A Computational Lexicon of Contemporary Hebrew

Alon Itai
Department of Computer Science
Technion, Israel Institute of Technology
32000 Haifa, Israel
itai@cs.technion.ac.il

Shuly Wintner and **Shlomo Yona**
Department of Computer Science
University of Haifa
31905 Haifa, Israel
shuly@cs.haifa.ac.il

Abstract

Computational lexicons are among the most important resources for natural language processing (NLP). Their importance is even greater in languages with rich morphology, where the lexicon is expected to provide morphological analyzers with enough information to enable them to correctly process intricately inflected forms. We describe the Haifa Lexicon of Contemporary Hebrew, the broadest-coverage publicly available lexicon of Modern Hebrew, currently consisting of over 20,000 entries.

While other lexical resources of Modern Hebrew have been developed in the past, this is the first publicly available large-scale lexicon of the language. In addition to supporting morphological processors (analyzers and generators), which was our primary objective, the lexicon is used as a research tool in Hebrew lexicography and lexical semantics. It is open for browsing on the web and several search tools and interfaces were developed which facilitate on-line access to its information. The lexicon is currently used for a variety of NLP applications.

Overview

Computational lexicons are among the most important resources for NLP. In languages with rich morphology, where the lexicon is expected to provide morphological analyzers with enough information to enable them to process intricately inflected forms correctly, a careful design of the lexicon is crucial. This paper describes the Haifa Lexicon of Contemporary Hebrew, the broadest-coverage publicly available lexicon of Modern Hebrew, currently consisting of over 20,000 entries. Table 1 lists the number of words in the lexicon by main part of speech.

noun	10332	preposition	100
verb	4485	conjunction	62
properName	4227	pronoun	60
adjective	1612	interjection	40
adverb	352	interrogative	9
quantifier	132	negation	6
Total:			21,417

Table 1: Size of the lexicon by part of speech

While other lexical resources of Modern Hebrew have been developed in the past (see Wintner (2004) for a survey), this is the first publicly available large-scale lexicon of the language. It is open for browsing on the web and several search tools and interfaces were developed which facilitate on-line access to its information. The lexicon was designed for supporting state of the art morphological processing of Hebrew, and it is now the core on which a morphological grammar (Yona and Wintner, 2005) is based. Additionally, it is currently used for a variety of applications, including a Hebrew to English machine translation system (Lavie et al., 2004) and monolingual and cross-lingual information retrieval. The lexicon is also used as a research tool in Hebrew lexicography and lexical semantics, as well as in psycho-linguistic research where word frequency and root frequency information is required.

Modern Hebrew

Hebrew is one of the two official languages of the State of Israel, spoken natively by half of the population and fluently by virtually all the (seven million) residents of the country. Hebrew exhibits clear Semitic behavior. In particular, its lexicon, word formation and inflectional morphology are typically Semitic. The major word formation machinery is root-and-pattern, where roots are sequences of consonants (typically three) and patterns are sequences of vowels and, sometimes, also consonants, with "slots" into which the root's consonants are inserted. Inflectional morphology is highly productive and consists mostly of suffixes, but sometimes of prefixes or circumfixes.

The Hebrew script, like the Arabic one, attaches several short particles to the word which immediately follows them. These include, *inter alia*, the definite article ה *h* "the", prepositions such as ב *b* "in", כ *k* "as", ל *l* "to" and מ *m* "from", subordinating conjunctions such as ש *š* "that" and כש *kš* "when", relativizers such as ש *š* "that" and the coordinating conjunction ו *w* "and". One of the reasons for the ambiguity of the Hebrew script is that in many words letters can be analyzed as either belonging to a prefix particle or to the stem.

An added complexity stems from the fact that there exist two main standards for the Hebrew script: one (*dotted* or *vocalized*) in which vocalization diacritics, known as *niqqud* "dots", decorate the words, and another (*undotted*) in which the dots are missing, and other characters represent some, but not all of the vowels. Most of the texts in Hebrew are of the latter kind; unfortunately, different authors use different conventions for the undotted script. Thus, the same word can be written in more than one way,

sometimes even within the same document. This fact adds significantly to the degree of ambiguity.

