Resources for Processing Hebrew

Shuly Wintner and Shlomo Yona

Department of Computer Science
University of Haifa, Israel
{shuly,shlomo}@cs.haifa.ac.il

Abstract

We describe work in progress whose main objective is to create a collection of resources and tools for processing Hebrew. These resources include corpora of written texts, some of them annotated in various degrees of detail; tools for collecting, expanding and maintaining corpora; tools for annotation; lexicons, both monolingual and bilingual; a rule-based, linguistically motivated morphological analyzer and generator; and a WordNet for Hebrew. We emphasize the methodological issue of well-defined standards for the resources to be developed. The design of the resources guarantees their reusability, such that the output of one system can naturally be the input to another.

1 Introduction

The state of the art in computational processing of Hebrew, as described by Wintner (2003), leaves much to be desired. Much of the infrastructure required both for practical applications and for computational linguistics research is either non-existent, lacking or proprietary. In this paper we describe work in progress whose main objective is to create a collection of resources and tools which are instrumental in most conceivable applications of natural language processing, in particular machine translation. These resources include corpora of written Hebrew, some of them annotated in various degrees of detail; tools for collecting, expanding and maintaining corpora; tools for annotation; lexicons, both monolingual and bilingual; a rule-based, linguistically motivated morphological analyzer and generator; and a WordNet for Hebrew.

We emphasize the methodological issue of well-defined standards for the resources to be developed. In particular, we use XML for defining the structure of corpora, annotated corpora, lexicons and morphological analyses. The design of the resources guarantees their reusability; in particular, it is essential that all the systems we develop adhere to the same standards, such that the output of one can naturally be the input to another. While the work we describe here is specific to Hebrew, the methodological principles which guide it are language independent.

In the next section we list some facts about the language. Section 3 describes the existing corpora, their structure and annotation, as well as tools for expanding and maintaining them. The ongoing work on the development of a morphological analyzer and generator is discussed in section 4. We briefly discuss the construction of a Hebrew WordNet in section 5. We conclude with plans for future research.

2 Facts about the language

Israeli Hebrew (also known as Modern Hebrew, henceforth *Hebrew*) is one of the two official languages of the State of Israel, spoken natively by half of the population (Izre'el, Hary, and Rahav, 2002) and fluently by virtually all the (over six million) residents of the country. In spite of some propositions that claim that it is a Creole, an Indo-European (Horvath and Wexler, 1997) or a hybrid language (Zuckermann, 2003), 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 three (typically) or more consonants and patterns are sequences of vowels and, sometimes, also consonants, with "slots" into which the root's consonants are being inserted. Inflectional morphology is highly productive and consists mostly of suffixes, but sometimes of prefixes or circumfixes. In general, inflectional morphology can be assumed to be concatenative, but derivational morphology is certainly non-concatenative (of course, even concatenative processes involve morph-phonological and orthographic alternations).

The Hebrew script, not unlike the Arabic one, attaches several short particles to the word which immediately follows them. These include, *inter alia*, the definite article *h* ("the"), prepositions such as *b* "in", *k* "as", *l* "to" and *m* "from", subordinating conjunctions such as *$* "that" and *k$* "when", relativizers such as *$* "that" and the coordinating conjunction *w* "and". The script is rather ambiguous as many of the prefix particles can also be parts of the stem. To facilitate readability we use a transliteration of Hebrew using ASCII characters in this paper; the tools we describe use the Hebrew script.

An added complexity stems from the fact that there exist two main standards for the Hebrew script: one in which vocalization diacritics, known as *niqqud* "dots", decorate the words, and another 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.

3 Corpora of Hebrew texts

The importance of large-scale corpora for automated language processing needs no introduction. To the best of our knowledge, no Hebrew corpora are publicly available (the only corpus we are aware of was developed as part of the Responsa project (Choueka, 1980), it does not contain modern texts and is not freely available). We have collected a set of more than 2500 articles from the Hebrew daily newspapers HaAretz, Ma'ariv and Yediot; in addition, we are constantly acquiring more data, mainly short newswire articles from the radio station Arutz 7. Our corpora currently contain more than seven million word tokens, and the number is constantly growing. See `http://cl.haifa.ac.il/corpora/`.

