# Hebrew Morphology-
# Models and Methods used in Search English/

Efraim Margalit

## Abstract

This paper review various Hebrew morphologic models embedded within search engines. It reviews the history of search engines, starting from using key-words, going through contemporary search engines. As part of reviewing the references, we'll try to evaluate the uniqueness of Hebrew, as of morphologic richness and the technological challenges it imposes.

It is well known, that Hebrew, as other Semitic languages, stands at the top of the morphologic complexity pyramid. This complexity originates from various aspects, mainly from the high number of forms, ranging from 70 to 100 million formal and proper forms.

There is a need for morphologic analysis models are available. The model that can be used for English is called "Stemming", and it deals with prefixes and suffixes of a vocabulary entries. For example, the form "misunderstanding" includes both a prefix and a suffix to a vocabulary entry. In English, prefix and suffixes are regular, so it is relatively simple to implement a stemming algorithm. For example, one can find an open source pseudo code for a stemming algorithm that includes as low as dozens of code lines.

Arabic, which is also Semitic language, resembles Hebrew by its reach complexity of formal linguistic form. Therefore, Arabic is generally treated as Hebrew regarding morphological analysis.

The huge number of forms results from the fact that though there is a small number of roots, namely about five thousands, each one of the roots can evolve to a large number of forms. For example, there are verbs that can be inflected into more than twenty thousand regular forms.

Yet another challenge in this field is the lexical ambiguity. Only forty to forty five percents of the words in the Hebrew language are unambiguous, while about one third of the words have more than two meanings. This means that any string within a Hebrew text has, on everage, more than one meaning.

Until here we received regular morphologic analysis of words in a language. However, when doing so, one cannot ignore considering the contemporary Internet revolution and its effects on our subject of interest. As part of this revolution, any person can publish on the Internet anything he\she desires. As part of this publications pluralism, we are witnessing many changes in writing styles. No more an official and well defined process of publication that includes a proofing process both for grammar and style, but rather new articles written in "Contemporary Hebrew".

The practical results of this trend are that are contemporary morphologic search engines must deal with various Hebrew slang style along with Hebrew texts that include linguistic errors.

This paper presents five different morphologic analysis models in details:

The first model is a statistic model that presents a method that combines three levels of morphological analysis and select the best analysis using statistic models. This model was developed by Segal and it presents a very high identification level when used with formal text.

The second model, developed by Ornan, presents a new Hebrew meanings vocabulary. This model is based on that this vocabulary includes both words entries as well semantic characteristics. In this manner, we can also implement the formal check of the semantic meaning of each sentence and clear most of the lexical ambiguity. For example, the words "Terminal refreshing" may have several formal linguistic analysis results. However, the correct analysis of a computer terminal is performed using the special vocabulary, that includes the attributes "Technology" for both entries.

The third model presents a Heuristic model, developed by Pinkas, that implements a type of constitution of the Hebrew grammar. This model acts without any vocabulary and base its linguistic analysis only on its constitution.

The forth model is based only on dictionary, assembled by Choueka, that implements an analysis process based on its full Hebrew dictionary, that includes additional information about the formal extensions for each entry.

The fifth model combines both a Heuristic model and a partial Hebrew dictionary, based on Carmel's method.

The chapter dealing with Methodology defines criteria for evaluating the models based on several aspects. It starts with the analysis of a single word, going through the analysis of a complete sentence, and eventually, evaluation of an analysis method of the spoken Hebrew language.

The comparative analysis that is based on this criteria shows that there are differences between the various methods with respect to the analysis types (formal texts, slang and text with spelling mistakes). It also shows that models that are based on a dictionary (statistcal, conceptual and dictionary) demonstrate an advantage when used with formsl text, while the heuristic models demonstrate an advantage when used with slang or text that include spelling mistakes.

The conclusion chapter presents minor enhancements to the to the described models. Yet, it looks like that the precision level of the various original models is quite high, so we do not accept a major breakthrough by introducing these enhancements.

Therefore, at the end of this paper, a new method is introced. This method introduces a text pre- analysis stage that classifies the text and select the linguistic analysis model that best fit the applied text. We assume this method can provide more accuratr linguistic analysis results over a broad scope of text types and styles.