Structure

The lexicon is represented in XML (Connolly, 1997) as a list of *item* elements, each with a base form which is the citation form used in conventional dictionaries. For nouns and adjectives it is the absolute singular masculine, whereas for verbs it is the third person singular masculine, past tense. Contemporary Hebrew dictionaries are ordered by lexeme rather than root, and we maintain, similarly to Dichy and Farghaly (2003), that this is a desirable organization. Still, the lexicon lists for each verb its root and pattern; this was made possible due to the way verbs were acquired, see below.

Lexicon items are specified for the following attributes: a unique *id*, three representations of the lexical entry (dotted, undotted and transliterated) and *script*, which encodes deviations from the standard script as well as register. In addition, every lexicon item belongs to a *part of speech* category, as listed in Table 1. The part of speech of an entry determines its additional attributes. For *nominals*, which are nouns, adjectives and numerals, these include number, gender and nominal status (absolute or construct). Verbs are specified for number, gender, person and tense, as well as for root and pattern. We also list the type of proper names (person, location, organization or date).

The lexicon specifies morpho-syntactic features (such as gender or number), which can later be used by parsers and other applications. But it also lists several lexical proerties which are specifically targeted at morphological analysis. A typical example is the plural suffix for nouns: while by default, this suffix is *im* for masculine nouns and *wt* for feminine, many lexical items are idiosyncratic. The lexicon lists information pertaining to non-default behavior with idiosyncratic entries.

The lexical representation of verbs is more involved. Here, the lexicon stores two main pieces of information: a root and an *inflection pattern* (IP). The latter is a combination of the traditional *binyan* with some information about peculiarities of the inflectional paradigm of verbs in this *binyan*. Such information is required because of some arbitrariness in the way verbs inflect, even in the regular patterns. For example, the second person singular masculine future form of the roots *p.s.l* and *š.k.b* in the first *binyan* (*pa'al*) is *tipswl* and *tiškb*, respectively. Note the additional '*w*' in the first form which is missing in the second: both roots are regular, and such information must be encoded in the lexicon to indicate the different inflected forms.

Irregularity and idiosyncrasy can be expressed directly in the lexicon, in the form of additional or alternative lexical entries. This is facilitated by the use of three optional elements in lexicon items: *add*, *replace* and *remove*. For example, the noun צהריים *chriim* "noon" is also commonly spelled צהרים *chrim*, so the additional spelling is specified in the lexicon, along with the standard spelling, using *add*. The verb יכול *ikwl* "can" does not have imperative inflections, which are generated by default for all verbs. To prevent the default behavior, the superfluous forms are *remove*d. Figure 1 lists a few (partial) lexicon items.

Sometimes the citation form which is specified in the lexicon is not the most convenient one for generating the inflection paradigm. For example, the preposition עם *&m* "with" is a citation form, whose entire inflection paradigm is much simpler if *&im* is used as the base. For such cases we use a mechanism based on an additional attribute, *inflectionBase*, which causes the entire paradigm to be generated with the alternative base. See Figure 2.

Interaction with Morphological Processing

The quality of a morphological analyzer greatly depends on the quality of the lexicon. A morphological analyzer must consult with the lexicon to check whether a theoretical analysis of a word indeed belongs to the language. Since searches in XML files are sequential, and hence very slow, we converted the XML files to a MySQL database (DuBois, 1999); morphological analyzers can thus access the lexicon via a standard query language (SQL). The current stable version of the lexicon is stored in the database, and its XML mirror is generated upon request.

This organization facilitates a modular development of morphological analysis and disambiguation systems. The morphological analyzer interacts with, but is separated from, the lexicon. Currently, the lexicon is used by two different morphological analyzers. It is also used independently by a morphological annotation tool and by a Hebrew to English machine translation system (Lavie et al., 2004).

Our current morphological analyzer performs *analysis by generation*: this is basically the same technique that was used by Shapira and Choueka (1964) in the first computational analyzer of Hebrew. The basic idea is to first generate all the inflected forms induced by the lexicon and store them in a database; then, analysis is simply a database lookup. It is common to think that for languages with rich morphology such a method is impractical. While this may have been the case in the past, contemporary computers can efficiently store and retrieve millions of inflected forms. Of course, this method would break in the face of an infinite lexicon (which can easily be represented with FST), but for most practical purposes it is safe to assume that natural language lexicons are finite.

