Language Resources for Hebrew

Alon Itai
Department of Computer Science, Technion, Israel Institute of Technology, 32000 Haifa, Israel

Shuly Wintner
Department of Computer Science University of Haifa 31905 Haifa, Israel

October 25, 2007

Abstract. We describe a suite of standards, resources and tools for computational encoding and processing of Modern Hebrew texts. These include an array of XML schemas for representing linguistic resources; a variety of text corpora, raw, automatically processed and manually annotated; lexical databases, including a broad-coverage monolingual lexicon, a bilingual dictionary and a WordNet; and morphological processors which can analyze, generate and disambiguate Hebrew word forms. The resources are developed under centralized supervision, so that they are compatible with each other. They are freely available and many of them have already been used for several applications, both academic and industrial.

1. Introduction

Language resources are crucial for research and development in theoretical, computational, socio- and psycho-linguistics, and for the construction of natural language processing (NLP) applications. Computational processing of Modern Hebrew (henceforth *Hebrew*) was until recently hindered by the lack of publicly available resources (Wintner, 2004). This paper describes a recent effort whose main goal is to develop, organize and maintain a large-scale set of resources and tools, including an array of XML schemas for representing linguistic resources; a broad-coverage monolingual lexicon; a variety of text corpora, raw, automatically processed and manually annotated; morphological analysis, generation and disambiguation systems; and a Hebrew WordNet. Most of the resources are distributed under the Gnu Public License, and are freely available for research and commercial purposes. They have been extensively used for both research and commercial applications in the past few years, and are regularly maintained and supported. While parts of this project have been presented elsewhere (Wintner and Yona, 2003; Yona and Wintner, 2005; Bar-Haim et al., 2005; Itai, 2006; Adler and Elhadad, 2006; Itai et al., 2006; Wintner, 2007; Yona and Wintner, 2007; Ordan and Wintner, 2007), this paper provides a general framework for the full-scale project, extending, updating and elaborating on previous discussions.

The main contribution of this paper is a report on a set of resources which will be of practical use to anyone with interest in linguistic investigation or computational processing of Hebrew and other Semitic languages. More generally, we believe that some of the design decisions we have made during the development of the resources (e.g., the use of XML; the organization of the XML schemas; the interactions between morphology and the lexicon; and the modularization of software development) have a more global scope and are applicable to similar projects for other languages with complex morphological and orthographic systems.

After a brief introduction to Hebrew morphology and orthography in Section 1.1, we discuss some design decisions that lead to the definition of various standards in Section 1.2. We then describe the lexical databases (Section 2), morphological processing tools (Section 3) and corpora (Section 4). We conclude with plans for further research.

1.1. Linguistic background

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 (over seven million) residents of the country. Hebrew exhibits clear Semitic behavior. In particular, its lexicon, word formation and inflectional morphology are typically Semitic.

Hebrew morphology is rich and complex.[1] The major word formation machinery is root-and-pattern, and inflectional morphology is highly productive and consists of prefixes, suffixes and circumfixes. Nouns, adjectives and numerals inflect for number (singular, plural and, in rare cases, also dual) and gender (masculine or feminine). In addition, all these three types of nominals have two phonologically and morphologically distinct forms, known as the *absolute* and *construct* states. In the standard orthography approximately half of the nominals appear to have identical forms in both states, a fact which substantially increases the ambiguity. In addition, nominals take possessive pronominal suffixes which inflect for number, gender and person.

Verbs inflect for number, gender and person (first, second and third) and also for a combination of tense and aspect/mood, referred to simply as 'tense' below. Verbs can also take pronominal suffixes, which are interpreted as direct objects, and in some cases can also take nominative

[1] To facilitate readability we use a straight-forward transliteration of Hebrew in this paper, where the characters (in Hebrew alphabetic order) are: abgdhwzxviklmnsypcqršt. In our resources, we use both a UTF-8 encoding of Hebrew and an ASCII transliteration, which differs from the above in two letters: '↔y and š↔e.

pronominal suffixes. A peculiarity of Hebrew verbs is that the participle form can be used as present tense, but also as a noun or an adjective.

These matters are complicated further due to two reasons: first, the standard Hebrew orthography (undotted script) leaves most of the vowels unspecified. On top of that, the script dictates that many particles, including four of the most frequent prepositions, the definite article, the coordinating conjunction and some subordinating conjunctions, all attach to the words which immediately follow them. When the definite article *h* is prefixed by one of the prepositions *b*, *k* or *l*, it is assimilated with the preposition and the resulting form becomes ambiguous as to whether or not it is definite. For example, *bth* can be read either as *b+th* "in tea" or as *b+h+th* "in the tea". Thus, the form *šbth* can be read as an inflected stem (the verb "capture", third person singular feminine past), as *š+bth* "that+field", *š+b+th* "that+in+tea", *š+b+h+th* "that in the tea", *šbt+h* "her sitting" or even as *š+bt+h* "that her daughter".

An added complexity stems from the fact that Hebrew can be written in two ways: one in which vocalization diacritics, known as *niqqud* "dots", decorate the words, and the undotted script, in which the dots are missing, and other characters represent some, but not all of the vowels.[2] Most of the texts in Hebrew are of the latter kind. Even though the Academy for the Hebrew Language has issued guidelines for transcribing undotted texts (Gadish, 2001), they are observed only partially. Thus, the same word can be written in more than one way, sometimes even within the same document. For example, *chriim* "noon" can be spelled *chrim* (see also Figure 1 below). This fact adds significantly to the degree of ambiguity.

1.2. Design decisions

In order to integrate the various resources a common interface had to be decided upon. The organization is motivated by the following principles:

Portability The format should be platform independent;

Readability The representation should allow for easy production of annotations, easy parsing and processing of the annotated data, by both machines and humans;

[2] The undotted script is sometimes referred to as *ktiv male* "full script", whereas the dotted script, without the diacritics, is called *ktiv xaser* "lacking script". These terms are misleading, as any representation that does not depict the diacritics lack many of the vowels.

Standardization Processing of the annotated data should be supported by a wide variety of environments (information processing tools, programming languages, etc.);

Reversibility The original data should be easily extracted from the annotated version if desired;

Openness The tools used to produce the resources and the production steps of the annotated data should be publicly available, to allow the recreation of the data or further development.

Our linguistic databases are represented in Extensible Markup Language (XML, Connolly (1997)) according to schemas (van der Vlist, 2002) that enforce structure and are also used for documentation and validation purposes. XML is a method of describing structured data. It is a simple, very flexible text format which is playing an increasingly important role in the exchange of a wide variety of data on the Web and elsewhere (Sperberg-McQueen and Burnard, 2002). We take advantage of the portability of XML documents and the wide availability of XML processing tools in order to facilitate access to our resources by users of any platform. Another advantage of using XML is that even though XML documents are meant to be easily processed by machines, it is possible for a human to view an XML document and understand its content simply by reading it, as XML documents are plain text files. In addition, there exist tools to visualize XML files via web browsers.

To achieve modularity, various language processing programs are integrated with our linguistic databases through XML: they are designed as standalone modules whose input and output is XML. We can, for instance, replace one morphological analyzer by another without affecting programs that use the output of the analyzer. We focus in this paper on the XML schemas used for representing the two major databases, namely corpora and lexicons.

1.3. Availability

All the resources that we list in this paper are publicly available and can be directly downloaded from the main website, `http://www.mila.cs.technion.ac.il/`. Non-profit use is allowed under the Gnu Public License; but we also enable incorporation of the resources in commercial products, under special licenses. In addition to members of this project, the resources were downloaded (and presumably used) by several academic institutions in Israel and many in the rest of the world, including the universities of Amsterdam and Utrecht in Holland, Manchester in the UK, The Ohio State University, Carnegie Mellon University, University of Illinois Urbana Champaign, University of California Berkeley, University of Pennsylvania and MITRE in the US, NAIST in Japan, and the Institute of Biodiagnostics in Canada.

2. Lexical databases

2.1. 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 section (which updates and extends Itai et al. (2006)) describes the Haifa Lexicon of Contemporary Hebrew, the broadest-coverage publicly available lexicon of Hebrew, currently consisting of over 22,000 entries. Table I lists the number of words in the lexicon by main part of speech (POS).

