

Resources for Processing Hebrew

Shuly Wintner and Shlomo Yona
Department of Computer Science
University of Haifa
{shuly, shlomo}@cs.haifa.ac.il

MT Summit IX, 23 Spetember 2003

Overview

We describe work in progress whose main objective is to create a collection of resources and tools for processing Hebrew, including:

- Corpora of written texts, 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;
- WordNet for Hebrew.

Motivation

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.

Objectives

The main objective of our work is to create a collection of resources and tools which are instrumental in most conceivable applications of natural language processing, in particular machine translation.

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, all the systems we develop adhere to the same standards, such that the output of one can naturally be the input to another.

While this work is specific to Hebrew, the methodological principles which guide it are language independent.

Structure

- Some facts about the language.
- Existing corpora, their structure and annotation, as well as tools for expanding and maintaining them.
- The development of a morphological analyzer and generator.
- The construction of a Hebrew WordNet.
- Plans for future research.

Facts about the language

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 six million) residents of the country.

Hebrew exhibits clear Semitic behavior. In particular, its lexicon, word formation and inflectional morphology are typically Semitic.

The major word formation machinery is root-and-pattern.

Inflectional morphology is highly productive and consists mostly of suffixes, but sometimes of prefixes or circumfixes.

Inflectional morphology can be assumed to be concatenative, but derivational morphology is certainly non-concatenative.

Facts about the language

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.

An added complexity stems from the fact that there exist two main standards for the Hebrew script: with or without vocalization diacritics, known as *niqqud* “dots”. Most of the texts in Hebrew are undotted; unfortunately, different authors use different conventions for the undotted script. This fact adds significantly to the ambiguity.

Corpora of Hebrew texts

Corpora of Hebrew texts

- Motivation
- Existing corpora
- Newly acquired corpora
- Our corpora currently contain more than seven million word tokens:
<http://cl.haifa.ac.il/corpora/>.

Corpora of Hebrew texts

Processing corpora:

- Cleaning up the texts;
- Segmenting the texts into sentences;
- Tokenization;
- Automatic morphological analysis ([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;
- Finally, texts are represented in XML, using a dedicated schema.

Corpora of Hebrew texts: results

- More than 2500 newspaper texts, comprising 1,307,244 tokens and 107,641 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.

Corpora of Hebrew texts: example

- Raw text
- After tokenization, formatted in XML
- Morphologically analyzed format in XML
- Morphologically analyzed and disambiguated format in XML
- Morphologically analyzed format in XML (new analyzer)

Morphological analysis and generation

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.

Morphological analysis and generation

The advantages of using finite-state technology (FST):

- It is beneficial to state the morphological, morpho-phonological and orthographic rules of the language in a way that is human-, as well as machine-readable.
- FST compiles rules into finite-state networks which are extremely efficient to process.
- 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.

Morphological analysis and generation

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 additional lexical information.

The rules implement inflectional morphology, morphological and morpho-phonological alternations, orthographic issues etc.

Lexicon

The structure of the lexicon is defined by an XML schema and the lexicon is represented in XML.

Our current lexicon contains a few hundred entries, including adjectives, adverbs, cardinal and ordinal numbers, conjunctions, existentials, nouns, particles, prepositions, pronouns, proper names and verbs.

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 lexicon is associated with a program which converts the XML lexicon representation to XFST.

Lexicon: example

- A schema for representing the lexicon
- An example lexicon

Morphological analysis and generation: results

The output of the analyzer is presented in the form of lexical strings, associated with the input surface string.

Example

Morphological analysis and generation: results

The output of the analyzer is converted 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.

Morphological analysis and generation: 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; the verb's weak paradigms have not been thoroughly tested yet.

The main challenge is the extension of the lexicon, and in particular provisions for dynamic addition of new entries (mostly proper names).

Morphological analysis and generation: evaluation

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 ensures 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.

A GUI for morphological annotation

GUI

Hebrew WordNet

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.

Hebrew WordNet

MultiWordNet: a methodology for parallel construction of multilingual WordNets, developed and implemented as a system.

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.

Hebrew WordNet

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.

Hebrew WordNet

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).

Hebrew WordNet: 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.

Conclusion

Ongoing work:

- Corpora
- Annotation schema and tools
- Morphological analyzer and generator
- Lexicon
- WordNet

Conclusion

Future work:

- Morphological disambiguation
- Machine learning techniques for expanding the lexicon
- A cascade of finite-state transducers , realizing rules for detection of numeral expressions, dates, addresses, geographical names etc.
- Shallow parsing

Bibliography

- [Beesley and Karttunen2003] Beesley, Kenneth R. and Lauri Karttunen. 2003. *Finite-State Morphology: Xerox Tools and Techniques*. CSLI, Stanford.
- [Bentur, Angel, and Segev1992] Bentur, Esther, Aviella Angel, and Danit Segev. 1992. Computerized analysis of Hebrew words. *Hebrew Linguistics*, 36:33–38, December. In Hebrew.
- [Choueka1993] Choueka, Yaacov. 1993. Response to “Computerized analysis of Hebrew words”. *Hebrew Linguistics*, 37:87, December. In Hebrew.
- [Choueka and Ne’eman1995] 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.
- [Dahan1997] Dahan, Hiya. 1997. *Hebrew–English English–Hebrew Dictionary*. Academon, Jerusalem.
- [Fellbaum1998] Fellbaum, Christiane, editor. 1998. *WordNet: An Electronic Lexical Database*. Language, Speech and Communication. MIT Press.
- [Ornan1985] 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.
- [Ornan1987] Ornan, Uzzi. 1987. Computer processing of Hebrew texts based on an unambiguous script. *Mishpatim*, 17(2):15–24, September. In Hebrew.
- [Segal1999] Segal, Erel. 1999. Hebrew morphological analyzer for Hebrew undotted texts. Master’s thesis, Technion, Israel Institute of Technology, Haifa, October. In Hebrew.
- [Wintner2003] Wintner, Shuly. 2003. Hebrew computational linguistics: Past and future. *Artificial Intelligence Review*, 19.