The morphological analyzer is obtained by inflecting the base forms in the lexicon. The number of inflected forms (before attaching prefixes) is 473,880 (over 300,000 of those are inflected nouns, and close to 150,000 are inflected verb forms). In addition to inflected forms, the analyzer also allows as many as 174 different sequences of prefix particles to be attached to words; of course, not all sequences combine with all forms (for example, the definite article cannot combine with an adverb). Theoretically, it could be possible to generate all the possible surface forms in Hebrew by combining prefix sequences with inflected words, but we estimate the number of such forms to be a few millions. The inflected forms are stored in a database and are used by the analysis program.

```
- <item dotted="צָהֳרַיִם" id="372" script="formal" transliterated="chriim" undotted="צהריים">
  - <noun gender="masculine" number="dual and plural">
      <add gender="masculine" number="dual and plural" script="colloquial" transliterated="chrim"
      undotted="צהרים"/>
    </noun>
  </item>
- <item id="17580" script="formal" transliterated="bwqr" undotted="בוקר">
  - <noun gender="masculine" number="singular" plural="im">
      <replace gender="masculine" number="plural" script="formal" transliterated="bqrim"
      undotted="בקרים"/>
    </noun>
  </item>
- <item id="4025" script="formal" transliterated="ikwl" undotted="יכול">
  - <verb binyan="Pa'al" feminine="t" inflectionPattern="1" root="יכל">
      <remove pgn="2p/M/Sg" tense="imperative" transliterated="ikwl" undotted="יכול"/>
      <remove pgn="2p/F/Sg" tense="imperative" transliterated="ikli" undotted="יכלי"/>
      <remove pgn="2p/M/Pl" tense="imperative" transliterated="iklw" undotted="יכלו"/>
      <remove pgn="2p/F/Pl" tense="imperative" transliterated="ikwlnh"
      undotted="יכולנה"/>
    </verb>
  </item>
```

Figure 1: Examples of lexicon entries

```
- <item dotted="עִם" id="8098" script="formal" transliterated="ym"
  undotted="עם">
    <preposition case="unspecified" inflectionBase="עימ"/>
  </item>
```

Figure 2: Examples of a lexicon entry with an alternative inflection base

Acquisition

The lexicon was initially populated with a small number of words in order to develop a morphological analyzer. Then, approximately 3000 nouns and adjectives were automatically acquired from the HSpell lexicon (Har'El and Kenigsberg, 2004). We also incorporated many of the lexical items of Segal (1997)'s morphological analyzer. Over 3500 verbs were added by typing in the roots and inflection bases of Zdaqa (1974), which is a list of the full inflection paradigms of all Hebrew verbs.

Remaining entries were added manually by a lexicographer using a graphical user interface specifically designed for this purpose (Figure 3). In adding new words we follow several strategies. First, we use the morphological analyzer on dynamic corpora (e.g., on-line newspapers) and manually inspect words which the analyzer does not recognize. Second, we use the morphological generator to produce certain derivations of existing forms and match them against the lexicon. For example, we automatically generated deverbal forms of all the verbs in the lexicon, and compared them with existing nominal forms; we also generated passive voices from active verbs and tested them in the same manner. Finally, we employ linguists who go over existing entries and suggest modifications and corrections. A recent change we introduced in this way is a treatment of present tense verbs as *middles*, which inflect like nominals. This process is still ongoing, although we currently focus mainly on named entities. Over 16,000 of the entries in the lexicon are dotted, and we continue to add dotted forms to the remaining entries.

Future work

We are currently working on two extensions of the lexicon. First, we add bilingual (Hebrew-English) word translation to its items, thereby extending it to a full bilingual dictionary. This process is ongoing, and approximately half of the entries are already associated with translations. Second, we are interested in extending the lexicon also to multiword tokens, which are abundant in Hebrew. We are currently designing this extension.

Acknowledgments

This work was funded by the Israeli Ministry of Science and Technology, under the auspices of the Knowledge Center for Processing Hebrew. We are grateful to Shira Schwartz, Danny Shacham and Michael Elhadad for their help.