The texts in the corpus are acquired from the Internet, and thus are "noisy" in several respects: they contain a combination of Hebrew, English, graphics and HTML tags; they do not adhere to any standard of spelling; they include a very high rate of proper names; etc. As a first stage of processing, we developed an algorithm for cleaning up the texts. The main guideline was to prefer accuracy over coverage: when the algorithm cannot decide what constitutes a sequence of Hebrew words, we prefer to skip the sequence. This results in texts which we believe are Hebrew paragraphs.

Using simple heuristics for detecting end of sentences, we segment the texts into sentences. (See the `Lingua::HE::Sentence` Perl module at `http://search.cpan.org/dist/Lingua-HE-Sentence/`.) Next, we tokenize each sentence. The guideline for tokenization is oversimplistic: we define as a token any sequence of alphabetic characters, including the '`"`' which are used to denote abbreviations, delimited by any other character. Clearly, this algorithm misses many of the tokens: for example, a token such as *tl 'bib* "Tel Aviv" will be tokenized as *two* independent tokens. The main reason for this decision is that it is easiest to implement. Any more elaborate definition of what a token is will inevitably run into problems which would necessitate, in the worst case, full morphological analysis before tokenization. As we wish to separate the two tasks, this seems to be a reasonable compromise. Multi-word tokens and collocations will be handled by later processing.

After tokenization, the texts are annotated morphologically using an automatic morphological analyzer (Segal, 1999); two versions of the analyzer exist: one in which each word is assigned all its analyses, independent of its context, and another in which morphological ambiguity is resolved by heuristics and short-context considerations. While the morphological analyzer is freely available, its current state is insufficient for serious applications: its dictionary is too limited and its accuracy, even without disambiguation, is rather poor (it is currently being upgraded, but the revised version is unavailable yet). We create two versions of the corpus, one with all the analyses and the other with disambiguation.

Finally, texts are represented in XML, using a dedicated schema that we have designed. A proposal for XML representation of Hebrew corpora is given by Sasaki (2002); however, this proposal assumes access to several features that an automatic processor is not likely to produce (including full disambiguated syntactic structure, for example). Instead, our XML schema lists information about the corpus, the specific article in a corpus, a paragraph in an article, a sentence in the article and a word token in the sentence. Then, each token is associated with its analyses according to the analyzer described above. We return to this XML schema in the next section.

Results: the tools we have developed in this part of the project include a program for "cleaning up" texts (mostly removing HTML tags, non-textual material, foreign language text etc.); a program for sentence boundary detection; and a tokenizer. The resources we currently have include more than 2500 newspaper texts, comprising 1307244 tokens and 107641 word types. The Arutz 7 corpus contains 55310 articles, 6,353,382 tokens and 188,798 types. The corpora are given in four formats: raw text; XML tokenized texts; XML morphologically annotated texts; and XML annotated and disambiguated texts. The XML schemas that we defined for the lexicon and the annotated corpora are available at `http://cl.haifa.ac.il/corpora/`. An example of an annotated, disambiguated sentence (*bn 'li&zr dwxh htxrwt b$wq htq$wrt*; "Ben Eliezer postpones the deregulation of the telecommunication market") is given in the Appendix.

4 Morphological analysis and generation

Existing morphological analyzers for Hebrew are either limited (Ornan, 1985; Ornan, 1987; Segal, 1999) or proprietary (Bentur, Angel, and Segev, 1992; Choueka, 1993; Choueka and Ne'eman, 1995). Our objective in this project is to create a morphological analyzer for Hebrew which will be (1) broad-coverage, (2) in the public domain and (3) based on finite-state linguistically motivated rules.