Table I. Size of the lexicon by part of speech

POS	#entries	POS	#entries
noun	11361	numeral	59
verb	4485	interjection	43
proper name	3408	quantifier	34
adjective	2492	modal	33
adverb	426	word prefix	19
preposition	109	interrogative	18
conjunction	83	negation	7
pronoun	77	existential	2
Total:			22,656

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 (Section 3) 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 (Szpektor et al., 2007). 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.

2.2. Structure

The structure of the lexicon is optimized for morphological processing of Hebrew, although a mapping of this structure to a more general one, such as the Lexical Markup Framework (ISO 24613), should be straight-forward. The lexicon is represented in XML as a list of *item* elements,[3] 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 (undotted, transliterated and dotted[4]) and *script*, which encodes deviations from the standard script as well as register. In addition, every lexicon item belongs to a (single) *part of speech* category, as listed in Table I. The part of speech of an entry determines its additional attributes. For *nominals*, i.e., nouns, adjectives and numerals, these include number and gender; verbs are specified for root and inflection pattern (see below). 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 properties 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 forms of the

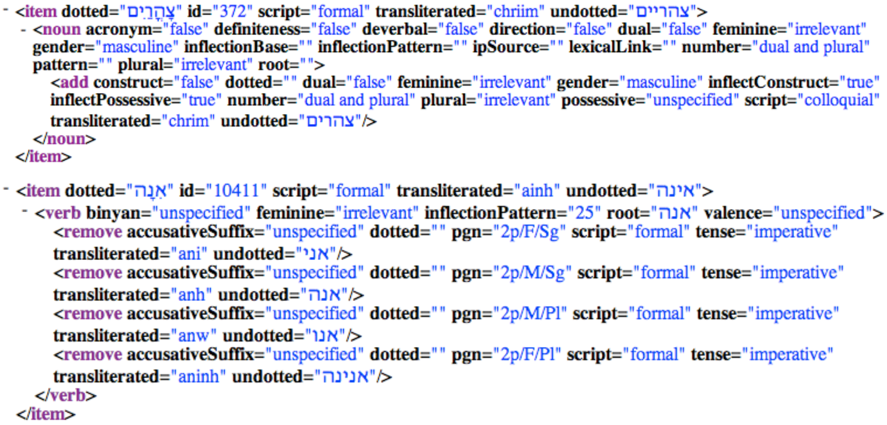
[3] These are often called *entry* in similar projects.

[4] 13,475 of the 22,656 entries in the lexicon are dotted, and we continue to add dotted forms to the remaining entries.

roots *p.s.l* and *š.k.b* in the first *binyan* (*pa'al*) are *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.

The lexicon also contains information concerning the valency of verbs. In order to avoid linguistic controversies, we distinguished only between transitive and intransitive verbs, and also noted whether the passive participle exists. More information should be added, hopefully incorporating and completing the monumental research conducted by Stern (1994).

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 *chriim* "noon" is also commonly spelled *chrim*, so the additional spelling is specified in the lexicon, along with the standard spelling, using *add*. The verb *anh* "harm" 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 demonstrates the structure of the lexicon.

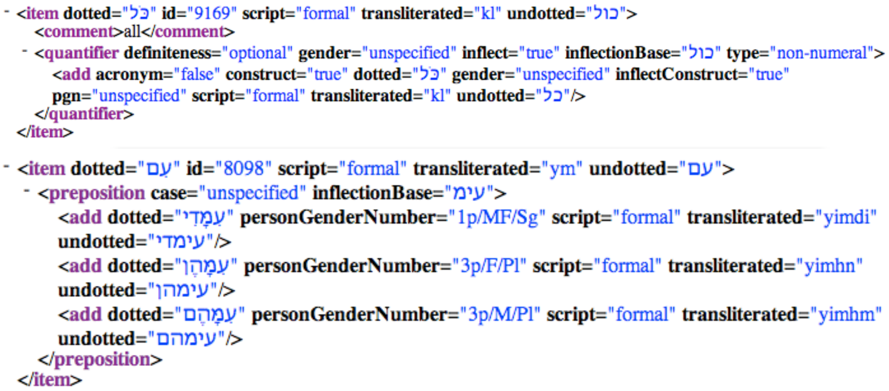


```
- <item dotted="צָהֳרַיִם" id="372" script="formal" transliterated="chriim" undotted="צהריים">
  - <noun acronym="false" definiteness="false" deverbal="false" direction="false" dual="false" feminine="irrelevant"
    gender="masculine" inflectionBase="" inflectionPattern="" ipSource="" lexicalLink="" number="dual and plural"
    pattern="" plural="irrelevant" root="">
      <add construct="false" dotted="" dual="false" feminine="irrelevant" gender="masculine" inflectConstruct="true"
      inflectPossessive="true" number="dual and plural" plural="irrelevant" possessive="unspecified" script="colloquial"
      transliterated="chrim" undotted="צהרים"/>
    </noun>
  </item>

- <item dotted="אִנָּה" id="10411" script="formal" transliterated="ainh" undotted="אינה">
  - <verb binyan="unspecified" feminine="irrelevant" inflectionPattern="25" root="אנה" valence="unspecified">
      <remove accusativeSuffix="unspecified" dotted="" pgn="2p/F/Sg" script="formal" tense="imperative"
      transliterated="ani" undotted="אני"/>
      <remove accusativeSuffix="unspecified" dotted="" pgn="2p/M/Sg" script="formal" tense="imperative"
      transliterated="anh" undotted="אנה"/>
      <remove accusativeSuffix="unspecified" dotted="" pgn="2p/M/Pl" script="formal" tense="imperative"
      transliterated="anw" undotted="אנו"/>
      <remove accusativeSuffix="unspecified" dotted="" pgn="2p/F/Pl" script="formal" tense="imperative"
      transliterated="aninh" undotted="אנינה"/>
    </verb>
  </item>
```

Figure 1. Examples of lexical entries

Sometimes the citation form which is specified in the lexicon is not the most convenient one for generating the inflection paradigm. For example, the quantifier *kl* "all" is a citation form, whose entire inflection paradigm is much simpler if *kwl* is used as the base. Similarly, the inflection paradigm of the preposition *'m* "with" is simpler if *'im* is used as the stem. 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.



```
- <item dotted="כֹּל" id="9169" script="formal" transliterated="kl" undotted="כול">
    <comment>all</comment>
  - <quantifier definiteness="optional" gender="unspecified" inflect="true" inflectionBase="כול" type="non-numeral">
      <add acronym="false" construct="true" dotted="כֹּל" gender="unspecified" inflectConstruct="true"
      pgn="unspecified" script="formal" transliterated="kl" undotted="כל"/>
    </quantifier>
  </item>
- <item dotted="עִם" id="8098" script="formal" transliterated="ym" undotted="עם">
  - <preposition case="unspecified" inflectionBase="עימ">
      <add dotted="עִמָּדִי" personGenderNumber="1p/MF/Sg" script="formal" transliterated="yimdi"
      undotted="עימדי"/>
      <add dotted="עִמָּהֶן" personGenderNumber="3p/F/Pl" script="formal" transliterated="yimhn"
      undotted="עימהן"/>
      <add dotted="עִמָּהֶם" personGenderNumber="3p/M/Pl" script="formal" transliterated="yimhm"
      undotted="עימהם"/>
    </preposition>
  </item>
```

Figure 2. Lexicon entries with alternative inflection bases

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 and other applications (in particular, the GUI that lexicographers use to manipulate the lexicon, see Section 2.3) 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. Our morphological processors interact with these resources indirectly: a finite-state morphological analyzer uses a converted version of the XML database, whereas a Java morphological generator uses the SQL database to generate a database of inflected forms, see Section 3.2.

