,000 fully annotated exemplar sentences
  - ~5,000 lexical units (targets)
- These figures enable us to train a Hebrew SRL system

## Quality

- We still need to assess the quality of the automatically produced data that we have pushed into Hebrew FrameNet
- Our experiment with annotating a seed of only about 100 sentences to determine sentence quality is encouraging in indicating practical methodology to identify different aspects of the data produced (LUs and Frame Annotations)

# Questions?

# Thank You!