The advantages of using finite-state technology (FST) for this task are straight-forward. First, it is beneficial to state the morphological, morphophonological and orthographic rules of the language in a way that is human-, as well as machine-readable. FST provides such capability by extending the language of regular expressions with a set of dedicated *replace rules*, which very naturally correspond to the way linguists think about morphological and phonological processes (Kaplan and Kay, 1994; Karttunen, 1997). An algorithmic formulation of morphological rules for Hebrew noun and verb inflections is provided by Ornan (2003). Second, FST compiles rules into finite-state networks which are extremely efficient to process. Finally, the technology is completely declarative: once an analyzer is given, it can immediately serve also as a generator. This property is extremely valuable for applications such as machine translation.

For this project we use the XFST finite-state toolbox (Beesley and Karttunen, 2003). We divide the design of the analyzer into two phases: the lexicon and the set of rules. The lexicon lists base forms (lexemes), with some additional information as discussed below. The rules implement inflectional morphology, morphological alternations, orthographic issues etc.

The structure of the lexicon is defined by an XML schema and the lexicon is represented in XML, so that its structure can be validated and easily extended and processed. At present, our lexicon contains relatively few entries, mainly to demonstrate each of the word categories that the morphological analyzer must be aware of. For each lexeme, the lexicon lists several features which are relevant for morphological analysis. Other lexical properties of words, e.g., definitions, glosses etc., can be easily added by extending the XML definition.

The most important property listed in the lexicon, from which all other features are derived, is the part of speech. Currently, the list of parts of speech includes adjective, adverb, auxiliaryVerb, conjunction, determiner, interrogative, noun, number, particle, preposition, pronoun, properName, punctuation and verb. More subtle classification is possible through a subcategory feature, which lists, e.g., valence for verbs, or nominalization for nouns.

Other features depend on the part of speech and include number, gender, person, root and pattern. More interestingly, certain features bear values which are instrumental for the morphological analyzer. These include, for example, the plural morpheme that a noun bears: by default this is *im* for the masculine, *wt* for the feminine, but as there are many idiosyncratic exceptions this information must be listed in the lexicon. The lexical entries of verbs specify not only their base forms but also secondary bases (e.g., the future base) which are usually hard to generate using finite-state machinery (Lavie et al., 1988) and involve a certain degree of arbitrariness.

The lexicon is associated with a program which converts the XML lexicon representation to XFST. The use of XML guarantees reusability in that the structure of the lexicon is application independent: other applications can use the same lexicon by supplying similar conversion programs. Out current lexicon contains a few hundred entries, including adjectives, adverbs, cardinal and ordinal numbers, conjunctions, existentials, nouns, particles, prepositions, pronouns, proper names and verbs.

The output of the analyzer is presented in the form of lexical strings, associated with the input surface string. See an example in figure 1. We convert this representation to XML format again. To this end, we use the XML schema which induces structure on morphologically annotated data. The schema is similar, but not identical, to the one used for the lexicon. Differences include an account of prefix particle sequences; morphological information such as status (absolute/construct) for nominals or tense for verbs; account of dependent pronominal suffixes, both in the noun (possessives) and in the verb (direct object markers); etc.

In order to evaluate the performance of the analyzer we are manually tagging a medium-sized corpus of newspaper articles (2000 sentences, approximately 30,000 word tokens). The annotation must be in a format that is consistent with the output of the analyzer: we simply use the same XML schema to define the format of the annotated data. Furthermore, we have implemented a graphical user interface for the annotator. The GUI is based on the XML schema and will ensure that the annotated data are always represented in a valid XML format, according to the specification of the schema. Note that one XML schema is used for three purposes here: representation of an analyzed corpus, the results of the morphological analysis (or the input for generation) and the annotation tool GUI.

Results: the morphological analyzer is still under development. All the inflectional morphology rules have been implemented, including closed-class words, the noun system and the verb system; however, the verb's weak paradigms have not been thoroughly tested yet. Of course, the main challenge is the extension of the lexicon, and in particular provisions for dynamic addition of new entries (mostly proper names). We currently have a small lexicon whose main purpose is to test the rule component; it contains most of the closed-class words (pronouns, prepositions, adverbs etc.), 350 nouns and 50 verbs, representative of most of the declension groups. Note that the analyzer is designed to produce all the possible analyses of each input form; disambiguation is deferred to a later stage of processing. The output of the morphological analyzer, formatted in XML, is exemplified in the Appendix.