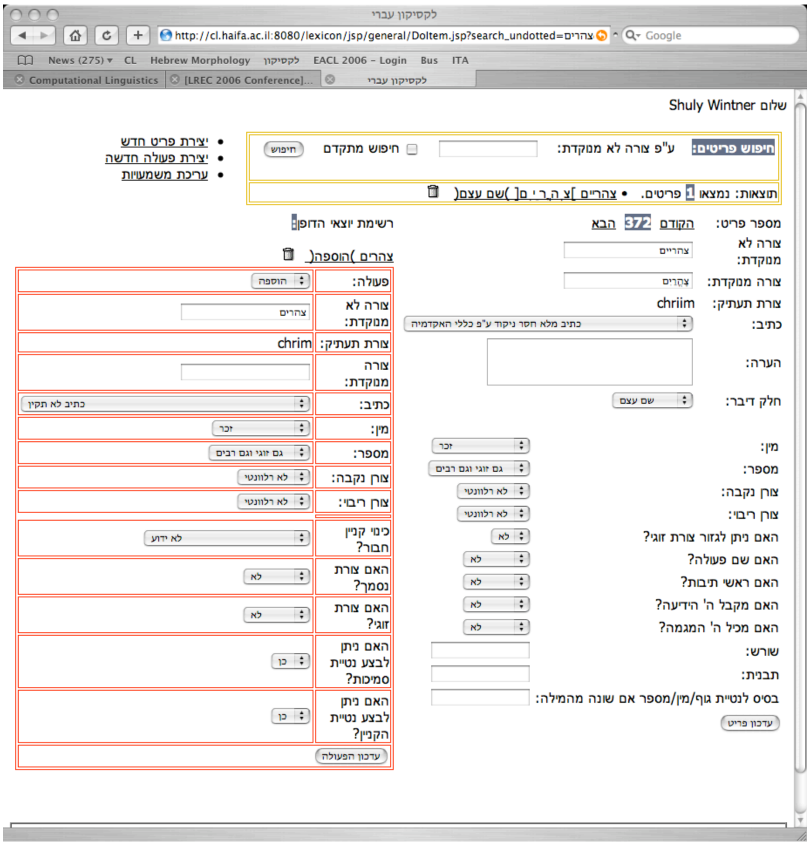
This organization facilitates a modular development of morphological analysis and disambiguation systems. The morphological analyzer interacts with, but is separated from, the lexicon. Currently, the lexicon is used by two different morphological analyzers (see Section 3) and by a morphological annotation tool (Section 4.4).

2.3. 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 used by the morphological analyzer of Segal (1997). Over 3500 verbs were added by typing in roots and inflection bases taken from complete lists of the full inflection paradigms of Hebrew verbs

(Zdaqa, 1974). In subsequent work we used more printed resources, including Barkali (2000a), Barkali (2000b).

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 transitive verbs and tested them in the same manner.



http://cl.haifa.ac.il:8080/lexicon/jsp/general/DoItem.jsp?search_undotted=צהרים

שלום Shuly Wintner

חיפוש פריטים: ע"פ צורה לא מנוקדת: □ חיפוש מתקדם

תוצאות: נמצאו 1 פריטים. • צהריים]צָהֳרַיִם[ )שם עצם(

מספר פריט: הקודם 372 הבא

צורה לא מנוקדת: צהריים

צורה מנוקדת: צָהֳרַיִם

צורת תעתיק: chriim

כתיב: כתיב מלא חסר ניקוד ע"פ כללי האקדמיה

הערה:

חלק דיבר: שם עצם

מין: זכר

מספר: גם זוגי וגם רבים

צורן נקבה: לא רלוונטי

צורן ריבוי: לא רלוונטי

האם ניתן לגזור צורת זוגי? לא

האם שם פעולה? לא

האם ראשי תיבות? לא

האם מקבל ה' הידיעה? לא

האם מכיל ה' המגמה? לא

שורש:

תבנית:

בסיס לנטיית גוף/מין/מספר אם שונה מהמילה:

רשימת יוצאי הדופן:

צהרים)הוספה(

פעולה:	הוספה
צורה לא מנוקדת:	צהרים
צורת תעתיק:	chrim
צורה מנוקדת:	
כתיב:	כתיב לא תקין
מין:	זכר
מספר:	גם זוגי וגם רבים
צורן נקבה:	לא רלוונטי
צורן ריבוי:	לא רלוונטי
כינוי קניין חבור?	לא ידוע
האם צורת נסמך?	לא
האם צורת זוגי?	לא
האם ניתן לבצע נטיית סמיכות?	כן
האם ניתן לבצע נטיית הקניין?	כן

Figure 3. Graphical user interface for lexicon maintenance

Finally, we employ linguists who go over existing entries and suggest modifications and corrections. Recent changes that we introduced in this way include a treatment of present tense verbs as *participles*, which inflect like nominals; and a finer classification of *modals*. The lexical acquisition process is still ongoing.

The vocabulary of Modern Hebrew is significantly smaller than that of English. In realistic evaluations on random texts the rate of out-of-vocabulary items is constantly below 5%, and the vast majority of those (80%) are proper names. See also Table IV in Section 3.2.

2.4. Multilingual extensions

The design of the lexicon is compatible with another language resource, the Hebrew WordNet (Section 2.5). To fully integrate the two databases we extended the lexicon schema to support also bilingual entries in the form of translation equivalents for each lemma. Following standard lexicographic conventions, each lexicon *item* is further divided into one or more *senses*; each sense, then, inherits from its *item* the morphological and morpho-syntactic information that is exemplified in Figure 1, but includes in addition a pointer to a WordNet *synset*, followed by a list of translation equivalents (to English). Each translation equivalent in the list is a pair consisting of an English lemma and a weight, which encodes information about the likelihood of the translation equivalent (where more frequent translations are heavier). Weight information has not been acquired yet.

English translation equivalents were acquired from a small bilingual dictionary (Dahan, 1997) for which we acquired the rights. Obtaining permission to use larger-scale dictionaries proved impossible, and hence we resorted to manual extension of the dictionary by lexicographers. Currently, 12,122 of the Hebrew lemmas are translated to English, yielding over 20,000 translation pairs.

In addition, we automatically acquired Hebrew-English term pairs from Wikipedia. Following the Wikipedia links to multiple languages, we extracted only the *title* of each document pair. This yielded 41,877 entries, most of which are proper names or technical terminology items. They will be added to the dictionary after manual confirmation by a lexicographer.

The Hebrew-English dictionary was instrumental for a Hebrew to English machine translation system (Lavie et al., 2004) and for the development of a cross-lingual information retrieval system (Szpektor et al., 2007).

2.5. WordNet for Hebrew

WordNet (Fellbaum, 1998) is a computational lexicographical resource which was motivated by psycholinguist concerns but turned out to be instrumental for a variety of computational tasks (Harabagiu, 1998). WordNet is used for information retrieval (Mandala et al., 1998), word-sense disambiguation (Agirre and Rigau, 1996), text categorization (de Buenaga Rodríguez et al., 1997), language generation (Jing, 1998), and semantic annotation (Fellbaum et al., 2001), to name a few examples. Furthermore, the success of the original English WordNet boosted the preparation of similar resources for other languages, and there are currently at least forty WordNet projects in other languages, completed or underway. There are obviously good reasons for compiling, maintaining and distributing WordNets for new languages.

We developed a medium-sized WordNet for Hebrew (Ordan and Wintner, 2007), cast in the MultiWordNet paradigm (Bentivogli et al., 2002). The network is thus synchronized with similar WordNets for other languages (currently, English, Italian, Spanish and Romanian). Hebrew is the first Semitic language for which a substantial WordNet has been designed (for preliminary attempts to create an Arabic WordNet, cf. Diab (2004) and Black et al. (2006)). The Hebrew WordNet currently contains 5261 synsets, with an average of 1.47 synonyms per synset, where nouns are much more frequent than other parts of speech (almost 78 percent, see Table II).

Table II. Current state of the Hebrew WordNet

POS	#synsets
Nouns	4090
Verbs	609
Adjectives	779
Adverbs	151
total	5261

3. Morphological Processing

This section describes a set of tools and programs for morphological processing, including tokenization, analysis, generation and disambiguation. All the resources interact with the lexicon discussed above.

3.1. Tokenization

Partitioning raw Hebrew data into tokens (words) is slightly more involved than in English due to issues of Hebrew encoding, mixed Hebrew/English, numbers, punctuation etc. We developed a tokenization module which operates on raw data (UTF-8 encoded) and produces an XML corpus. The module is capable of segmenting texts into paragraphs, sentences and tokens. The XML format of the output is discussed in Section 4.2.

