Knowledge Center for Processing Hebrew

Alon Itai
Head of the Knowledge Center for Processing Hebrew
Department of Computer Science
Technion, Israel Institute of Technology
32000 Haifa, Israel
Email: itai@cs.technion.ac.il

Introduction

To preserve cultural diversity it is necessary to preserve underrepresented languages. Such languages suffer from the dominance of English. This is aggravated by the Internet and the personal computer whose tools are tailored to English. Many tools cannot handle native alphabets, and these languages lack specific tools, such as spellers. Thus, for example, email and chats are often conducted in (pidgin) English. Moreover, even users of languages which use the Latin alphabet need language specific tool such as spellers, word processors and web searchers; these in turn often require linguistic tools such as morphological parsers, syntactic parsers, lexicons and bilingual dictionaries. Since the number of native speakers is relatively small there is little economic incentive for commercial companies to develop the tools. Thus there is a real danger that the more sophisticated users will abandon their native tongue in their professional work. It is the role of national and regional governments to carry the burden of preserving the native languages and to that end develop computer tools: software and databases.

It is also important to conduct linguistic research in such languages and order to do so one needs publicly available linguistic tools with open access source code software programs. For example, it is argued that one cannot meaningfully search documents in a language with complex morphology without using a morphological analyzer. If there is no publicly available morphological analyzer then every researcher has to reconstruct such a tool. Moreover, in order to build a high quality morphological analyzer one needs a high quality lexicon. Thus every researcher conducting corpus linguistics has to invest in a morphological analyzer and lexicon before starting her research.

This paper describes the Knowledge Center for Processing Hebrew whose aim is to make computer tools and databases for Hebrew available to the public and thus enhance the linguistic research of Modern Hebrew in both computational and theoretical linguistics, and to promote the commercial usage of NLP systems for Hebrew.

The Knowledge Center Model

In 2003, the Israeli Ministry of Science and Technology established a Knowledge Center for Processing Hebrew. Its aim was to develop products (software and databases) for processing Hebrew and make them available to the public, both in academia and industry. Researchers from four universities are involved with the Center's activities.

Before the establishment of the Center, the lack of standardization and centralization caused much duplication of effort. For example, several morphological analyzers of Hebrew were developed by different teams, using different methodologies and different output formats, and based on different lexicons (Choueka and Shapira 1964, Ornan 1987, Lavie et al. 1988, Bentur et al. 1992, Segal 1999, Yona and Wintner 2005). Since their output was different, and the source code was not available, it was impossible to compare them or reuse their resources. Furthermore, many of the developed tools were unavailable to the entire community.

Much of the Center's efforts are dedicated to transform software developed in academia for research purposes into tools available to the public. In research one wants to prove a concept, not to provide a commercial tool. Providing documentation, user interface and making the programs platform independent require a lot of work with little academic reward. Thus most often tools created in academia cannot be reused. The Center upgrades software and other tools created by academia and private researchers to make them reusable. The Center also provides a depository, thus researchers know from where to download tools.

The ministry's aim was to make the center self sustainable, i.e., the revenues from selling products and services should provide funds to maintain the Center. However, since the market is small, such revenues proved to be insufficient. Furthermore, had there been a commercial market there would have been no need to establish the Center.

Since our aim was to make the products available to the entire community in order to encourage research we have made all our products available under the GPL (Gnu public license http://www.gnu.org/copyleft/gpl.html), including the source code of software. This license allows free use but requires that all products that embed GPL products also be under GPL thus limiting the commercial use of our products. In order to enable commercial development, we allow the commercial use of products that do not contain embedded GPL components. This use is non-exclusive, i.e., the same products are also available for free use under the GPL.

Modern Hebrew

Modern Hebrew is one of the two official languages of the State of Israel, spoken natively by half of the population and fluently by virtually all the (seven million) residents of the country. The language is strongly related to (though linguistically distinct from) biblical Hebrew, and thus has raised the interests of both linguists and religious scholars.

Modern Hebrew exhibits clear Semitic behavior. In particular, its lexicon, word formation and inflectional morphology are typically Semitic. Its morphology is inflectional and highly productive and consists mostly of suffixes, but sometimes of prefixes or circumfixes. Often connectives and prepositions are prepended to words.

In the standard Hebrew script, like Arabic, most of the vowels are not represented. Thus Hebrew texts are highly ambiguous. 55% of the tokens are ambiguous; some tokens have up to 13 analyses, while the average number of analyses is over 2.

Thus a major difficulty in processing Hebrew is to morphologically disambiguate the text, i.e., choose the right analyses according to the context.

Products

The development of the products was 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;

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;

Suboptimal Efficiency: The resources and tools are not meant to compete with industrial products but instead to be easy to understand, easy to use and easy to expand. Thus, the resources and tools we provide are not always optimized for space and time.