5 Hebrew WordNet

WordNet (Fellbaum, 1998) is an on-line lexical reference system whose design is inspired by current psycholinguistic theories of human lexical memory. English nouns, verbs, adjectives and adverbs are organized into synonym sets (synsets), each representing one underlying lexical concept. Different relations, such as synonyms, antonyms, hypernyms, hyponyms, holonyms and meronyms, link the synonym sets. The system can be used for searching concepts, as well as the relations which link them.

Following the success of the English WordNet, similar networks have been developed for a variety of languages. In particular, a methodology for parallel construction of multilingual WordNets was developed and implemented as a system, called MultiWordNet (Bentivogli, Pianta, and Girardi, 2002). It contains information on several aspects of multilingual dictionaries, including lexical relationships between words, semantic relations over lexical concepts, several mappings of lexical concepts in different languages etc. MultiWordNet now contains lexical databases for English, Italian and Spanish, all aligned and synchronized.

```
BN[noun][Gender=Masculine][Number=Singular]
BN[noun][Gender=Masculine][Number=Singular][construct]
BN[proper noun]

ALI&ZR[proper noun]

DWXH[adjective][Gender=Feminine][Number=Singular]
DWXH[adjective][Gender=Masculine][Number=Singular]
DWXH[adjective][Gender=Masculine][Number=Singular][construct]
DWX[noun][Gender=Masculine][Number=Singular][construct]
   [dependent possessive pronoun][3P/F/Sg]
DXH[participle][Person=All][Gender=Feminine][Number=Singular]
DXH[participle][Person=All][Gender=Masculine][Number=Singular]
DXH[participle][Person=All][Gender=Masculine][Number=Singular][construct]


H[definite article]TXRH[noun][Gender=Feminine][Number=Plural]
H[definite article]TXRWT[noun][Gender=Feminine][Number=Singular]
HTXRWT[noun][Gender=Feminine][Number=Singular]
HTXRWT[noun][Gender=Feminine][Number=Singular][construct]

B[preposition][definite article]ewq[noun][Gender=Masculine][Number=Singular]
B[preposition]$WQ[noun][Gender=Masculine][Number=Singular]
B[preposition]$WQ[noun][Gender=Masculine][Number=Singular][construct]
B$WQ[proper noun]

H[definite article]TQ$WRT[noun][Gender=Feminine][Number=Singular]

L[preposition]PXT[verb][Tense=to-infinitive]
L[preposition]PXWT[adjective][Gender=Masculine][Number=Singular]
L[preposition]PXWT[adjective][Gender=Masculine][Number=Singular][construct]
L[preposition]PXWT[adverb]
L[preposition]PXWT[quantifier]
LPXWT[adverb]

&D[noun][Gender=Masculine][Number=Singular]
&D[noun][Gender=Masculine][Number=Singular][construct]
&D[preposition]

MARS[proper noun]
M[preposition]ARS[noun][Gender=Masculine][Number=Singular]
M[preposition]ARS[noun][Gender=Masculine][Number=Singular][construct]

2001[numeral]

.[punctuation]
```

Figure 1: An example of the morphological analysis

MultiWordNet has a variety of applications, including:

- Information retrieval: lexical relations can significantly improve the performance of query answering systems, for example; multilingual relationships facilitate multilingual information extraction and retrieval.

- Semantic annotation: since words in the network are tagged by the semantic concepts to which to relate, a multilingual WordNet can be used for semantic annotation and classification of texts.

- Disambiguation: semantic relationships can assist in determining the semantic distance between words and concepts, thereby assisting in lexical disambiguation.

- Terminology: the system can be used for developing structured terminologies for specific applications.

- Machine translation: as the different WordNets are aligned, word-sense accurate translation is a feasible possibility.