3.2. Morphological analysis and generation

Morphological analysis is a crucial component of most NLP systems. Whether the goal of an application is information retrieval, question answering or machine translation, NLP applications must be aware of word structure. For some languages and for some applications, simply stipulating a list of surface forms is a viable option; this is not the case for languages with complex morphology, in particular Hebrew, both because of the huge number of potential forms and because of the complete inability of such an approach to handle out-of-lexicon items. The number of such items in Hebrew is significantly larger than in many European languages due to the combination of prefix particles with open-class words such as proper names. An alternative solution would be a dedicated morphological analyzer, implementing the morphological and orthographic rules of the language.

We developed a large-scale morphological grammar of Hebrew, HAMSAH[5] (Yona and Wintner, 2005; Yona and Wintner, 2007), based on finite-state technology (Beesley and Karttunen, 2003). The grammar consists of a finite-state version of the lexicon described in Section 2, and a set of linguistically motivated morphological rules. HAMSAH is the broadest-coverage and most accurate publicly available morphological analyzer of Modern Hebrew. To the best of our knowledge, this is the first formal grammar for the morphology of Modern Hebrew.

The finite-state solution, however, turned out to be sub-optimal. Several problems were encountered during the development and maintenance of the grammar, including poor compile-time performance, unreasonable memory requirements and lack of abstraction which resulted in maintenance difficulties (Wintner, 2007). Consequently, we reimplemented the analyzer in Java. 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

[5] HAifa Morphological System for Analyzing Hebrew.

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, but for most practical purposes it is safe to assume that natural language lexicons are finite. This is certainly the case for Hebrew.

Our 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 157 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 over 100 million, making it impractical to store them all in main memory. Similarly, it would have been possible to separately store a list of suffixes in addition to prefixes, and have a lexicon of stems not unlike the Arabic lexicon of Buckwalter (2002). Our choice balances between time and space requirements in a reasonable way.

The inflected forms are stored in a database and are used by the analysis program. As it turns out, storing a database of half a million inflected forms (along with their analyses) is inexpensive, and retrieving items from the database can be done very efficiently. We experimented with two versions: one uses MySQL as the database and the other loads the inflected forms into a hash table. In this latter version, most of the time is spent on loading the database, and retrieval time is negligible. We compared the performance of the two systems on four tasks, analyzing text files of 10, 100, 1,000 and 10,000 tokens. The results are summarized in Table III. Thus using a hash table at peak performance we are able to analyze 4,000 tokens per second.

Table III. Time performance of morphological analysis (in seconds)

#Tokens	10	100	1,000	10,000
MySQL	1.24	3.04	8.84	44.94
Hash	5.00	5.15	5.59	7.64

Table IV. Total number and percentage of tokens for which the correct analysis was found.

	tokens	proper names	punct.	numerals	prefixes	other
total	1612	128	314	25	22	1123
recognized	1512	64	314	25	22	1087
% recognized	93.8	50	100	100	100	96.8

To evaluate the coverage of the morphological analyzer, we collected a set of Hebrew documents from three sources, comprising 1612 tokens. Approximately 40% of the tokens were taken from news articles in the newspaper *HaAretz*; this is the domain for which the morphological analyzer was originally developed. 40% were taken from news articles in two other Hebrew on-line newspapers, *Ynet* and *NRG*, whose language register and style are rather different. The remaining 20% were taken from out-of-domain texts, including older Hebrew (texts were collected from the *Ben-Yehuda Project*, comparable to the *Gutenberg Project*), blogs, etc.

The results of the evaluation are listed in Table IV. The correct analysis was produced for almost 94% of the tokens. The major omission, as expected, is of proper names, of which only 50% were recognized. Ignoring proper names and punctuation, the correct analysis was produced for 1134/1170, or 96.9% of the tokens. Note that this is a measure of precision; it is much more difficult to measure the recall, i.e., what percentage of possible analyses of a word was produced by the analyzer. We need to compare the analyses produced by our analyzer on a representative corpus to all the analyses of that corpus. The source of the difficulty is that human annotators tend to overlook rare but possible analyses. Thus we are unable to manually produce a corpus with all possible analyses of each word.

The output of the morphological analyzer is subsequently translated to XML, following the specification of a dedicated schema (see Section 4). The schema facilitates the specification of several analyses for each surface form, including an associated weight (which can be set by morphological disambiguation, see below).

3.3. Morphological disambiguation

As noted in Section 1.1, the standard Hebrew script is highly ambiguous. In an annotated corpus of newspaper articles (see Section 4), the average number of analyses per word form is 2.64. Table V lists a histogram of the number of analyses.

Table V. Histogram of analyses

# analyses	# tokens	# analyses	# tokens
1	38468	7	1977
2	15480	8	1309
3	11194	9	785
4	9934	10	622
5	5341	11	238
6	3472	>12	397

Consequently, the output of morphological analysis is ambiguous. The output produced by the analyzer for the form *šbth* is illustrated in Table VI. In general, it includes the part of speech (POS) as well as sub-category, where applicable, along with several POS-dependent features such as number, gender, tense, nominal state, definitness, etc.

Table VI. The analyses of the form *šbth*

#	ID	lemma	POS	Num	Gen	Per	Tense	State	Def	Pref	Suf
1	17280	*šbt*	noun	sing	fem	n/a	n/a	abs	no		h
2	1379	*bt*	noun	sing	fem	n/a	n/a	abs	no	*š*	h
3	19130	*bth*	noun	sing	fem	n/a	n/a	abs	no	*š*	
4	19804	*th*	noun	sing	masc	n/a	n/a	abs	yes	*š+b+h*	
5	19804	*th*	noun	sing	masc	n/a	n/a	abs	no	*š+b*	
6	19804	*th*	noun	sing	masc	n/a	n/a	cons	no	*š+b*	
7	1541	*šbh*	verb	sing	fem	3	past	n/a	n/a		
8	9430	*šbt*	verb	sing	fem	3	past	n/a	n/a		

Identifying the correct morphological analysis of a given word in a given context is an important and non-trivial task. Compared with POS tagging of English, morphological disambiguation of Hebrew is a much more complex endeavor due to the following factors:

Segmentation A single token in Hebrew can actually be a sequence of more than one lexical item. For example, analysis 4 of Table VI (*š+b+h+th* "that+in+the+tea") would correspond to the tag sequence consisting of a subordinating conjunction, followed by a preposition, a determiner and a noun.

Large tagset The number of different tags in a language such as Hebrew (where the POS, morphological features and prefix and suffix particles are considered) is huge. The analyzer produces 22 different parts of speech, some with subcategories; 6 values for the number feature (including disjunctions of values), 4 for gender, 5 for person, 7 for tense and 3 for nominal state. Possessive pronominal suffixes can have 15 different values, and prefix particle sequences can theoretically have hundreds of different forms. While not all the combinations of these values are possible, we estimate the number of possible analyses to be in the thousands.

Ambiguity Hebrew is highly ambiguous: the analyzer outputs on average approximately 2.64 analyses per word token. Oftentimes two or more alternative analyses share the same part of speech, and in some cases two or more analyses are completely identical, except for their lexeme (see analyses 7 and 8 in Table VI). Morphological disambiguation of Hebrew is hence closer to the problem of word sense disambiguation than to standard POS tagging.

Anchors High-frequency function words are almost always morphologically ambiguous in Hebrew. Many of the function words which help boost the performance of English POS tagging are actually prefix particles which add to the ambiguity in Hebrew.

Word order Hebrew word order is relatively free, and in any case freer than in English.

Adler and Elhadad (2006) have developed an HMM-based method to morphologically disambiguate Hebrew texts. They report results on a large scale corpus (6M words) with fully unsupervised learning to be 92.32% for POS tagging and 88.5% for full morphological disambiguation, i.e., finding the correct lexical entry.