Our linguistic databases are represented in Extensible Markup Language – XML (Connolly 1997) according to schemas that enforce structure and are also used for documentation and validation purposes. The output of the morphological analyzers and taggers is also in XML format. Thus we can use the software modularly and compare the outputs of different implementations.

The products include

1. XML standards for representing lexicons and corpora.
2. Segmentizers: Tokenizer, sentencizer (a program that partitions the corpus to sentences), word-segmentizer (a program that partitions the word into morphemes).
3. Morphological analyzers and taggers: The analyzers list all the possible analyses, whereas the taggers attempt to find the correct analysis in context.
4. Part of speech taggers (Bar Haim, Sima'an and Winter 2005).
5. Corpora: 20 million word corpora of printed press, 17 million words of Parliamentary proceedings, 1.3 million word corpora of printed press with partial niqqud (diacritical marks for vowels). All these corpora appear in XML format, and include morphological analysis and automatic tagging. A 6000 sentence morphologically manually tagged corpus is also available.
6. Graphical User Interfaces (GUI) for tagging and preparing lexicons.
7. Tree bank: 6000 syntactically parsed sentences (Sima'an 2001).
8. Lexicon: A full lexicon of Modern Hebrew containing over 21,000 entries (Itai, Wintner and Yona 2006).
9. Tools for processing phonemic script (Ornan 1987).
10. Speech analysis databases.

A full list of products is available at
http://mila.cs.technion.ac.il/website/english/resources/index.html

Conclusions

The Knowledge Center for Processing Hebrew was created for the sole purpose of developing a research infrastructure for language resources. It is a good example of a government-funded entity that functions as a language resource center and focuses on defining and enforcing standards, as well as developing and archiving linguistic databases (such as corpora and lexicons) and tools (such as morphological analyzers). It facilitates easy access to and sharing of resources through an open-source policy. The products developed at the Center have so far proved useful both for commercial applications and for linguistic, psycholinguistic and literary research.

The Knowledge Center provides a model that can be applied to other languages. Some of our products are language specific, whereas others can be adapted to other languages. However, this is not a solution for languages with very few speakers, since the cost of establishing a center is large and it is essential to have skilled professionals — linguists and programmers.

Acknowledgement

The Center for Processing Hebrew was funded by the Israeli Ministry of Science and Technology. We wish to thank the Center's members: Yoad Winter (Technion), Shuly Wintner (Haifa University), Michael Elhadad and the late Arnon Cohen (Ben Gurion University), Yoram Singer and Eli Shamir (Hebrew University) and to the numerous graduate students who were involved in the development of the resources. We also wish to thank the technical staff: Dalia Bojan, Adi Cohen-Milea, Danny Shacham, Shira Schwartz, and Shlomo Yona.

References

Bar-Haim, Roy, Khalil Sima'an and Yoad Winter 2005: Choosing an Optimal Architecture for Segmentation and POS-Tagging of Modern Hebrew. *ACL 2005 Workshop on Computational Approaches to Semitic Languages*.

Bentur, Esther, Aviella Angel, and Danit Segev, 1992. Computerized Analysis of Hebrew Words. *Hebrew Linguistics* 36, 33-38. (In Hebrew.)

Choueka Yaacov 1998, (Chief Editor, with the Rav-Milim team), *Rav-Milim - the complete dictionary of contemporary Hebrew*, Steimatzki, C.E.T. and Miskal, Tel-Aviv.

Choueka, Yaacov and M. Shapiro 1964, *Machine analysis of Hebrew morphology: potentialities and achievements*, Leshonenu (Journal of the Academy of the Hebrew Language), Vol. 27, 1964, 354 -372 (Hebrew).

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

Itai, Alon, Shuly Wintner and Shlomo Yona 2006. A Computational Lexicon of Contemporary Hebrew. In *Proceedings of LREC-2006*, Genoa, Italy, May 2006.

Lavie, Alon, Alon Itai, Uzzi Ornan, and Mori Rimon. 1988. On the applicability of two-level morphology to the inflection of Hebrew verbs. ALLC June 1988, Jerusalem, (TR 513, Department of Computer Science, Technion, 32000 Haifa, Israel).

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

Segal, Erel 1999 Hebrew Morphological Analyzer for Hebrew undotted texts *M.Sc. Theses.* Dept. of Computer Science,Technion, Israel Institute of Technology Haifa, Israel. October 1999.

Sima'an, Khalil, Alon Itai, Yoad Winter, Alon Altman and Noa Nativ 2001. Building a Tree-Bank of Modern Hebrew Text. *Traitment Automatique des Langues*, 42, 347-380.

Yona, Shlomo and Shuly Wintner 2005. A finite-state morphological grammar of Hebrew. In *Proceedings of the ACL-2005 Workshop on Computational Approaches to Semitic Languages*, Ann Arbor, MI, June 2005.