Our goal in this project is to use the MultiWordNet methodology for constructing a Hebrew WordNet, integrated with the one described above (and, therefore, aligned with English, Italian and Spanish). We will investigate the appropriateness of the methodology for adding a language of a completely different family into the unified framework; use bilingual dictionaries in order to semi-automatically acquire new synsets into the system; validate the consistency of the system by cross-checking it with the added Hebrew entries; and explore avenues for new applications of a multilingual lexical database for multilingual applications, such as cross-language information retrieval and machine translation.

Results: currently, very few word senses have been added to the system, mainly to demonstrate the support of a language which is written in a completely different character set, right-to-left. The main bottleneck is the acquisition of an on-line bilingual dictionary, which is essential for the methodology described above. We are currently in the last phases of adapting an existing dictionary (Dahan, 1997) for our needs. Once this is done, we will start adding word senses semi-automatically.

6 Conclusion

We described a set of linguistic tools and resources for computational processing of Hebrew. Much of the work discussed above is still ongoing, and it is early to provide a detailed evaluation of its results. However, we hope to have been able to show a consistent methodology in developing the resources which guarantees compatibility and reusability. The raw, tokenized and morphologically annotated corpora are already available. We expect the morphological analyzer to be fully functional by the end of 2003; the Hebrew WordNet is expected to be complete by the end of 2004.

Even when the projects described above are completed, much is still left to be done. In particular, we are determined to improve the morphological analyzer in two main ways: extension of the lexicon and disambiguation. The former will be done by defining a (supervised) machine learning algorithm for extracting the base, as well as significant morphological information, from unseen word forms. The latter will be implemented mostly in finite-state technology and will apply short-context rules, heuristics and statistical measures. We intend to implement a cascade of finite-state transducers (Abney, 1996) on top of the existing analyzer, realizing rules for detection of numeral expressions, dates, addresses, geographical names etc. Then, we will define rules for phrase boundary detection, culminating in a system which can perform shallow parsing efficiently and accurately.

Upon completion of these projects, Hebrew will have a set of resources and tools which will set up an infrastructure for both research and commercial applications. The developed resources will be available for research (and, hopefully, also for commercial) purposes in order to improve the quality of future NLP applications for Hebrew.

References

Abney, Steven. 1996. Partial parsing via finite-state cascades. In *Workshop on Robust Parsing, 8th European Summer School in Logic, Language and Information*, pages 8–15, Prague, Czech Republic.

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

Bentivogli, Luisa, Emanuele Pianta, and Christian Girardi. 2002. MultiWordNet: developing an aligned multilingual database. In *Proceedings of the First International Conference on Global WordNet*, Mysore, India, January.

Bentur, Esther, Aviella Angel, and Danit Segev. 1992. Computerized analysis of Hebrew words. *Hebrew Linguistics*, 36:33–38, December. In Hebrew.

Choueka, Yaacov. 1980. Computerized full-text retrieval systems and research in the humanities: The Responsa project. *Computers and the Humanities*, 14:153–169.

Choueka, Yaacov. 1993. Response to “Computerized analysis of Hebrew words”. *Hebrew Linguistics*, 37:87, December. In Hebrew.

Choueka, Yaacov and Yoni Ne’eman. 1995. “Nakdan-T”, a text vocalizer for modern Hebrew. In *Proceedings of the Fourth Bar-Ilan Symposium on Foundations of Artificial Intelligence*, June.

Dahan, Hiya. 1997. *Hebrew–English English–Hebrew Dictionary*. Academon, Jerusalem.

Fellbaum, Christiane, editor. 1998. *WordNet: An Electronic Lexical Database*. Language, Speech and Communication. MIT Press.

Horvath, Julia and Paul Wexler. 1997. *Relexification in Creole and Non-Creole Languages – With Special Attention to Haitian Creole, Modern Hebrew, Romani, and Rumanian*, volume xiii of *Mediterranean Language and Culture Monograph Series*. Harrassowitz, Wiesbaden.

Izre’el, Shlomo, Benjamin Hary, and Giora Rahav. 2002. Designing CoSIH: The corpus of Spoken Israeli Hebrew. *International Journal of Corpus Linguistics*, 6(2):171–197.