Shacham and Wintner (2007) recently developed a morphological disambiguation module for Hebrew. Following Daya et al. (2004) and Habash and Rambow (2005), they approach the problem of morphological disambiguation as a complex classification task. They train a classifier for each of the attributes that can contribute to the disambiguation of the analyses produced by the analyzer (e.g., POS, tense, state). Each classifier predicts a small set of possible values and hence can be highly accurate. In particular, the basic classifiers do not suffer from problems of data sparseness. Of course, each simple classifier cannot fully disambiguate the output of the analyzer, but it does induce a ranking on the analyses. Then, the outcomes of the simple classifiers are combined to produce a consistent ranking which induces a linear order on the analyses. The results are 91.44% accuracy.

These disambiguation modules are fully compatible with the morphological analyzer: they receive as input an XML file consistent with the schema described below (Section 4), where each surface form is analyzed morphologically and all its analyses are listed. The output is a file in the same format, in which each analysis is associated with a weight, reflecting its likelihood in the context. This facilitates the use of the output in applications which may not commit to a *single* correct analysis in a given context.

In addition to full morphological disambiguation, we have adapted a recently developed part of speech tagger for Hebrew (Bar-Haim et al., 2005) to the format of the XML corpus. The tagger is based on a Hidden Markov Model trained on the annotated corpus described in Section 4. Our adaptation of the tagger takes as input a morphologically analyzed corpus (possibly with multiple analyses per word) and produces a corpus in the same format, with only the morphological analyses that are consistent with the most probable POS tagging of the input. The most updated version of the tagger, trained on a treebank of 4500 sentences, boasts 97.2% accuracy for segmentation (detection of underlying morphomes, including a possibly assimilated definite article), and 90.8% accuracy for POS tagging (Bar-haim et al., 2008).

4. The Corpus of Contemporary Hebrew

4.1. General description

Large text corpora are fundamental resources for linguistic and computational linguistic investigations (Abney, 1996; Manning and Schütze, 1999, chapter 4). The Corpus of Contemporary Hebrew is the first large-scale, publicly available corpus of Hebrew. It is available in four levels of annotation:

Raw Raw text with no annotations.

Morphologically analyzed The raw text is tokenized and morphologically analyzed (Section 3).

Morphologically disambiguated Same as above, but the correct analysis in context is manually annotated.

Syntactically parsed A tree-bank of syntactically parsed sentences.

Table VII displays the size (in words) of the corpora. The column under ‘Raw’ indicates the size of the raw corpus, which is also morphologically analyzed. The ‘Manually annotated’ column refers to the corpus which is morphologically disambiguated and syntactically parsed (the size of the tree-bank). Table VIII depicts the distribution of POS in the annotated corpus.

Table VII. The current sizes of the various corpora

	Raw	Manually annotated
Tokens	41,965,058	89,347
Types	510,940	23,947

Note that the main obstacle that prevents the extension of the corpus is copyright: our negotiations with producers of dynamic contents in Israel, notably newspapers and publishing houses, proved futile, but we are constantly seeking other sources of on-line texts which can be added to the corpus.

Table VIII. POS frequencies

POS	# tokens	% tokens
Noun	25836	28.92
Punctuation	13793	15.44
Proper Noun	7238	8.10
Verb	7192	8.05
Preposition	7164	8.02
Adjective	5855	6.55
Participle	3213	3.60
Pronoun	2688	3.01
Adverb	2226	2.49
Conjunction	2021	2.26
Numeral	1972	2.21
Quantifier	951	1.06
Negation	848	0.95
Interrogative	80	0.09
Prefix	29	0.03
Interjection	12	0.01
Foreign	6	0.01
Modal	5	0.01

4.2. Organization

Several initiatives in recent years attempted to define criteria for organizing language resources, and in particular for representing linguistic corpora. These include the *Text Encoding initiative* (Ide and Veronis, 1995) and the *XCES Corpus Encoding Standard* (Ide et al., 2000), as well as a proposed ISO standard (ISO/TC 37/SC 4). The Hebrew corpus generally follows the directives of the proposed ISO standard, as laid out by Ide et al. (2003). Our corpus representation XML schema induces the following structure. A corpus is a sequence of articles, each of which is a sequence of paragraphs which are sequences of sentences.[6] A sentence is a sequence of tokens, and a token contains at least two attributes: *id* and *surface* form (the word in Hebrew script, UTF-8 encoded). In addition, a token may contain morphological analyses. A morphologically analyzed corpus contains all the analyses of a word (as produced by the morphological analyzer), regardless of context. Figures 4, 5 depict all the analyses that are produced by a morphological analyzer for the form *šbth*. Each analysis consists of zero or more *prefix*es, a *base* and an optional *suffix*. The base specifies the properties of the lemma of the token, including its form (both in Hebrew and transliterated), part of speech and POS-dependent features (such as number, gender and nominal state in the case of nouns).

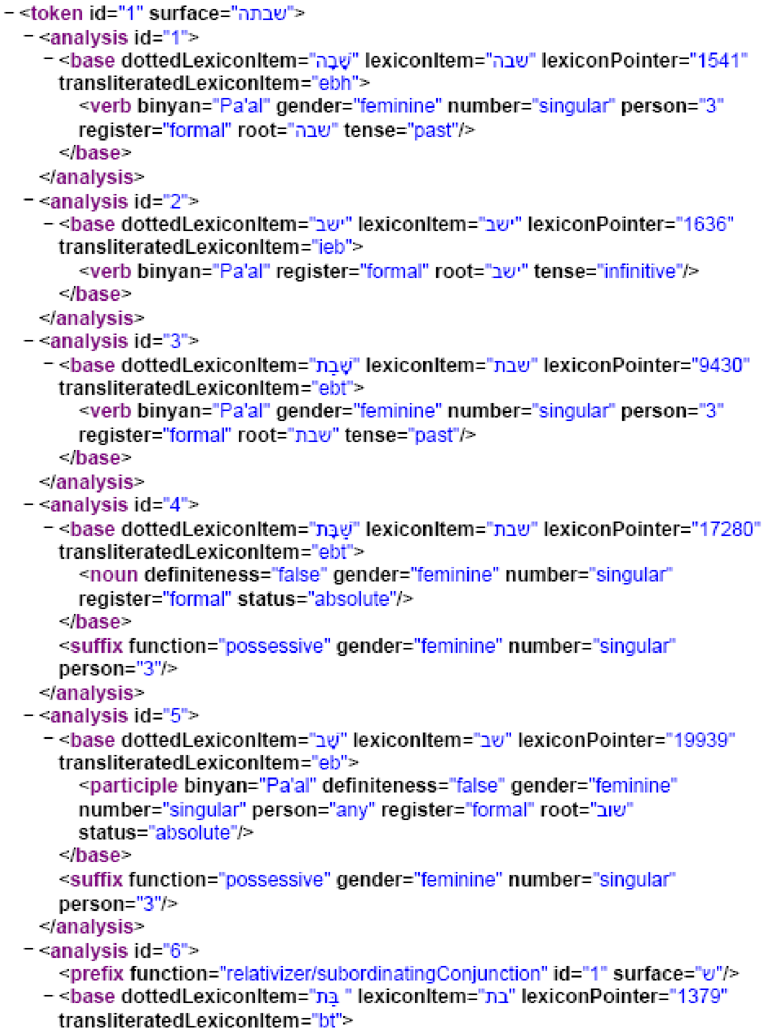
In order to facilitate morphological disambiguation tasks, the corpus representation schema must encode information pertaining to the correct analyses, when contextual information can be used. To this end, we have added an additional attribute, *score*, to each *analysis* element. In a manually annotated corpus, the value of this attribute is 1 for the correct analysis and 0 for all other analyses. Automatic disambiguation tools can use any values for this attribute, to rank the analyses. The *score* element is not provided in our automatically analyzed corpora yet, and is systematically removed from Figures 4, 5.

4.3. Acquisition

Our goal is to acquire a representative corpus of contemporary Hebrew. Due to copyright and budget limitations we used resources that are freely available. The seed of our corpus was acquired from archives of the *Ha'aretz* daily newspaper[7] from 1991, dealing mainly with foreign affairs. Attempts to obtain more recent archives of the three main Israeli newspapers proved futile, and we resorted to collecting copyright-free

[6] An article also includes meta-data, such as its source, the author, the date of production, etc.