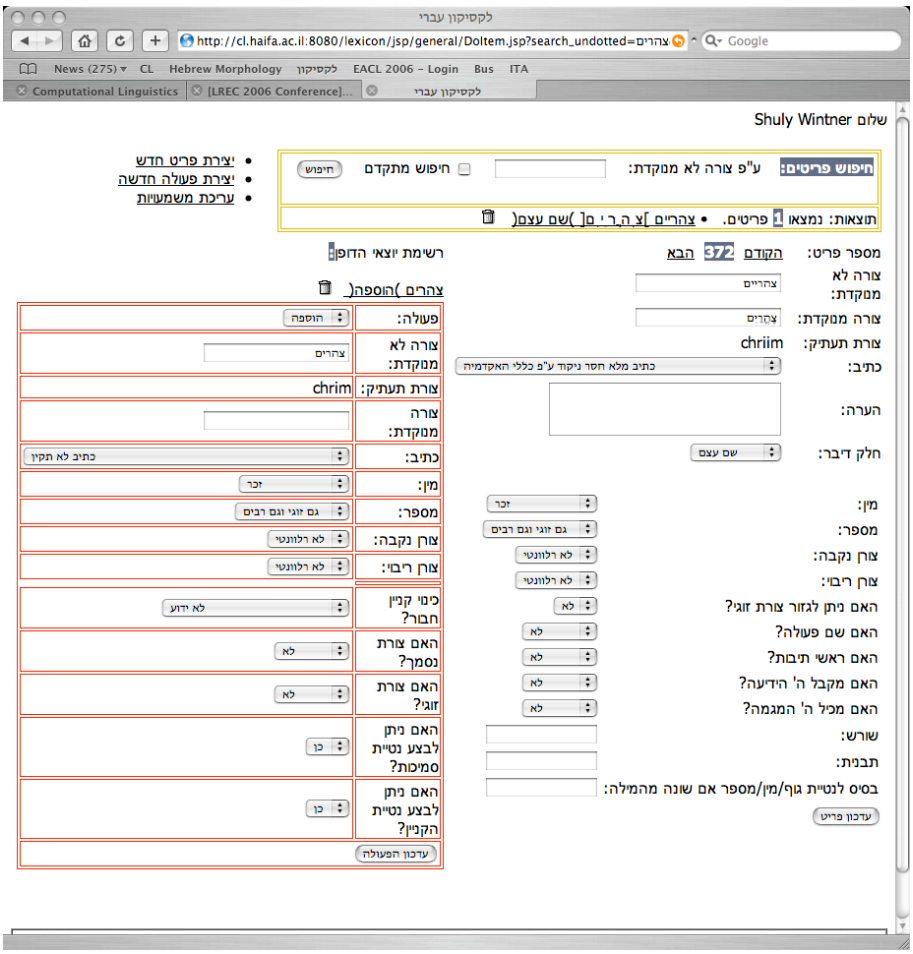

Figure 3: Graphical user interface for lexicon maintenance

References

Connolly, Dan. 1997. *XML: Principles, Tools, and Techniques*. O'Reilly.

Dichy, Joseph and Ali Farghaly. 2003. Roots and patterns vs. stems plus grammar-lexis specifications: on what basis should a multilingual lexical databas centered on Arabic be built. In *Proceedings of the MT-Summit IX workshop on Machine Translation for Semitic Languages*, pages 1–8, New Orleans, September.

DuBois, Paul. 1999. *MySQL*. New Riders.

Har'El, Nadav and Dan Kenigsberg. 2004. Hspell: a free Hebrew speller. Available from `http://www.ivrix.org.il/projects/-spell-checker/`.

Lavie, Alon, Shuly Wintner, Yaniv Eytani, Erik Peterson, and Katharina Probst. 2004. Rapid prototyping of a transfer-based Hebrew-to-English machine translation system. In *Proceedings of TMI-2004: The 10th International Conference on Theoretical and Methodological Issues in Machine Translation*, Baltimore, MD, October.

Segal, Erel. 1997. Morphological analyzer for unvocalized hebrew words. Unpublished work, available from `http://www.cs.technion.ac.il/-~erelsgl/hmntx.zip`.

Shapira, Meir and Yaacov Choueka. 1964. Mechanographic analysis of Hebrew morphology: possibilities and achievements. *Leshonenu*, 28(4):354–372. In Hebrew.

Wintner, Shuly. 2004. Hebrew computational linguistics: Past and future. *Artificial Intelligence Review*, 21(2):113–138.

Yona, Shlomo and Shuly Wintner. 2005. A finite-state morphological grammar of Hebrew. In *Proceedings of the ACL Workshop on Computational Approaches to Semitic Languages*, pages 9–16, Ann Arbor, Michigan, June. Association for Computational Linguistics.

Zdaqa, Yizxaq. 1974. *Luxot HaPoal (The Verb Tables)*. Kiryath Sepher, Jerusalem. In Hebrew.