Kaplan, Ronald M. and Martin Kay. 1994. Regular models of phonological rule systems. *Computational Linguistics*, 20(3):331–378, September.

Karttunen, Lauri. 1997. The replace operator. In Emmanuel Roche and Yves Schabes, editors, *Finite-State Language Processing*, Language, Speech and Communication. MIT Press, Cambridge, MA, chapter 4, pages 117–147.

Lavie, Alon, Alon Itai, Uzzi Ornan, and Mori Rimon. 1988. On the applicability of two-level morphology to the inflection of Hebrew verbs. In *Proceedings of the International Conference of the ALLC*, Jerusalem, Israel.

Ornan, Uzzi. 1985. Indexes and concordances in a phonemic Hebrew script. In *Proceedings of the Ninth World Congress of Jewish Studies*, pages 101–108, Jerusalem, August. World Union of Jewish Studies. In Hebrew.

Ornan, Uzzi. 1987. Computer processing of Hebrew texts based on an unambiguous script. *Mishpatim*, 17(2):15–24, September. In Hebrew.

Ornan, Uzzi. 2003. *The Final Word*. University of Haifa Press, Haifa, Israel. In Hebrew.

Sasaki, Tsuguya. 2002. Building of an annotated corpus and a lexical database of Modern Hebrew in XML. In *34th Annual Conference of the Association for Jewish Studies*, Los Angeles, CA.

Segal, Erel. 1999. Hebrew morphological analyzer for Hebrew undotted texts. Master’s thesis, Technion, Israel Institute of Technology, Haifa, October. In Hebrew.

Wintner, Shuly. 2003. Hebrew computational linguistics: Past and future. *Artificial Intelligence Review*, 19.

Zuckermann, Ghil‘ad. 2003. *Language Contact and Lexical Enrichment in Israeli Hebrew*. Palgrave Macmillan, London – Ney York.

Appendix: Morphologically analyzed text in XML

```
<?xml version="1.0" encoding="UTF-8"?>

<corpus name="Example instance of hebrew_corpus.xsd" version="0.01"
  maintainer="Shlomo Yona" email="shlomo@cs.haifa.ac.il">

  <article id="1014" takenFrom="n151341.txt.fribidi.trimmed.tr.converted">
```

```
    <paragraph>
      <sentence>
        <token surface="" transliterated="BN">
          <analysis>
            <base lexiconItem="BN">
              <properName/>
            </base>
          </analysis>
        </token>
        <token surface="" transliterated="ALI&ZR">
          <analysis>
            <base lexiconItem="ALI&ZR">
              <properName/>
            </base>
          </analysis>
        </token>
        <token surface="" transliterated="DWXH">
          <analysis>
            <base lexiconItem="DXH">
              <verb tense="present"
                gender="masculine and feminine"
                number="singular" person="1 and 2 and 3"/>
            </base>
          </analysis>
        </token>
        <token surface="" transliterated="HTXRWT">
          <analysis>
            <prefix surface="h" function="definiteArticle"/>
            <base lexiconItem="TXRWT">
              <noun gender="feminine"
                number="singular"
                status="absolute"/>
            </base>
          </analysis>
        </token>
        <token surface="" transliterated="B$WQ">
          <analysis>
            <prefix surface="b" function="preposition"/>
            <base lexiconItem="$WQ">
              <noun gender="masculine"
                number="singular"
                status="construct"/>
            </base>
          </analysis>
        </token>
        <token surface="" transliterated="HTQ$WRT">
          <analysis>
            <prefix surface="h" function="definiteArticle"/>
            <base lexiconItem="TQ$WRT">
              <noun gender="feminine"
                number="singular"
                status="absolute"/>
            </base>
          </analysis>
        </token>
        <token surface="." transliterated=".">
          <analysis>
            <base>
              <punctuation/>
            </base>
          </analysis>
        </token>
      </sentence>
    </paragraph>
  </article>
</corpus>
```