[7] `http://www.haaretz.co.il/`

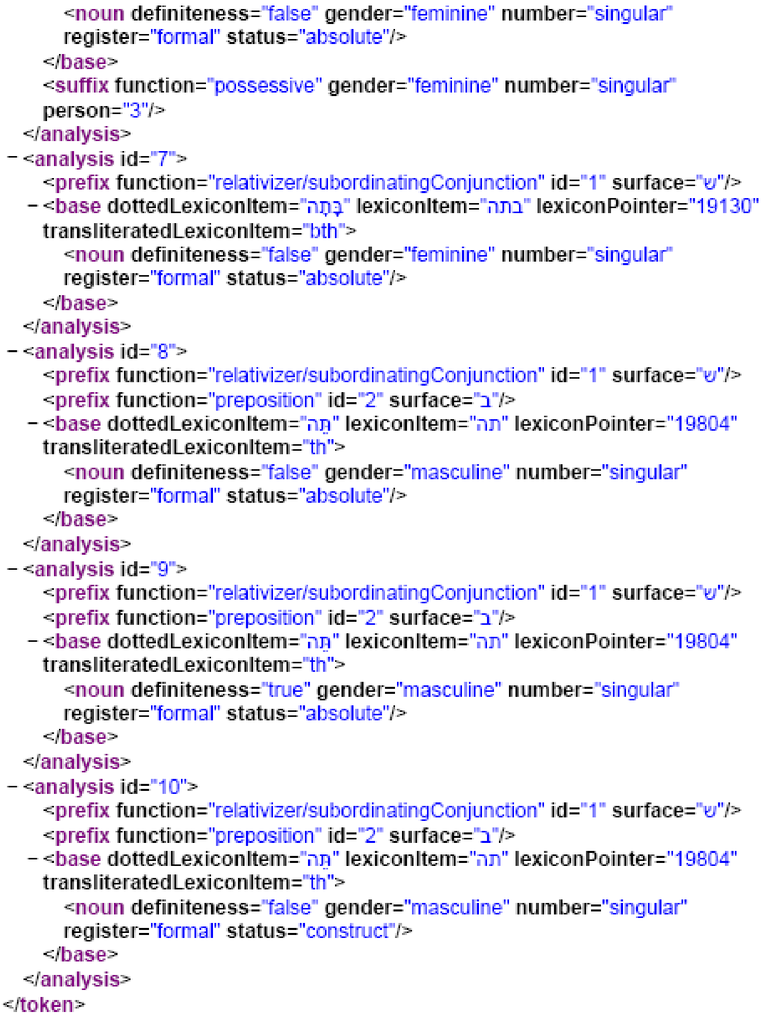


```
- <token id="1" surface="שבתה">
  - <analysis id="1">
    - <base dottedLexiconItem="שָׁבָה" lexiconItem="שבה" lexiconPointer="1541"
      transliteratedLexiconItem="ebh">
        <verb binyan="Pa'al" gender="feminine" number="singular" person="3"
        register="formal" root="שבה" tense="past"/>
      </base>
    </analysis>
  - <analysis id="2">
    - <base dottedLexiconItem="ישֵׁב" lexiconItem="ישב" lexiconPointer="1636"
      transliteratedLexiconItem="ieb">
        <verb binyan="Pa'al" register="formal" root="ישב" tense="infinitive"/>
      </base>
    </analysis>
  - <analysis id="3">
    - <base dottedLexiconItem="שָׁבַת" lexiconItem="שבת" lexiconPointer="9430"
      transliteratedLexiconItem="ebt">
        <verb binyan="Pa'al" gender="feminine" number="singular" person="3"
        register="formal" root="שבת" tense="past"/>
      </base>
    </analysis>
  - <analysis id="4">
    - <base dottedLexiconItem="שַׁבָּת" lexiconItem="שבת" lexiconPointer="17280"
      transliteratedLexiconItem="ebt">
        <noun definiteness="false" gender="feminine" number="singular"
        register="formal" status="absolute"/>
      </base>
      <suffix function="possessive" gender="feminine" number="singular"
      person="3"/>
    </analysis>
  - <analysis id="5">
    - <base dottedLexiconItem="שָׁב" lexiconItem="שב" lexiconPointer="19939"
      transliteratedLexiconItem="eb">
        <participle binyan="Pa'al" definiteness="false" gender="feminine"
        number="singular" person="any" register="formal" root="שוב"
        status="absolute"/>
      </base>
      <suffix function="possessive" gender="feminine" number="singular"
      person="3"/>
    </analysis>
  - <analysis id="6">
      <prefix function="relativizer/subordinatingConjunction" id="1" surface="ש"/>
    - <base dottedLexiconItem="בַּת " lexiconItem="בת" lexiconPointer="1379"
      transliteratedLexiconItem="bt">
```

Figure 4. An example of a fully analyzed corpus, the token *šbth*

texts from the Web. Over a period of several months, we collected all the articles that were published on the website of the online newspaper[8] *Arutz 7*. The texts are mostly short newswire articles, dealing mainly with domestic politics. Obviously, the domain and the source of this corpus bias word frequency distribution to some extent, but its availability facilitated the collection of over 15 million word tokens. A small corpus was collected from on-line articles of *The Marker*, a financial

[8] http://www.inn.co.il/



```
      <noun definiteness="false" gender="feminine" number="singular"
      register="formal" status="absolute"/>
    </base>
    <suffix function="possessive" gender="feminine" number="singular"
    person="3"/>
  </analysis>
- <analysis id="7">
    <prefix function="relativizer/subordinatingConjunction" id="1" surface="ש"/>
  - <base dottedLexiconItem="בָּתָה" lexiconItem="בתה" lexiconPointer="19130"
    transliteratedLexiconItem="bth">
      <noun definiteness="false" gender="feminine" number="singular"
      register="formal" status="absolute"/>
    </base>
  </analysis>
- <analysis id="8">
    <prefix function="relativizer/subordinatingConjunction" id="1" surface="ש"/>
    <prefix function="preposition" id="2" surface="ב"/>
  - <base dottedLexiconItem="תֵּה" lexiconItem="תה" lexiconPointer="19804"
    transliteratedLexiconItem="th">
      <noun definiteness="false" gender="masculine" number="singular"
      register="formal" status="absolute"/>
    </base>
  </analysis>
- <analysis id="9">
    <prefix function="relativizer/subordinatingConjunction" id="1" surface="ש"/>
    <prefix function="preposition" id="2" surface="ב"/>
  - <base dottedLexiconItem="תֵּה" lexiconItem="תה" lexiconPointer="19804"
    transliteratedLexiconItem="th">
      <noun definiteness="true" gender="masculine" number="singular"
      register="formal" status="absolute"/>
    </base>
  </analysis>
- <analysis id="10">
    <prefix function="relativizer/subordinatingConjunction" id="1" surface="ש"/>
    <prefix function="preposition" id="2" surface="ב"/>
  - <base dottedLexiconItem="תֵּה" lexiconItem="תה" lexiconPointer="19804"
    transliteratedLexiconItem="th">
      <noun definiteness="false" gender="masculine" number="singular"
      register="formal" status="construct"/>
    </base>
  </analysis>
</token>
```

Figure 5. An example of a fully analyzed corpus, the token *šbth* (cont.)

newspaper.[9] We have also acquired a corpus similar to the Hansard corpus: transcripts of two years of the Knesset (Israeli parliament) proceedings.[10] Table IX details the sizes of these corpora.

The final source is a corpus of partially dotted newspaper items. *Shaar la-Matxil*[11] is a newspaper for students of Hebrew. It is written in simple Hebrew and is partially dotted, i.e., every morphologically

[9] http://www.themarker.com/
[10] http://www.knesset.gov.il/
[11] http://slamathil.allbiz.co.il/

Table IX. Corpora sizes

	Haaretz	Arutz 7	Knesset	The Marker
Tokens	11,097,790	15,107,618	15,066,731	692,919
Types	305,545	323,943	204,967	62,216

ambiguous word contains sufficient information to disambiguate it. We have encountered technical difficulties in cleaning this corpus and thus far only part of it (approximately one million tokens) is publicly available. We are in the process of automatically supplying the missing dots, and providing a nearly full morphological disambiguation. When completed we shall have a large disambiguated Hebrew corpus.

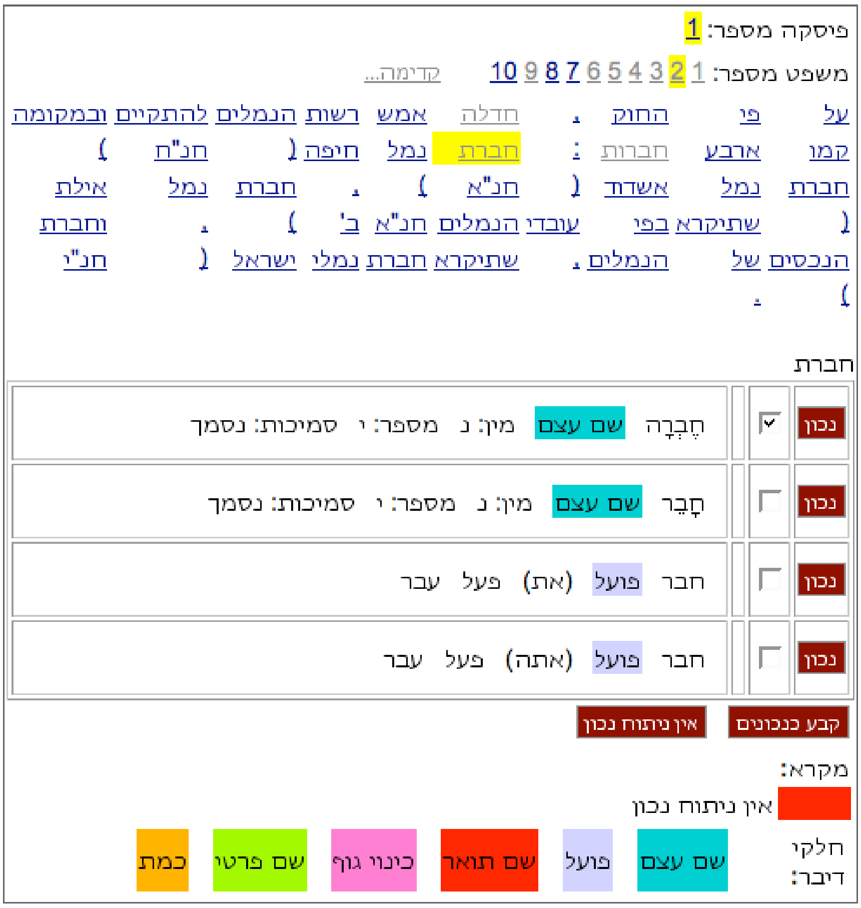
For the syntactically annotated corpus, we continued the work of Sima'an et al. (2001), who manually annotated a small treebank using a morphological analyzer (Segal, 1999) and the SEMTAGS annotation tool of Bonnema (1997). We added some 4,000 syntactically annotated sentences to the corpus of Sima'an et al. (2001), and slightly changed their annotation scheme according to the problems encountered while working on these additional data.

We plan to increase and diversify the corpora using additional resources. Since publishing houses have refused to cooperate we are negotiating with other creators of dynamic content on the Web.

4.4. Annotation tools

Annotated corpora are among the most important resources for training and evaluating NLP applications. The morphologically annotated corpus discussed above proved invaluable for training our morphological disambiguation module (Section 3.3).

To aid the annotators, we developed a graphical user interface which reads a morphologically analyzed corpus, displays it sentence by sentence, presents all the analyses for each word and allows the annotator to select the correct one. The tool is web-based to facilitate portability, and is written in JSP. A major design decision was to enable the annotator to make simple decisions fast, so that when a valid analysis is available, a single mouse click suffices to select it and move to the following word. If no analysis is correct, again a single click marks all analyses as wrong and moves to the following word. Finally, if the annotator is undecided among several analyses, more than one can be selected. This can happen, among other reasons, because sometimes two analyses are identical up to the lemma, and since not all the lemmas in our lexicon are dotted, two analyses can appear to be completely identical. Figure 6 depicts the annotation user interface.



פיסקה מספר: 1

משפט מספר: 1 2 3 4 5 6 7 8 9 10 קדימה...

על פי החוק , חדלה אמש רשות הנמלים להתקיים ובמקומה
קמו ארבע חברות : חברת נמל חיפה (חנ"ח)
חברת נמל אשדוד (חנ"א) , חברת נמל אילת
(שתיקרא בפי עובדי הנמלים חנ"א ב') , וחברת
הנכסים של הנמלים , שתיקרא חברת נמלי ישראל (חנ"י
) .

חברת

נכון	☑	חֶבְרָה	שם עצם	מין: נ	מספר: י	סמיכות: נסמך
נכון	☐	חָבֵר	שם עצם	מין: נ	מספר: י	סמיכות: נסמך
נכון	☐	חבר	פועל	(את)	פעל	עבר
נכון	☐	חבר	פועל	(אתה)	פעל	עבר

קבע כנכונים אין ניתוח נכון

מקרא:
🟥 אין ניתוח נכון

חלקי דיבור: שם עצם פועל שם תואר כינוי גוף שם פרטי כמות

Figure 6. Graphical user interface for morphological annotation

4.5. Named entities

To facilitate Named Entity (NE) Recognition tasks we extend the corpus schema such that NEs can be represented. There are at least two ways to encode NEs: by adding a *named entity* element "between" *sentence* and *token*, such that NEs are enclosed by *named entity* tags, whereas other tokens are not; or by adding a *named entity* attribute to *token* elements. We opted for the latter in order to minimally affect existing tools that were built for the corpus schema. We add two optional attributes to *token* elements: *enamex*, whose value indicates the type of the named entity (currently, *person, location, organization* or *none*); and *neid*, whose value is the serial number of the named entity on the sentence. Since NEs can span more than one token, this uniquely determines when a sequence of tokens forms one or more NEs.

We developed a graphical user interface for annotating named entities: the GUI takes as input a morphologically disambiguated corpus and presents its text to the annotator. Using simple mouse-operated actions, the annotator can mark NEs which are then recorded in the corpus following the enhancements described above. The output is a new corpus which can then be used to train and evaluate NER tasks. We are currently annotating a 2,000 sentence corpus for named entities.

5. Conclusions and further research

We have presented resources and tools for processing Hebrew, outlining the design principles underlying them and emphasizing the role of XML as a means for facilitating inter-operability of the resources and systems. The described resources are still under development and are updated on a daily basis. All the resources are available in their current state for both research and commercial uses.

We plan to diversify the corpora to make them more representative, and to extend the lexicon by adding more entries, dotted lemmas, translation equivalents and, eventually, also definitions. In addition to these extensions, our current research focuses on NLP applications which are compatible with the described resources, such as named entity recognition, shallow parsing, machine translation etc. Our main goal is to provide a centralized, high-quality repository of resources for processing Hebrew, to be used by researchers and software developers.

Acknowledgments

This work was funded by the Israeli Ministry of Science and Technology. Parts of this project were supported by THE ISRAEL SCIENCE FOUNDATION (grant No. 137/06); by the Israel Internet Association; and by the Caesarea Rothschild Institute for Interdisciplinary Application of Computer Science at the University of Haifa. Several people were involved in this work, and we are extremely grateful to all of them: Meni Adler, Roy Bar-Haim, Dalia Bojan, Ido Dagan, Michael Elhadad, Nomi Guthmann, Adi Milea, Noam Ordan, Erel Segal, Danny Shacham, Shira Schwartz, Yoad Winter, and Shlomo Yona. We are grateful to the reviewers for useful comments.

References

Abney, S.: 1996, ‘Statistical Methods and Linguistics’. In: J. Klavans and P. Resnik (eds.): *The Balancing Act: Combining Symbolic and Statistical Approaches to Language*. Cambridge, MA: The MIT Press.

Adler, M. and M. Elhadad: 2006, ‘An Unsupervised Morpheme-Based HMM for Hebrew Morphological Disambiguation’. In: *Proceedings of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the Association for Computational Linguistics*. Sydney, Australia, pp. 665–672.

Agirre, E. and G. Rigau: 1996, ‘Word sense disambiguation using Conceptual Density’. In: *Proceedings of the 16th conference on Computational linguistics*. Morristown, NJ, USA, pp. 16–22.

Bar-Haim, R., K. Sima’an, and Y. Winter: 2005, ‘Choosing an Optimal Architecture for Segmentation and POS-Tagging of Modern Hebrew’. In: *Proceedings of the ACL Workshop on Computational Approaches to Semitic Languages*. Ann Arbor, Michigan, pp. 39–46.

Bar-haim, R., K. Sima’an, and Y. Winter: 2008, ‘Part-of-speech tagging of Modern Hebrew text’. *Natural Language Engineering*. To appear.

Barkali, S.: 2000a, *Lux HaP’alim HaShalem (The Complete Verbs Table)*. Jerusalem: Rubin Mass, 51 edition. In Hebrew.

Barkali, S.: 2000b, *Lux HaShemot (The Nouns Table)*. Jerusalem: Rubin Mass, 18 edition. In Hebrew.

Beesley, K. R. and L. Karttunen: 2003, *Finite-State Morphology: Xerox Tools and Techniques*. Stanford: CSLI.

Bentivogli, L., E. Pianta, and C. Girardi: 2002, ‘MultiWordNet: developing an aligned multilingual database’. In: *Proceedings of the First International Conference on Global WordNet*. Mysore, India.

Black, W., S. Elkateb, H. Rodriguez, M. Alkhalifa, P. Vossen, A. Pease, and C. Fellbaum: 2006, ‘Introducing the Arabic WordNet Project’. In: *Proceedings of the Third Global WordNet Meeting*.

Bonnema, R.: 1997, ‘Data Oriented Semantics’. Master’s thesis, University of Amsterdam.

Buckwalter, T.: 2002, ‘Buckwalter Arabic Morphological Analyzer’. Distributed through LDC as LDC2002L49.

Connolly, D.: 1997, *XML: Principles, Tools, and Techniques*. O’Reilly.

Dahan, H.: 1997, *Hebrew–English English–Hebrew Dictionary*. Jerusalem: Academon.

Daya, E., D. Roth, and S. Wintner: 2004, ‘Learning Hebrew Roots: Machine Learning with Linguistic Constraints’. In: *Proceedings of EMNLP’04*. Barcelona, Spain, pp. 357–364.

de Buenaga Rodríguez, M., J. M. G. Hidalgo, and B. Díaz-Agudo: 1997, ‘Using WordNet to Complement Training Information in Text Categorization’. In: *Proceedings of the 2nd International Conference on Recent Advances in Natural Language Processing*.

Diab, M.: 2004, ‘The Feasibility of Bootstrapping an Arabic WordNet leveraging Parallel Corpora and an English WordNet’. In: *Proceedings of the Arabic Language Technologies and Resources*. Cairo, Egypt.

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

DuBois, P.: 1999, *MySQL*. New Riders.

Fellbaum, C. (ed.): 1998, *WordNet: An Electronic Lexical Database*, Language, Speech and Communication. MIT Press.

Fellbaum, C., M. Palmer, H. T. Dang, L. Delfs, and S. Wolf: 2001, 'Manual and Automatic Semantic Annotation with WordNet'. In: *Proceedings of WordNet and Other Lexical Resources Workshop*.

Gadish, R. (ed.): 2001, *Klalei ha-Ktiv Hasar ha-Niqqud*. Academy for the Hebrew Language, 4th edition. In Hebrew.

Habash, N. and O. Rambow: 2005, 'Arabic Tokenization, Part-of-Speech Tagging and Morphological Disambiguation in One Fell Swoop'. In: *Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics (ACL'05)*. Ann Arbor, Michigan, pp. 573–580.

Harabagiu, S. (ed.): 1998, *Usage of WordNet in Natural Language Processing Systems: Proceedings of the Coling-ACL 1998 Workshop*. Montreal, Canada: Association for Computational Linguistics.

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

Ide, N., P. Bonhomme, and L. Romary: 2000, 'XCES: An XML-based Encoding Standard for Linguistic Corpora'. In: *Proceedings of the Second International Language Resources and Evaluation Conference*. Paris.

Ide, N., L. Romary, and E. de la Clergerie: 2003, 'International standard for a linguistic annotation framework'. In: *SEALTS '03: Proceedings of the HLT-NAACL 2003 workshop on Software engineering and architecture of language technology systems*. Morristown, NJ, USA, pp. 25–30.

Ide, N. M. and J. Veronis (eds.): 1995, *Text Encoding Initiative: Background and Contexts*. Norwell, MA, USA: Kluwer Academic Publishers.

Itai, A.: 2006, 'Knowledge Center for Processing Hebrew'. In: *Proceedings of the LREC-2006 Workshop "Towards a Research Infrastructure for Language Resources"*. Genoa, Italy.

Itai, A., S. Wintner, and S. Yona: 2006, 'A Computational Lexicon of Contemporary Hebrew'. In: *Proceedings of The fifth international conference on Language Resources and Evaluation (LREC-2006)*. Genoa, Italy.

Jing, H.: 1998, 'Usage of WordNet in Natural Language Generation'. In: S. Harabagiu (ed.): *Usage of WordNet in Natural Language Processing Systems: Proceedings of the Coling-ACL 1998 Workshop*. pp. 128–134.

Lavie, A., S. Wintner, Y. Eytani, E. Peterson, and K. 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.

Mandala, R., T. Tokunaga, H. Tanaka, A. Okumura, and K. Satoh: 1998, 'Ad Hoc Retrieval Experiments Using WordNet and Automatically Constructed Thesauri'. In: *TREC*. pp. 414–419.

Manning, C. D. and H. Schütze: 1999, *Foundations of Statistical Natural Language Processing*. Cambridge, MA: MIT Press.

Ordan, N. and S. Wintner: 2007, 'Hebrew WordNet: a test case of aligning lexical databases across languages'. *International Journal of Translation, special issue on Lexical Resources for Machine Translation* **19**(1), 39–58.

Segal, E.: 1997, 'Morphological Analyzer for Unvocalized Hebrew Words'. Unpublished work.

Segal, E.: 1999, 'Hebrew Morphological Analyzer for Hebrew undotted texts'. Master's thesis, Technion, Israel Institute of Technology, Haifa. In Hebrew.

Shacham, D. and S. Wintner: 2007, 'Morphological disambiguation of Hebrew: a case study in classifier combination'. In: *Proceedings of EMNLP-CoNLL 2007, the Conference on Empirical Methods in Natural Language Processing and the Conference on Computational Natural Language Learning.* Prague.

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

Sima'an, K., A. Itai, Y. Winter, A. Altman, and N. Nativ: 2001, 'Building a Tree-Bank of Modern Hebrew Text'. *Traitment Automatique des Langues* **42**(2).

Sperberg-McQueen, C. M. and L. Burnard (eds.): 2002, *Guidelines for Text Encoding and Interchange.* Oxford: University of Oxford.

Stern, N.: 1994, *Milon ha-Poal.* Bar Ilan University. In Hebrew.

Szpektor, I., I. Dagan, A. Lavie, D. Shacahm, and S. Wintner: 2007, 'Cross Lingual and Semantic Retrieval for Cultural Heritage Appreciation'. In: *Proceedings of the ACL-2007 Workshop on Language Technology for Cultural Heritage Data (LaTeCH 2007).* Prague.

van der Vlist, E.: 2002, *XML Schema.* O'Reilly.

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

Wintner, S.: 2007, 'Finite-state Technology as a Programming Environment'. In: A. Gelbukh (ed.): *Proceedings of the Conference on Computational Linguistics and Intelligent Text Processing (CICLing-2007)*, Vol. 4394 of *Lecture Notes in Computer Science.* Berlin and Heidelberg, pp. 97–106.

Wintner, S. and S. Yona: 2003, 'Resources for processing Hebrew'. In: *Proceedings of the MT-Summit IX workshop on Machine Translation for Semitic Languages.* New Orleans, pp. 53–60.

Yona, S. and S. Wintner: 2005, 'A Finite-State Morphological Grammar of Hebrew'. In: *Proceedings of the ACL Workshop on Computational Approaches to Semitic Languages.* Ann Arbor, Michigan, pp. 9–16.

Yona, S. and S. Wintner: 2007, 'A Finite-State Morphological Grammar of Hebrew'. *Natural Language Engineering.* To appear.

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